Correlation between American Mortality and DJIA Index Price

Ong, Li Kee

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

MASTER OF SCIENCE

Asper School of Business

The University of Manitoba

Winnipeg

Abstract

For an equity-linked insurance, the death benefit is linked to the performance of the company's investment portfolio. Hence, both mortality risk and equity return shall be considered for pricing such insurance. Several studies have found some dependence between mortality improvement and economy growth. In this thesis, we showed that American mortality rate and Dow Jones Industrial Average (DJIA) index price are negatively dependent by using several copulas to define the joint distribution. Then, we used these copulas to forecast mortality rates and index prices, and calculated the payoffs of a 10-year term equity-linked insurance. We showed that the predicted insurance payoffs will be smaller if dependence between mortality and index price is taken into account.

Keywords: mortality, DJIA (abbreviation of Dow Jones Industrial Average), Time series models, Outlier models, Copulas, Equity-linked Securities.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

For a guaranteed equity-linked insurance contract, the death benefit is linked to the performance of an investment portfolio, and hence the risk will be shared between the policyholder and the insurer (Bacinello, n.d.). If there exists negative dependence between the investment portfolio and mortality, a positive change in the investment portfolio may be accompanied by a negative change in the mortality. Consequently, if insurers do not consider the dependence between the investment portfolio and mortality, the equity-linked insurance may not be priced properly.

It is suggested that a good economy or a positive equity return may lead to mortality improvement as people will have more money to spend on healthcare resources. The converse may also be true. When people live longer, the economy would benefit from the strong labor markets and competitions; as a result, this may cause equity return to rise.

Several studies on the relationships between mortality improvement and economic growth have been conducted. For example, Ribeiro and Pietro (2009) concluded that, based on the data from 1900 to 2008, increases in the US mortality rate is followed by negative DJIA returns in the same year as well as the next five years. Furthermore, large decreases in the mortality rates also lead to positive asset returns in the same year and the next five years.

In addition, based on the data on mortality rates for different age groups and GDP of certain countries, Kalemli-Ozcan (2002) suggested that mortality decline working through the channels of education and fertility promotes economic growth, as mortality improvements have been an important incentive to increase investment in the education of children. On the other hand, Jamison, Jamison, and Hanushek (2007) showed that improvements in education quality is associated with declines in infant mortality. Their result was based on the input data, including GDP per capita, capital stock per capita, and years of education, of 62 countries.

Moreover, Preston (2007) examined the relationships between level of average life expectancy and national income per capita in the 1930s and 1960s of several countries; the linear correlation between life expectancy and logarithm of income per capita was found to be 0.885 in the 1930s and 0.880 in the 1960s. Adelman (1963) also examined the coefficients of correlation between infant mortality rates and levels of income. It has been found that the coefficients are the order of -0.8. Pritchett and Summers (1996) estimated the income elasticity of infant and child health with respect to infant mortality and concluded that increases in income lead to improvements of health status. Furthermore, Ettner (1996) also suggested that increases in income significantly improve mental and physical health based on estimations of structural impact of income on various measures of health.

Finally, based on age-adjusted[a] mortality rates over 1901-2000 in the United States and independent variables including real GDP per capita and the unemployment rate, Brenner (2005) argued that economic growth has been a strong factor in American mortality improvements over the 20th century.

Since equity return is linked to the economy, it is reasonable to also investigate the dependence between mortality and equity return for equity-linked insurance pricing purposes. This thesis examines the dependence and correlation, instead of cause/effect relationship, between American mortality rate and DJIA index price. In order to examine the correlations, instead of time series data, innovations from the time series models should be used as the inputs. In other words, if there exist some autocorrelations in the time series data, the correlation analysis should be performed on the residuals or standardized residuals, which are obtained from appropriate time series models (Patton, 2012).

Firstly, time series models will be fitted to Log index and Log mortality of age group "55 to 64 years". Then, linear as well as rank correlations between the residuals or standardized residuals can be examined.

[a] Age-adjusted death rate is calculated by giving different weights (instead of equal weights) to the death rate of different age groups. Usually the weight represents the proportion of the age group to the entire population.

Apart from the linear and rank correlations, for the extreme cases, tail dependence can be examined by using copulas.

Our analysis shows that, under some copulas, tail correlations/coefficients are larger than linear and rank correlations. In addition, these tail correlations are significant at 0.05 while linear and rank correlations are not.

Chapter 2

Time Series Models

This chapter reviews some important time models including ARIMA models, GARCH models, and general time-series outlier models. Our analysis show that, for our data series Log index and Log mortality, outlier models are preferred to ARIMA/GARCH models.

2.1 Data

Grouped data for 1900-1932 American mortality are available on `www.mortality.org/cgi-bin/hmd/country.php?cntr=USA&level=1`, and data for 1933-2010 can be found on `www.cdc.gov/nchs/data/databriefs/db88.htm#x2013;2010</a>`. In addition, Annual DJIA index price can be obtained from `www.measuringworth.com/DJIA_SP_NASDAQ/`. Figures 2.1 and 2.2 display the annual DJIA index price and annual mortality rates of age groups "55 to 64 years". While index price increased throughout the time, mortality rate decreased steadily. The reason mortality data of age group "55 to 64 years" is used is that the people from this age group are about to reach their retirement

Figure 2.1: Annual DJIA index price, 1900-2010

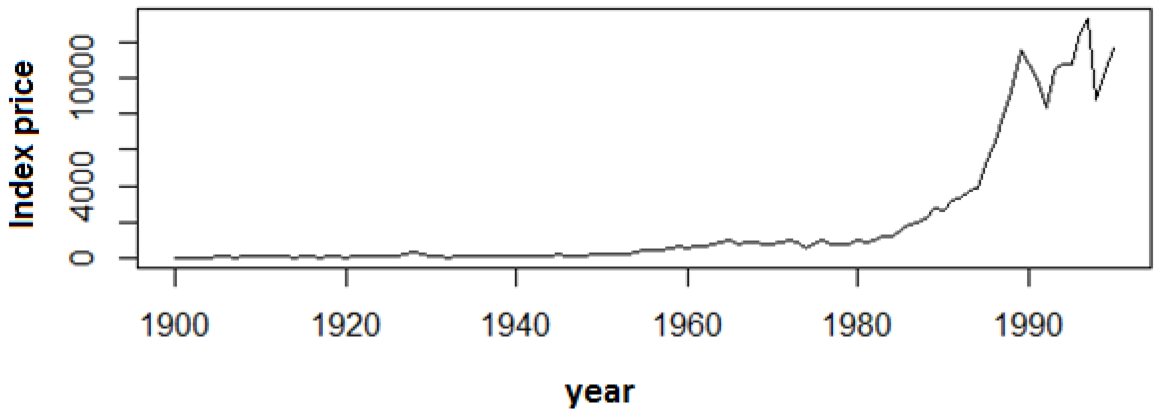
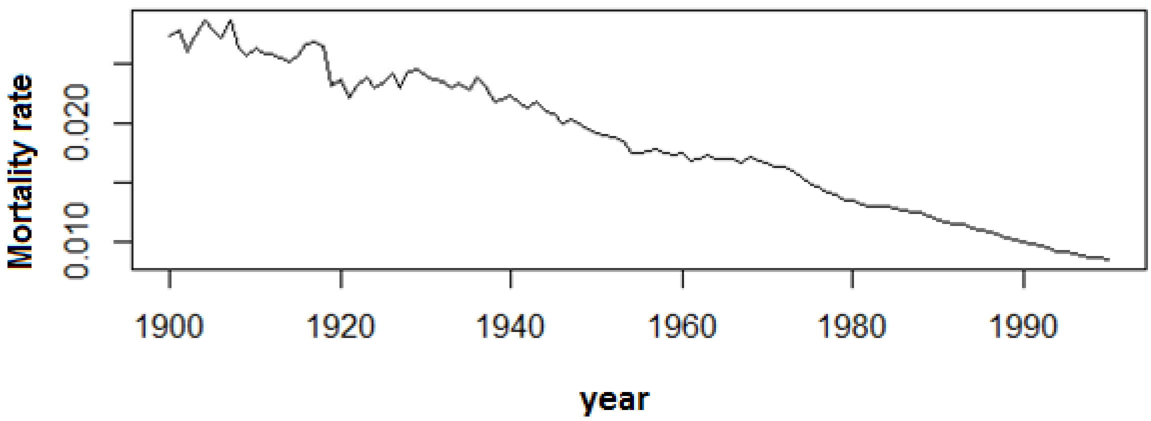


Figure 2.2: Annual US mortality rates of the age group "55 to 64 years", 1900-2010



age at 65 and may receive insurance, annuity or pension benefits which are linked to the performance of the investment portfolio.

2.2 ARIMA Models

The Autoregressive Integrated Moving Average (ARIMA) model is defined by (Asteriou and Hall, 2011)

$$z_t = \sum_{i=1}^{p} \phi_i z_{t-i} + e_t + \sum_{i=1}^{q} \theta_i e_{t-i} \tag{2.1}$$

where $\phi_1 ... \phi_p$ are the autoregressive (AR) parameters, $\theta_1 ... \theta_p$ are the moving-average (MA) parameters, and $e_t, e_{t-1}, \ldots$ are the white noise error terms with zero mean and variance σ_e. Furthermore, $z_t = Z_t$ for $d = 0$ and $z_t = Z_t - Z_{t-d}$ for $d = 1, 2, ...$ denotes the d^{th} difference of the original data series Z_t. Generally, the error terms are assumed to be independent and follow a normal distribution identically. As such, the estimated residuals $\hat{e}_t$ can be obtained from the model. In an ARIMA model, the conditional variance is constant while the conditional mean is not.

Box and Jenkins (1976) proposed a three phase iterative procedure for modelling an ARIMA time series:

1. Identification
2. Estimation
3. Diagnostic checks

Firstly, both autocorrelation function and partial autocorrelation function plots can be used to identify the potential AR and MA order of the model.

The final model may be selected by comparing the AIC (Akaike Information Criterion) (Akaike, 1974) or BIC (Bayesian Information Criterion) (Schwarz, 1978) values which are based on information theory:

$$AIC = 2k - 2\,ln(L) \tag{2.2}$$

$$BIC = k\,ln(n) - 2\,ln(L) \tag{2.3}$$

where n is the total number of observations, k is the number of parameters in the model, and L is the maximum value of the likelihood function for the model. Basically both AIC and BIC measure the relative quality of a model by taking its complexity into account. That is, increasing the number of parameters in the model will result in increase in the maximum likelihood value, but this will also lead to overfitting. By taking the penalty term for the number of parameters against the term for the maximum likelihood value, the model with the lowest AIC or BIC is preferred.

Once the orders are determined, then, ARIMA model can be fitted to time series data and AR as well as MA parameter estimates can be obtained and inferences on the parameter estimates can be examined.

Finally, some diagnostic tests should be conducted. For example, normality tests for model residuals should be performed. The Ljung–Box test (Box and Pierce, 1970), for which under the null hypothesis the autocorrelations of a time series are different from zero, can be used to determine if there exist serial correlations in the residuals. Furthermore, both ACF and PACF plots should be checked so that they do not display significant spikes at all lags, as a significant spike means a large autocorrelation. Also, checking whether there exists some cluster of volatility in the residuals can be done by plotting the residuals or squared residuals. If there is some serial correlation in the residuals, the residuals should also be modeled.

Since there is clearly an increasing trend for the index price as well as a decreasing trend for mortality rate as shown in Figures 2.1 and 2.2, they are not stationary time series and hence differences shall be taken. In addition, since index price and mortality rate will be forecasted by using time series model, they may go below 0 which is inappropriate. Therefore, the inputs Z_t of the time series model shall be Log index price and Log mortality rate.

In other words, ARIMA models with first difference will be fitted to both Log index and Log mortality, and the final models are determined by using the lowest BIC values. Table 2.1 shows the summary of the selected models for Log index and Log mortality of age group "55 to 64 years". For Log index, there is no AR or MA order; for Log mortality, the estimated parameters are all statistically significant.

After ARIMA models are determined, the residuals can then be obtained. However, it is also important to check if the residuals need to be modeled as well. Figures 2.3 and 2.4 display the residuals of Log index and Log mortality. For Log index, there are some significant spikes between the year 1920 and 1940. And for Log mortality, the spike of the year 1920 is the highest one.

There are at least two options to model the residuals of an ARIMA model. Firstly, the spikes on the squared residuals plot can be some cluster of volatility, and hence another time series model shall be fitted to the resid-

Table 2.1: Summary of the selected ARIMA models

Data series	ARIMA Model	BIC
Log index	(0,1,0)	-15.49
Log mortality	(1,1,2)	-460.47

Log mortality			
Parameter	Coefficient	Standard Error	Statistic
ϕ_1	0.9978	0.0045	237.57
θ_1	-1.2604	0.0988	-12.76
θ_2	0.287	0.0974	2.95

Figure 2.3: Residuals of Log index

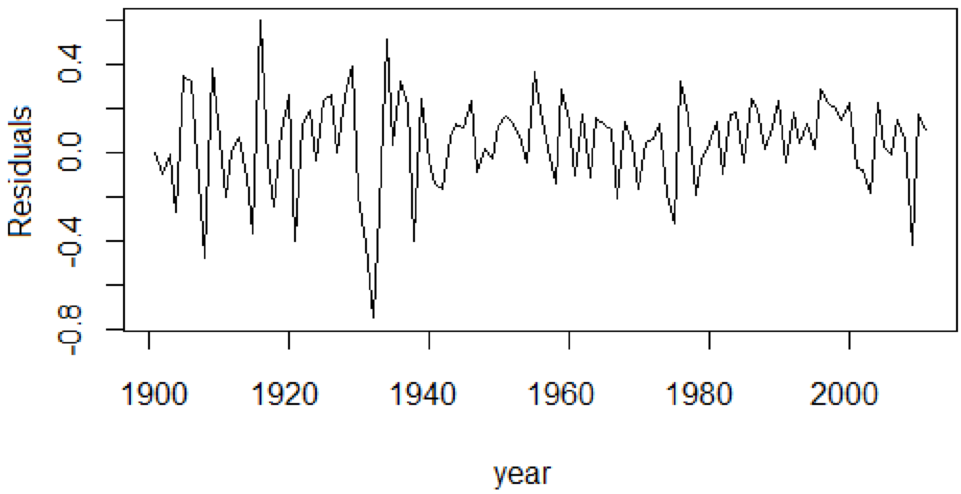
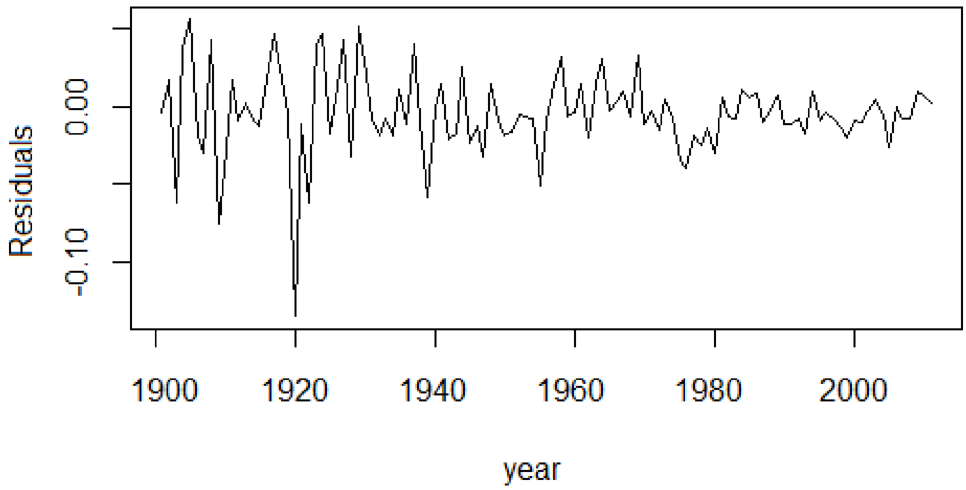


Figure 2.4: Residuals of Log mortality



uals. However, if the spikes on the squared residuals plot are not some cluster of volatility, then secondly, we may treat them as outliers effects in the data series.

2.3 GARCH Models

If there are some serial correlations in the residuals, the generalized autoregressive conditional heteroskedasticity (GARCH) model should be used in addition to ARIMA model. A GARCH model is defined by (Bollerslev, 1986)

$$\hat{e}_t = \sigma_t \epsilon_t \tag{2.4}$$

where $\sigma_t = \sqrt{\omega + \sum_{i=1}^{P} \alpha_i e_{t-i}^2 + \sum_{i=1}^{Q} \beta_i \sigma_{t-i}^2}$ and ϵ_t is a white noise process. When the error terms of an ARIMA model has GARCH effects, we can have a combination of ARIMA and GARCH models for which both the conditional mean and conditional variance are non-constant. That is, the conditional mean and conditional variance are not independent of time. As such, we can estimate the standardized residuals $\hat{\epsilon}_t$ and examine correlations analysis. Again, we should check whether the standardized residuals exhibit autocorrelation, as ϵ_t is assumed to be independent and identically distributed. This can be done by plotting the ACF of the standardized residuals and/or conducting Box-Ljung test (Box and Pierce, 1970) for which the autocorrelations are not different from zero under the null hypothesis.

Table 2.2 shows the selected ARIMA/GARCH models for both Log index and Log mortality. Again, the models are justified by using the lowest BIC values. For both data series, the orders of GARCH model are determined to be (1,1). Note that these models have a lower BIC value than that of the previous pure ARIMA models. Figures 2.5 and 2.6 display the standardized residuals of both data series. It turns out that the patterns of residuals and standardized residuals do not really differ. For Log mortality, the spike for the year 1920 remains to be very significant; for Log index, the result is slightly improved, but the spikes between 1900 and 1940 remains higher. Moreover, Ljung-Box tests (Box and Pierce, 1970) for residuals and standardized residuals of both models are also conducted. Since these p-values are significantly large, the null hypothesis that autocorrelation is 0 shall not be rejected. It turns out that the GARCH model does not address the issue we have on residuals. Since there is no serial correlation in the residuals, we may explain that the spikes on the squared residual plots are the outliers, and hence outlier models can be used.

Table 2.2: Summary of the selected ARIMA/GARCH models

Data series	ARIMA/GARCH Model	BIC
Log index	(0,1,0)/(1,1)	-25.01
Log mortality	(1,1,2)/(1,1)	-507.17

Log index			
Parameter	Coefficient	Standard Error	Statistic
ω	0.0178	0.01316	1.30
α_1	0.29585	0.15263	1.94
β_1	0.36275	0.34851	1.04

Log mortality			
Parameter	Coefficient	Standard Error	Statistic
ϕ_1	0.9978	0.0045	237.57
θ_1	-1.2604	0.0988	-12.76
θ_2	0.287	0.0974	2.95
ω	0.0178	0.01316	0.04
α_1	0.29585	0.15263	1.86
β_1	0.36275	0.34851	22.96

Ljung-Box Test p-value		
	Residuals	Standardized Residuals
Log index	0.8649	0.7358
Log mortality	0.9199	0.2986

Figure 2.5: Standardized Residuals of Log index

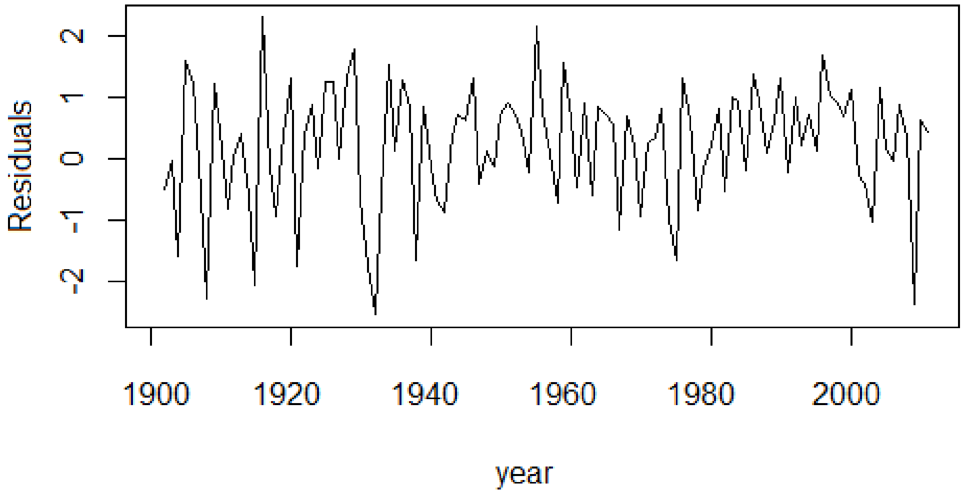
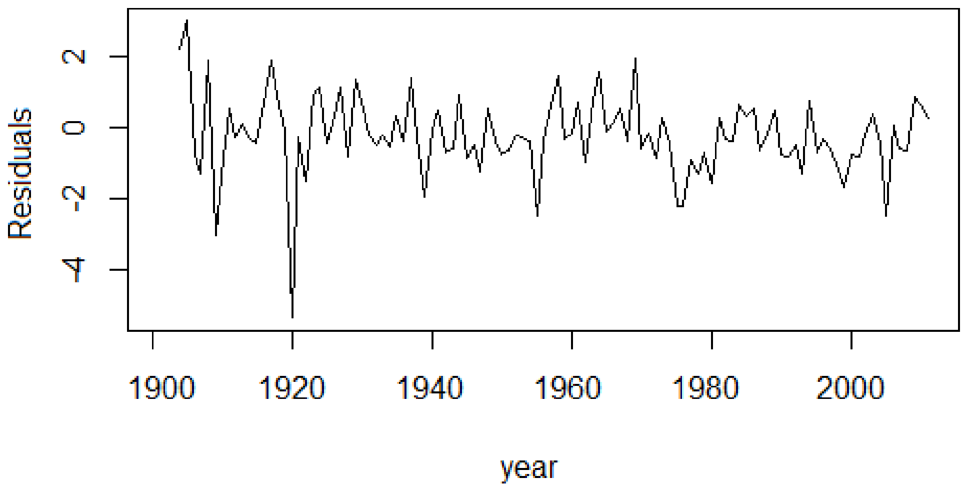


Figure 2.6: Standardized Residuals of Log mortality



2.4 Outlier Models

Based on the ARIMA time series model, Chen and Liu (1993) proposed an outlier model to address the issue of outlier effects. Following (2.1), an ARIMA model can be expressed in the following equation:

$$\frac{\phi(B)(1-B)^d}{\theta(B)} Z_t = \pi(B) Z_t = e_t \tag{2.5}$$

where $\phi(B)$, $\theta(B)$ and $\pi(B)$ are polynomials in B, and d represents the difference order. A general time-series outlier model is defined by

$$Y_t = Z_t + \sum_{m=1}^{M} \triangle_{t,i,m} \tag{2.6}$$

where Y_t represents Log index price or Log mortality rate, $Z_t = \frac{e_t}{\phi(B)}$ as defined in (2.3), M is the total number of outlier events, $\triangle_{t,i,m}$ is the outlier effects at time t, and i represents the type of the outliers. There are four types of outliers:

1. Additive outlier $\triangle_{t,AO} = \omega D_t^{(T)}$
2. Innovational outlier $\triangle_{t,IO} = \frac{\theta(B)}{\phi(B)(1-B)^d} \omega D_t^{(T)}$
3. Level Shift $\triangle_{t,LS} = \frac{1}{1-B} \omega D_t^{(T)}$
4. Temporary Change $\triangle_{t,TC} = \frac{1}{1-\delta B} \omega D_t^{(T)}$

where $D_t^{(T)} = 1$ for $t = T$ and 0 otherwise. An additive outlier (AO) only affects a single observation of a time series. An innovational outlier (IO) affects the whole time series structure, starting at a series point $t = T$. A level shift (LS) outlier has permanent effects on all observations, starting at a series point $t = T$. Similarly, a temporary change (TC) outlier has temporary effects on all observations, starting at a series point $t = T$. For the temporary change, the value of δ will be 0.7 as recommended by Chen and Liu (1993); however, it can also be specified depending on the time series analysis.

In order to fit a general time-series outlier model to the data, we start with the residuals. Similar to the residuals of an ARIMA model, the fitted residuals of an outlier model is given by

$$\hat{e}_t = \frac{\phi(B)(1-B)^d}{\theta(B)} Y_t \tag{2.7}$$

Since Y_t can be expressed in terms of the ARIMA model and outlier effects, the fitted residuals of an outlier model can be rewritten as

$$\hat{e}_t = \pi(B)(Z_t + \sum_{m=1}^{M} \triangle_{t,i,m}) = \varepsilon_t + \pi(B) \sum_{m=1}^{M} \triangle_{t,i,m} \tag{2.8}$$

Assume that there is only one outlier, the general time-series outlier model and the fitted residuals are given by

$$Y_t = Z_t + \triangle_{t,i,1} \tag{2.9}$$

$$\hat{e}_t = \varepsilon_t + \pi(B) \triangle_{t,i,1} \tag{2.10}$$

Therefore, assuming there is only one outlier, the fitted residuals of each of the four types of outliers is given by

1. $\hat{e}_{t,AO} = \varepsilon_t + \omega D_t^{(T)} \pi(B)$
2. $\hat{e}_{t,IO} = \varepsilon_t + \omega D_t^{(T)}$
3. $\hat{e}_{t,LS} = \varepsilon_t + \omega D_t^{(T)} \frac{\pi(B)}{1-B}$

4. $\hat{e}_{t,TC} = \varepsilon_t + \omega D_t^{(T)} \frac{\pi(B)}{1-\delta B}$

for $t = T, T+1, T+2, \ldots, n$. Note that, for example, for an additive outlier case, the outlier effects on the data series is $\omega D_t^{(T)}$, and this effects on the residuals is $\omega D_t^{(T)}$ multiplied by the polynomial $\pi(B)$. Hence, $\hat{e}_{t,AO} = \varepsilon_t + \omega D_t^{(T)} \pi(B)$ can be rewritten as $\hat{e}_{t,AO} = \varepsilon_t + X_t$, where $X_T = 1$ and $X_{T+k} = -\pi_k$. Here, note that $\hat{e}$ is actually a linear function of X_t. When an ARIMA time series model is fitted to the data, we can obtain both the polynomial $\pi(B)$ (and hence X_t) and the residuals. If X_t is the input and the residuals $\hat{e}$ is the output, then, the least squared estimation can be used to estimate the parameter ω (Chen and Liu, 1993):

$$\hat{\omega}_{t,AO} = \frac{\sum_{t=T}^{n} \hat{e}_{t,AO} X_t}{\sum_{t=T}^{n} (X_t)^2} \tag{2.11}$$

for $T = 1, 2, \ldots, n$. The estimation of the other three types of outlier effects are similar: Fit an ARIMA model to the data series, obtain the residuals and polynomials, and estimate the parameters $\hat{\omega}_{t,i}$ by using least squared estimation. As such, the standardized statistic of the corresponding outlier effects is given by

$$\hat{\tau}_{t,i} = \frac{\hat{\omega}_{t,i}}{\hat{\sigma}} \tag{2.12}$$

for $i = AO, IO, LS, TC$ and $T = 1, 2, \ldots, n$. There are various methods which can be used to estimate the residual standard deviation $\hat{\sigma}$. One of the methods is called median absolute deviation (MAD) and is given by (Hoaglin, Mosteller and Tukey, 1983)

$$\hat{\sigma} = \frac{1.483 * \text{median}\{|\hat{e}_t - \tilde{e}_t|\}}{\sqrt{\sum_{t=T}^{n} (X_t)^2}} \tag{2.13}$$

where $\tilde{e}_t$ is the median of the estimated residuals.

In addition, back to the initial fitted residuals of an outlier model $\hat{e}_t = \varepsilon_t + \pi(B)\sum_{m=1}^{M} \triangle_{t,i,m}$, similarly, this can also be treated as a multiple linear regression model and the parameters $\hat{\omega}_{T,i}$ and standardized statistic $\hat{\tau}_{T,i}$ of each outlier effect can estimated accordingly.

In order to fit an outlier model to a data series, Chen and Liu (1993) proposed a two stage joint estimation iterative procedure:

Stage I

1. For first iteration: Fit ARIMA model to original data series and obtain residuals.

   For other iterations: Fit ARIMA model to adjusted data series (data series after outlier effects are removed) and obtain residuals.

2. For $t = 1, 2 \ldots n$, compute $\hat{\tau}_{t,AO}$, $\hat{\tau}_{t,IO}$, $\hat{\tau}_{t,LS}$ and $\hat{\tau}_{t,TC}$.

   To find a possible outlier: $L = max_{t,i}\{\hat{\tau}_{t,i} > C\}$, where C is a defined critical value.

3. If an outlier is found, remove the outlier effect from the residuals, go to 2. to find another possible outlier. If not, immediately stop here.

4. If outliers are found under the current ARIMA model, remove the outlier effects from the data series. Go to 1. to revise ARIMA parameters. If not, immediately stop here and conclude that there is no outlier in this data series.

Stage II

1. Suppose M outliers are found: jointly estimate the outlier effects by using the equation $\hat{e}_t = \varepsilon_t + \pi(B) \sum_{m=1}^{M} \triangle_{t,i,m}$

2. Compute τ_m, for $m = 1 \ldots M$. If $min_{t,i}\{\hat{\tau}_m < C\}$, delete the outlier and go back to 1. If the minimum of the statistics is greater than C, then immediately stop here.

3. Obtain the adjusted data series by removing the remaining outlier effects from the original data series.

Based on the simulation results, Chen and Liu (1993) suggest that if the number of observations is between 100 and 200, the appropriate critical value C will be 3.0. If the number of observations is greater than 200, the critical value C should also be greater than 3.0.

Note that there are two types of residuals. Firstly, $\hat{e}_t$ is the residuals obtained from the ARIMA model fitted to the original data series. That is, $\hat{e}_t$ is the residuals inclusive of outlier effects. On the other hand, after outlier effects are determined, $\hat{\varepsilon}_t$ can be obtained from the ARIMA model fitted to the adjusted data series. Hence, $\hat{\varepsilon}_t$ is analogous to the standardized residuals in the ARIMA/GARCH model.

Li and Chan (2007) observed the presence of outliers or outlier effects in the mortality data series. Therefore, other than the traditional ARIMA time series model and the recent Lee-Carter model (Lee and Carter, 1992), mortality can also be studied by using the general time-series outlier model: Firstly, Lee-Carter model is fitted to the American and Canadian mortality data, and the time-varying parameter k_t is obtained. Then, a general time-series outlier model can be fitted to the k_t of both American and Canadian cases. However, Li and Chan noticed one problem arising from the outlier model and the estimation.

> *"Even if the model is correctly specified, outliers may lead to biases in the estimation of parameters hence affecting the detection of outliers and ultimately obscure the re-estimation of parameters, the whole process repeating itself indefinitely" – Li and Chan (2005)*

For the joint estimation proposed by Chen and Liu (1993), firstly, an ARIMA model is fitted to the original data series; the ARIMA parameters are estimated, and the residuals are obtained. If some outliers are detected (under the initial ARIMA model), then the outlier effects are removed from the original data series, and new ARIMA parameters are estimated by using the adjusted data series. In other words, this new ARIMA model has the same orders as the initial ARIMA model, but the estimated parameters are different.

One problem over here is that the orders of the new ARIMA model (for the adjusted data series) may not necessarily be the same as the orders of the initial ARIMA model (for the original data series). While an outlier model can address the issue of outlier effects, Li and Chan (2005) claim that the outlier model do not address the problem that the outliers can cause potential erroneous model selection. In order to overcome such situation, an external iteration cycle proposed by Li and Chan (2005) will be employed.

1. **Identify tentative ARIMA model**

   Fit an ARIMA model to the original data series and obtained the residuals.

2. **Detect outliers and make adjustments**

   Use Chen and Liu's (1993) joint estimation iterative procedure to detect the potential outliers in Stage I, and delete the outliers that are not significant in Stage II. Remove the remaining outlier effects from the original data series and obtain the adjusted data series.

3. **Re-identify the model**

   Identify the ARIMA model using the adjusted data series. The ARIMA orders for the adjusted data series may not be the same as the ARIMA orders of the original data series. If the ARIMA orders are different, repeat Step 2 by using the adjusted ARIMA orders and original data series. If the orders are the same, then immediately stop here, the ARIMA model will be the final model.

Based on Li and Chan's (2007) analysis, several outliers are found for the case of American mortality. Among the most significant is the 1918 additive outlier which has a magnitude of 24.473 and standardized statistic of 13.06. Other detected outliers generally have magnitude and standardized statistic lower than 10 and 5 respectively.

For our data series, the initial model for Log index and Log mortality are identified to be ARIMA(0,1,0) and ARIMA(1,1,2) respectively. Then, right after the first iteration, after the outlier effects have been removed from the original data series, the models for the adjusted data series are identified to ARIMA(0,1,0) and ARIMA(1,1,2) respectively, which are the same as the initial models. Therefore, the outlier models for Log index and Log mortality are ARIMA(0,1,0) plus outlier effects and ARIMA(1,1,2) plus outlier effects respectively. Summary of the selected outlier models as well as detected outliers for both data series are also shown in the Table 2.3. Note that the BIC of the outlier models are lower than those of the ARIMA and ARIMA/GARCH models. And therefore, it may be reasonable to favor outlier models in terms of model selections for the data.

For Log index, there is a total of three outliers which happen to be in 1907, 1931, and 1932. A temporary change in 1931 is followed by an additive outlier in 1931; This is no surprise as the Great Depression which swept the United States took place between 1929 and 1932. The minus sign of the coefficients explains that the Log index decreased due to the extraordinary event. Although these outliers are detected, however, the standardized statistics are not really large. For Log mortality of age group "55 to 64 years", there is a total of ten outliers. The level shift in 1919 is among the most significant one, with a standardized statistic of -8.71; it was around the time when the flu event occurred. Note that Li and Chan (2007) also detected a total of seven outliers by using Lee-Carter model k_t of American mortality as the input; five outliers happen to be the same in terms of years: 1916, 1921, 1928, 1936, and 1954.

Table 2.3: Summary of the selected outlier models

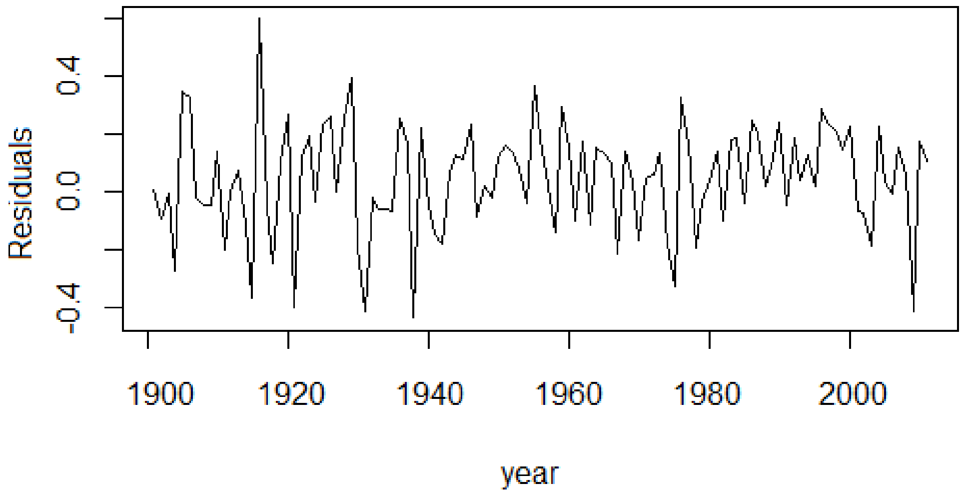
Data series	Outlier Model	BIC
Log index	(0,1,0) + outlier effects	-31.12
Log mortality	(1,1,2) + outlier effects	-514.2

Log index			
Parameter	Coefficient	Standard Error	Statistic
AO33	-0.4194	0.1369	-3.06
AO8	-0.4283	0.1367	-3.13
TC32	-0.7277	0.1785	-4.08

Log mortality			
Parameter	Coefficient	Standard Error	Statistic
ϕ_1	0.9322	0.0397	23.48
θ_1	-1.2575	0.0915	-13.74
θ_2	0.5645	0.0979	5.77
AO8	0.0654	0.0117	5.59
TC37	0.0507	0.0114	4.45
TC55	-0.0487	0.0113	-4.31
AO3	-0.0499	0.0124	-4.02
TC5	0.0578	0.0128	4.52
TC17	0.0446	0.0129	3.46
LS20	-0.1237	0.0142	-8.71
AO22	-0.0559	0.0115	-4.86
LS29	0.0526	0.0121	4.35
AO39	-0.0391	0.0113	-3.46

**AOK, LSK and TCK represent additive outlier, level shift and temporary change, respectively, of year 1989+K.

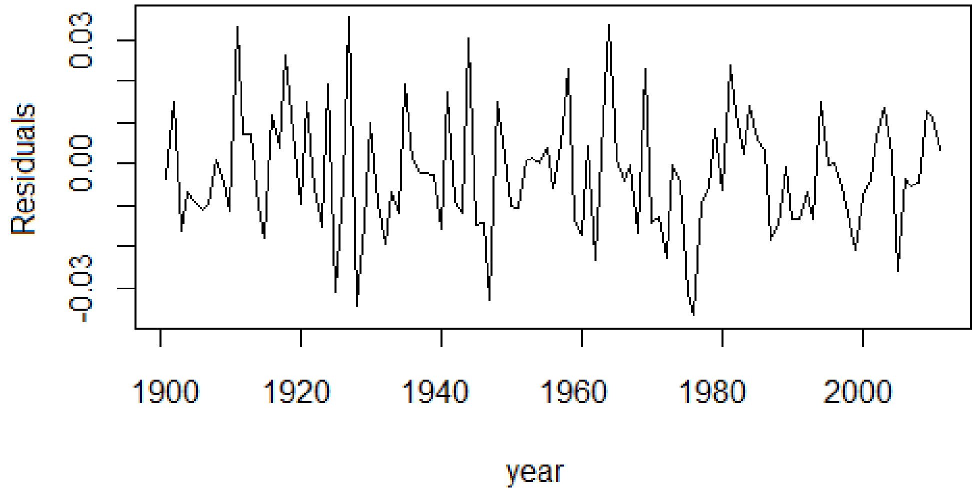
Figure 2.7: Residuals of Log index, outlier model



Figures 2.7 and 2.8 display the residuals of the outlier models fitted to both data series. It can be observed that the spikes on the residuals plots are removed. Therefore, based on these results, we conclude that the spikes on the residual plots are not some cluster of volatility; instead, they are some outliers effects. Hence, we will use outlier models for our data series Log index and Log mortality. Then, residuals can be obtained from outlier models for correlation analysis.

However, note that there are two types of residuals for an outlier model. An outlier model is a combination of an ARIMA model and outliers effects. One can estimate the outlier magnitudes by using the joint estimation procedure on the equation of fitted residuals $\hat{e}_t = \varepsilon_t + \pi(B)\sum_{m=1}^{M} \triangle_{t,i,m}$. Note that in this case, ε_t will not be estimated along with the magnitude of outliers (although it could be). Rather, $\hat{\varepsilon}_t$ will be obtained by using

Figure 2.8: Residuals of Log mortality, outler model

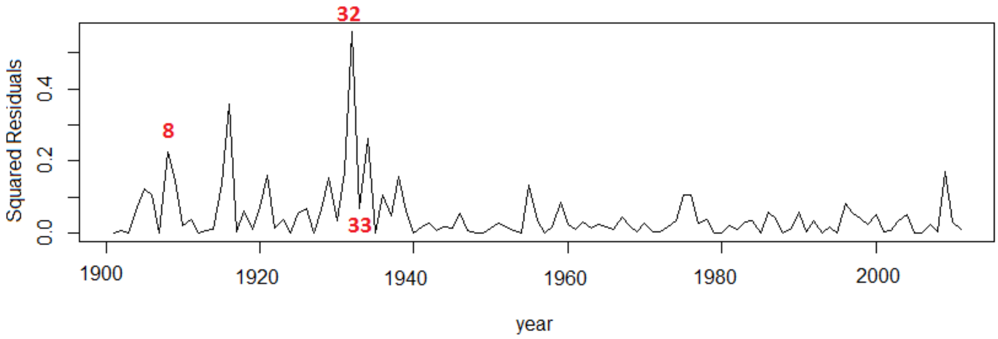


the adjusted data series. In other words, subject to the external iteration cycle (Li and Chan, 2005), $\hat{e}_t$ is the residuals obtained from the ARIMA model of the very first iteration in the first stage of the joint estimation procedure fitted to the original data series; on the other hand, $\hat{\varepsilon}_t$ is the residuals obtained from the ARIMA model fitted to the adjusted data series (after outliers are removed).

Therefore, in order to capture the tail dependence between the two data series, instead of $\hat{\varepsilon}_t$, $\hat{e}_t$ should be used as the inputs as it contains outlier effects in the outlier model. For the rest of this paper, we denote X the residuals inclusive of outlier effects of Log mortality, and Y the residuals inclusive of outlier effects of Log index.

It is worth to also note the differences between an ARIMA/GARCH model and a general time-series outlier model. Firstly, after an ARIMA model

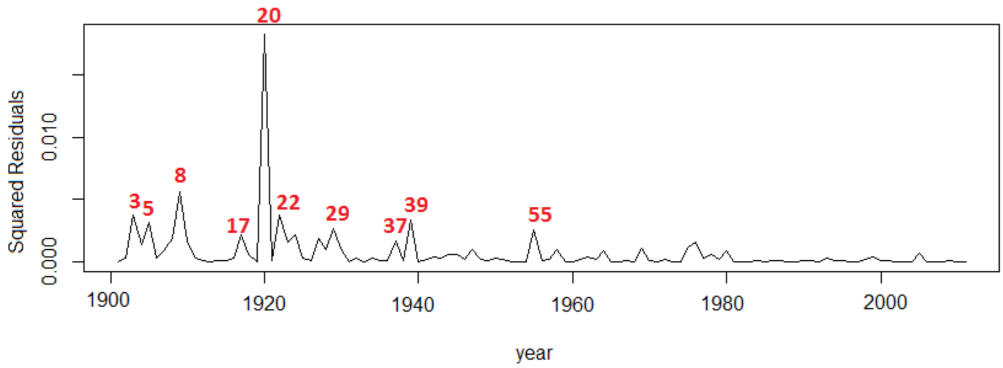
Figure 2.9: Squared residuals of Log index with outlier effects labeled



is fitted and residuals or squared residuals are obtained, while a GARCH model explains that the spikes on the squared residual plot are some cluster of volatility, an outlier model treats the spikes as outliers or outlier effects. Our results showed that GARCH model is not able to deal with these spikes; however, they are captured as outlier effects by outlier model. This can be illustrated on Figures 2.9 and 2.10 where the locations of outliers are labeled on the squared residual plots. Especially in the squared residual plot of Log mortality, almost every significant spike matches the detected outliers of the outlier model.

Secondly, although both models consist of an ARIMA model, the ARIMA orders will be different if two or more iterations of external iteration cycle are required. Also, even if the ARIMA orders of both models are the same, the estimated parameters will be different as both models seek to explain the residuals in two different ways.

Figure 2.10: Squared residuals of Log mortality with outlier effects labeled



Chapter 3

Correlations and Copulas

This chapter reviews linear and rank correlations as well as empirical and parametric tail correlations. Parametric tail correlations can be calculated by fitting copulas to our data series. Goodness of fit tests as well as inferences on tail correlations will be conducted. Our analysis show that Gumbel copula is preferred for our data series.

3.1 Correlations

Correlations measure the strength of the relationship between two variables or two data sets. The Pearson correlation coefficient is defined as

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} \tag{3.1}$$

Rank correlation Kendall's Tau is a measure of concordance for bivariate random vectors; for $(\tilde{X},\tilde{Y})$ an independent copy of (X,Y), Kendall's Tau is given by (Kendall, 1938)

$$\rho_\tau(X,Y) = E(sign((X-\tilde{X})(Y-\tilde{Y}))) \tag{3.2}$$

Note that for $i \neq j$, any pair of observations (x_i, y_i) and (x_j, y_j) are said to be concordant if both $x_i > x_j$ and $y_i > y_j$ or if both $x_i < x_j$ and $y_i < y_j$. Otherwise, they are discordant. The other rank correlation Spearman's Rho is simply the correlations between the cumulative function of the two variables (Spearman, 1904).

$$\rho_S(X, Y) = \rho(F(X), F(Y)) \tag{3.3}$$

Correlations can be used to explain the relationship between residuals of two models. If there exists a significant correlation between the residuals of two models or variables, we may expect that a sudden change in one variable will be accompanied by a change in the other.

For a sample, linear correlation can be calculated by (Rummel, n.d.)

$$r = \frac{\sum x_i y_i - n\bar{X}\bar{Y}}{(n-1)s_X s_Y} \tag{3.4}$$

and the test statistic for linear correlation is given by (Rummel, n.d.)

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{3.5}$$

and follows a student-t distribution with $n - 2$ degrees of freedom. Sample Kendall's tau can be calculated by (Abdi, 2007)

$$r_\tau = \frac{\text{number of concordant pairs-number of discordant pairs}}{n(n-1)0.5} \tag{3.6}$$

and the test statistic for Kendall's Tau is given by (Abdi, 2007)

$$z = 3\tau\sqrt{\frac{n(n-1)0.5}{2n+5}} \tag{3.7}$$

and follows a standard normal distribution. Finally, sample Spearman's Rho can be calculated by (Daniel, 1990)

$$r_S = \frac{\sum u_i v_i - n\bar{U}\bar{V}}{(n-1)s_U s_V} \tag{3.8}$$

where $U = F(X)$ and $V = F(Y)$, and the test statistic for Spearman's Rho is given by (Daniel, 1990)

$$t = r\sqrt{\frac{n-2}{1-r^2}} \tag{3.9}$$

and follows a Student-t distribution with $n-2$ degrees of freedom, which is the same as the test statistic for linear correlation.

Tail correlations shall also be examined in case that linear and rank correlations do not exist, or if they do they are not significant. Tail correlations can be empirical or parametric. For an empirical case, suppose X and Y are two random variables with empirical CDF F_X and F_Y, then if $x_q = F_X^{-1}(q)$ is the $100q^{th}$ percentile of X and $y_q = F_Y^{-1}(q)$ is the $100q^{th}$percentile of Y, the upper and lower tail correlations are given by (McNeil et al., 2015)

$$\lambda_u = Pr(Y > y_q | X > x_q) = \frac{Pr(Y > y_q, X > x_q)}{Pr(X > x_q)} \tag{3.10}$$

$$\lambda_l = Pr(Y < y_q | X < x_q) = \frac{Pr(Y < y_q, X < x_q)}{Pr(X < x_q)} \tag{3.11}$$

Note that $Pr(X < x_q) = Pr(Y < y_q) = q$. In addition, tail correlation is always between 0 and 1. For upper tail correlation, q should be large enough (close to 1) and for lower tail correlation, it should be small (close to 0).

As mentioned in the last chapter, we denote by X the residuals inclusive of the outlier effects of Log mortality, and Y the residuals inclusive of the outlier effects of Log index. Figures 3.1 and 3.2 show the plots of X against Y as well as empirical CDF of X against empirical CDF of Y. While the existence of linear and rank correlations are ambiguous as shown in the plots, there is clearly an upper tail dependence.

Figure 3.1: Plot of X against Y

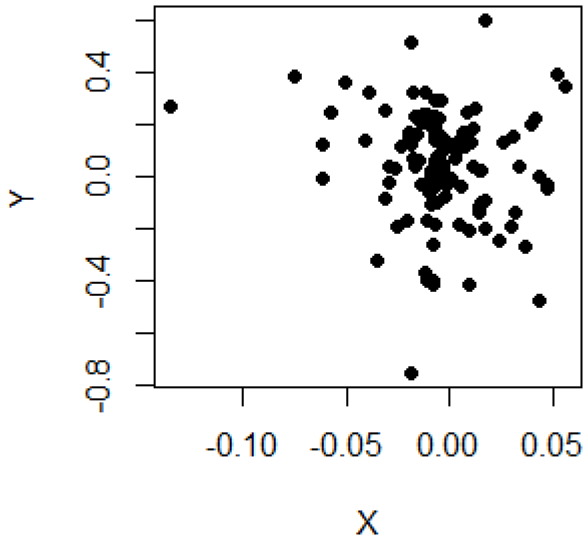


Figure 3.2: Plot of $F(X)$ against $F(Y)$

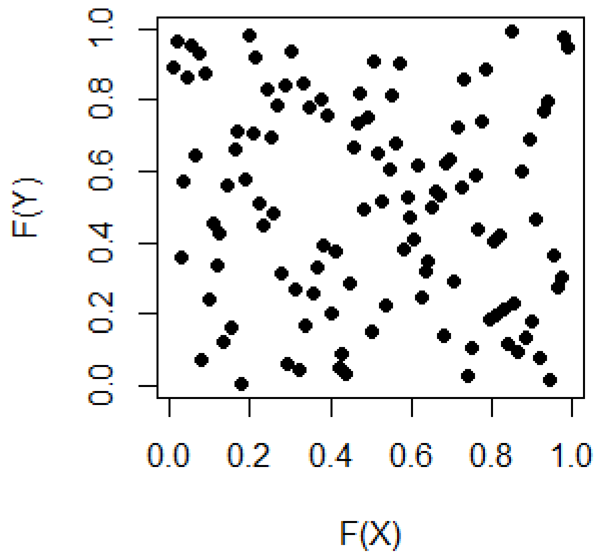


Table 3.1: Summary of linear and rank correlations between X and Y

	Correlation	Correlation Inference p-value
Pearson coefficient	-0.152	0.112
Kendall's Tau	-0.116	0.072
Spearman's Rho	-0.163	0.086

Table 3.1 summarizes the linear and rank correlations between X and Y. These correlation values are all between -0.1 and -0.17, implying some degree of dependence between the two data series. However, although p-values for the inferences are small, they are not small enough to be statistically significant. That is, we cannot reject the null hypothesis that the correlations are not different from 0.

Table 3.2 also shows the empirical tail correlations between $-X$ and Y. The results seem to conform to the empirical CDF plot of X against Y. As q becomes larger and close to 1, upper tail correlation increases. However, at the lower tail side, when q is smaller and close to 0, the tail correlation goes to 0. This result implies that a very low mortality may be accompanied by a very high index price.

Table 3.2: Summary of empirical tail correlations between $-X$ and Y

Lower			
q	0.05	0.1	0.15
Tail Correlation	0	0.091	0.125

Upper			
q	0.85	0.9	0.95
Tail Correlation	0.125	0.182	0.2

3.2 Copulas

Tail correlations (or tail coefficients) can be parametric. In this case, tail coefficients can be derived by using copula. According to Sklar (1959) theorem, every multivariate cumulative distribution function can be written in the form of copula C. For a bivariate case, for $F_X(x) = u$ and $F_Y(y) = v$, we have

$$C(u, v) = F(x, y) \tag{3.12}$$

One useful fact is that the marginal distributions $F_X(x)$ and $F_Y(y)$ can be determined independently of the copula distribution. Some popular classes of copula include Gaussian, student-t, Gumbel, Clayton, and Joe. Both Gaussian and student-t are elliptical copulas; they are also known as implicit copulas as they do not have a simple closed form (McNeil et al., 2015). Since we are only interested in observing the dependence between two variables, we only consider the bivariate case. Bivariate Gaussian and student-t copulas are given by (McNeil et al., 2015)

$$C_{\rho}^{Gaussian}(u,v) = \Phi(\phi^{-1}(u), \phi^{-1}(v)) \tag{3.13}$$

$$C_{\rho,\nu}^{t}(u,v) = \boldsymbol{t}_{\nu}(t^{-1}(u), t^{-1}(v)) \tag{3.14}$$

where ρ and ν are the parameters, Φ joint standard normal cumulative distribution function, ϕ standard normal cumulative distribution function, $\boldsymbol{t}_{\nu}$ standard student-t cumulative distribution function with ν degrees of freedom, and t standard student-t cumulative distribution function.

Archimedean copulas include Gumbel and Clayton. These are all explicit copulas since they have a simple closed form. All bivariate Archimedean copulas are defined by (McNeil et al., 2015)

$$C(u,v) = \varphi^{-1}(\varphi(u), \varphi(v)) \tag{3.15}$$

where φ^{-1} is the Archimedean generator and it is a mapping from $[0,1]$ to $[0,\infty]$. More specifically, Gumbel and Clayton copulas are given by (McNeil et al., 2015)

$$C_{\theta}^{Gu}(u,v) = e^{-((-lnu)^{\theta}+(-lnv)^{\theta})^{1/\theta}}, \ \ 1 \leq \theta < \infty \tag{3.16}$$

$$C_{\theta}^{Cl}(u,v) = (u^{\theta} + v^{\theta} - 1)^{-1/\theta}, \ \ 0 < \theta < \infty \tag{3.17}$$

where θ is the copula parameter. Since $C(u,v) = F(x,y)$ for $F(x) = u$ and $F(y) = v$, the upper and lower tail coefficients can be derived by applying limits on the copulas (McNeil et al., 2015):

$$\lambda_u = \lim_{q \to 1^-} Pr(Y > F_Y^{-1}(q) | X > F_X^{-1}(q)) = \lim_{q \to 1^-} \frac{1 - 2q + C(q,q)}{1-q} \tag{3.18}$$

$$\lambda_l = \lim_{q \to 0^+} Pr(Y < F_Y^{-1}(q) | X < F_X^{-1}(q)) = \lim_{q \to 0^+} \frac{C(q,q)}{q} \tag{3.19}$$

In order to obtain the tail coefficients, we need to firstly fit different a copula to the data, and maximum likelihood estimation can be used to estimate the copula parameters. Once the parameter estimates are obtained, we can proceed to tail coefficients calculations by using equations (3.12) and (3.13).

Another class of copula is called extreme value copula. Just like the Archimedean copulas, all extreme value copulas are of the same form. Specifically, bivariate extreme value copulas take the form (McNeil et al., 2005)

$$C(u,v) = exp((ln(u) + ln(v))A(\frac{ln(u)}{ln(u) + ln(v)})) \tag{3.20}$$

where $A(w) = \int_0^1 max((1-x)w, x(1-w))dH(x)$ is the Pickand's dependence function, $H(x) = \lim_{n \to \infty} F^n(C_n x + D_n)$, and C_n and D_n are the normalizing constants.

Gumbel copula is both Archimedean and extreme value copula as it can be expressed in these two different forms. The dependence function $A(w)$ of Gumbel, Galambos, and HuslerReiss are given by (McNeil et al., 2015)

$$A^{Gu}(w) = (w^\theta + (1-w)^\theta)^{1/\theta} \tag{3.21}$$

$$A^{Ga}(w) = 1 - (w^{-\theta} + (1-w)^{-\theta})^{-1/\theta} \tag{3.22}$$

$$A^{HR}(w) = (1-w)\Phi(\frac{1}{\theta} + \frac{\theta}{2} ln\frac{1-w}{w}) + w\Phi(\frac{1}{\theta} + \frac{\theta}{2} ln\frac{w}{1-w}) \tag{3.23}$$

From the dependence functions of the copula families, we can derive the upper and lower tail coefficients for a given extreme value copula by using

Table 3.3: Summary of linear and rank correlations between X and Y

Copula	Lower Tail	Upper Tail
Gaussian	0	0
Student-t	$2t_{\nu+1}(-\frac{\sqrt{\nu+1}\sqrt{1-\rho}}{\sqrt{1+\rho}})$	$2t_{\nu+1}(-\frac{\sqrt{\nu+1}\sqrt{1-\rho}}{\sqrt{1+\rho}})$
Gumbel	0	$2-2^{1/\theta}$
Clayton	$2^{-1/\theta}$	0
Galambos	0	$2^{-1/\theta}$
Husler Reiss	0	$2(1-\Phi(\frac{1}{\theta}))$

equations (3.12) and (3.13) (McNeil et al., 2015):

$$\lambda_u = \lim_{q\to 1^-} \frac{1-2q+C(q,q)}{1-q} = 2(1-A(0.5)) \quad (3.24)$$

$$\lambda_l = \lim_{q\to 0^+} \frac{C(q,q)}{q} = 0 \quad (3.25)$$

Therefore, the lower tail coefficient of an Extreme Value copula is always equal to 0; the upper tail coefficient of Gumbel, Galambos and HuslerReiss copulas can be calculated by using the equation (3.24).

Tale 3.3 shows the derived lower and upper tail coefficients for copulas. For the Gaussian copula, both upper and lower tail are always equal to 0; and for stuent-t copula, both upper and lower tail are always equal. Moreoever, Gumbel copula has an upper tail, and Clayton has a lower tail.

Note that for the Gumbel copula, the upper tail coefficient equals $2-2^{1/\theta}$, which is the same as the one in the Archimedean form. In order to calculate the upper tail coefficients, again, the parameters for each extreme value copula are to be estimated, and this can be done by using maximum log-likelihood estimation. Then, by using the equations above, the upper tail coefficients can be obtained.

Moreover, one can simply replace u and/or v by $1-u$ and/or $1-v$ of the equation of a copula to obtain a rotated copula:

1. For a 90^o rotated copula, replace u by $1-u$
2. For a 180^o rotated copula, replace u and v by $1-u$ and $1-v$ respectively
3. For a 270^o rotated copula, replace v by $1-v$

The reason rotated copulas are used is that some ordinary copulas that we discussed earlier can not be fitted to negatively dependent bivariate data (McNeil et al., 2015). By rotating the copula or reverse the order of data, tail dependence for negatively dependent data can hence be captured.

3.3 Fitting Copulas

Before a copula is fitted, the marginal distributions of residuals should be defined. We may choose empirical marginals for the residuals. The cumulative empirical distribution is given by

$$\hat{F}(e) = \frac{1}{N+1} \sum_{t=1}^{N} 1_{\{\hat{e}_t \leq e\}} \tag{3.26}$$

where N is the total number of observations and $1_{\{\}}$ is the indicator function.

One standard estimation method, maximum likelihood estimation will be used to fit the copulas (Patton, 2012). The bivariate copula density function is given by

$$c(u,v) = \frac{\partial^2 C(u,v)}{\partial u \partial v} \tag{3.27}$$

Then, if $\hat{U}$ and $\hat{V}$ are the empirical marginals, in order to obtain the estimated values of the parameters Θ, we have the log-likelihood function that shall be maximized (Patton, 2012):

$$L(\Theta; \hat{U}, \hat{U}) = \sum_t ln(c(\hat{u}_t, \hat{v}_t; \Theta)) \tag{3.28}$$

where $\Theta = (\theta_1, \theta_2, ...)$ is the parameter vector.

A semi-parametric copula-based model is composed of empirical marginal distributions for the residuals and parametric model for the copula (Patton, 2012). For Archimedean or Elliptical copulas, parametric bootstrap procedure (Remillard, 2010) may be used to estimate the p-value of the goodness of fit test. Cramer-von Mises test will be used for goodness of fit test of copula models, and the test statistic is defined by

$$\hat{G} = \sum_{i=1}^{T} (C(u_i, v_i; \hat{\Theta}) - C_{EMP}(u_i, v_i))^2 \tag{3.29}$$

where $C(u_i, v_i; \hat{\Theta})$ is the parametric copula under the null hypothesis and $C_{EMP}(u_i, v_i) = \frac{1}{T}\sum_{i=1}^{T} 1(u_i \leq u, v_i \leq v)$ is the empirical copula. The bootstrap procedure is as follows:

Algorithm 3.1. Parametric bootstrap procedure (Remillard, 2010)

(i) Obtain empirical CDF of the actual data (X, Y), and then use the empirical CDF (U, V) to obtain the the parameter estimates of a specified copula

(ii) Compute the Cramer-von Mises test statistic $\hat{G}$

(iii) Repeat the following a total of n times $(k = 1, 2, \ldots, n)$:

   (a) Generate a bivariate random uniform variables from the specified copula with the estimated parameters

   (b) Compute empirical CDF of the random uniform variables

   (c) Use the empirical CDF of the random uniform variables to obtained parameter estimates of the specified copula.

   (d) Compute the Cramer-von Mises test statistic $\hat{G}_k$

(iv) Compute the approximated p-value $= \frac{1}{n}\sum_{k=1}^{n} 1_{\{\hat{G} \leq \hat{G}_k\}}$

The idea of the Parametric bootstrap procedure is to calculate the test statistic, or the discrepancy between the empirical copula and the underlying copula, and then simulate data from the underlying copula and calculate the discrepancy between the empirical copula and the underlying copula. The p-value measures how compatible the data is with the underlying copula.

For extreme value copulas, goodness of fit test p-values can also be obtained by using parametric bootstrap procedure (Genest et al., 2011). In this procedure, Cramer-von Mises test based on the dependence function of each copula, for which under the null hypothesis the dependence function of the data is of a specified copula family, will be used. The test statistic is given by

$$\hat{H} = \sum_{i=1}^{T} (A(w; \hat{\Theta}) - A_{EMP}(w))^2 \tag{3.30}$$

where $A_{EMP}(w) = (\frac{1}{T}\sum_{i=1}^{T} min(\frac{-ln(\hat{u}_i)}{1-w}, \frac{-ln(\hat{v}_i)}{w}))^{-1}$ is the non-parametric estimator and $A(w;\hat{\Theta})$ is the parametric estimator by using maximum likelihood method. The parametric bootstrap procedure for extreme value copula goodness of fit test is as follows:

Algorithm 3.2. Parametric bootstrap procedure for extreme value copula (Genest et al., 2011)

(i) Obtain empirical CDF of the actual data (X, Y), and then use the empirical CDF (U, V) to obtain the parameter estimates of a specified copula

(ii) Compute the Cramer-von Mises test statistic $\hat{H}$

(iii) Repeat the following a total of n times $(k = 1, 2, \ldots, n)$:

  (a) Generate a bivariate random uniform variables from the specified copula with the estimated parameters

  (b) Compute empirical CDF of the random uniform variables

  (c) Use the empirical CDF of the random uniform variables to obtained parameter estimates of the specified copula.

  (d) Compute the Cramer-von Mises test statistic $\hat{H}_k$

(iv) Compute the approximated p-value $= \frac{1}{n}\sum_{k=1}^{n} 1_{\{\hat{H} \leq \hat{H}_k\}}$

The parametric bootstrap procedure is analogous to the Parametric bootstrap procedure for the Elliptical and Archimedean copulas as discussed earlier, except that the Cramer-von Mises test statistic here is applied to the dependence function of each extreme value copula

For a semi-parametric model, inference on the parameter estimate was studied by Genest, Ghoudi and Rivest (1995). The asymptotic distribution of the parameter estimates is given by

$$\sqrt{T}(\hat{\Theta} - \Theta*) \longrightarrow^{a} N(0, V*_{SPML}) \tag{3.31}$$

where

$$V*_{SPML} = A_{CF}^{-1}\Sigma_{CF}A_{CF}^{-1} \tag{3.32}$$

is the asymptotic covariance matrix. According to Chen and Fan (2006), A_{CF} and Σ_{CF} can be estimated by

$$\hat{A}_{CF,T} \equiv -\frac{1}{T}\sum_{t=1}^{T}\frac{\partial^2 ln(c(\hat{u}_t, \hat{v}_t); \hat{\Theta})}{\partial\Theta\partial\Theta'} \tag{3.33}$$

$$\hat{\Sigma}_{CF} = \frac{1}{T}\sum_{t=1}^{T}\mathbf{s}_t\mathbf{s}_t' \tag{3.34}$$

where

$$\mathbf{s}_t\mathbf{s}_t' \equiv \frac{\partial}{\partial\Theta}ln(c(\hat{u}_t, \hat{v}_t); \hat{\Theta}) + \hat{P}_t + \hat{Q}_t \tag{3.35}$$

$$\hat{P}_t \equiv \frac{1}{T}\sum_{s=1, s\neq t}^{T}\frac{\partial^2 ln(c(\hat{u}_s, \hat{v}_s); \hat{\Theta})}{\partial\Theta\partial u}(1\{\hat{u}_t \leq \hat{u}_s\} - \hat{u}_s) \tag{3.36}$$

$$\hat{Q}_t \equiv \frac{1}{T}\sum_{s=1, s\neq t}^{T}\frac{\partial^2 ln(c(\hat{u}_s, \hat{v}_s); \hat{\Theta})}{\partial\Theta\partial v}(1\{\hat{v}_t \leq \hat{v}_s\} - \hat{v}_s) \tag{3.37}$$

for bivariate copula.

For inferences on the tail coefficients, since each of the above tail coefficient is a function of the parameter estimate(s), the Delta method (Oehlert, 1992) can be employed to calculate the standard error. If the tail coefficient is a function of one parameter estimate, as for many of the copulas mentioned above, we denote $g(\hat{\theta})$ the tail coefficient which is a function of

$\hat{\theta}$; then, the standard error of $g(\hat{\theta})$ will be $se(g(\hat{\theta})) = se(\hat{\theta})\frac{\partial g(\hat{\theta})}{\partial \hat{\theta}}$. Finally, the statistic is $z = \frac{g(\hat{\theta})}{se(g(\hat{\theta}))} \sim^a N(0, 1)$.

For our data series, empirical distributions will be used as the marginals. We denote $\hat{U}$ and $\hat{V}$ the empirical CDF of X and Y respectively. We showed that the linear and rank correlations between X and Y are negative. Hence, we shall used some 90^o and/or 270^o rotated copulas for our data series. More specifically, Gaussian, Student-t, 90^o rotated Gumbel, 90^o rotated Clayton, 270^o rotated Gumbel, 270^o rotated Clayton, 90^o rotated Galambos and 90^o rotated Husler Reiss copulas will be fitted to the empirical CDF of X and Y, by using maximum likelihood estimation.

Note that since empirical marginals are used, the model will be semi-parametric. Goodness of fit tests will be examined by using parametric bootstrap procedures (Remillard, 2010) (Genest et al., 2011). For Elliptical and Archimedean copulas, the test statistics will be based on the copula functions; for extreme value copulas, the dependence functions will be used to conduct the tests. In addition, delta method (Oehlert, 1992) will be used to calculate the tail correlation inference statistics and hence the corresponding p-values.

Table 3.4 shows the summary of the copulas fitted to $\hat{U}$ and $\hat{V}$ as well as the tail coefficients calculated by using the MLE of copula parameters. First, Student-t has the highest log-likelihood value, and is followed by 90^o rotated Gumbel copula. For the copulas that cannot capture the upper tail coefficient, their log-likelihood values are pretty small.

Second, for the goodness of fit tests by using parametric bootstrap pro-

Table 3.4: Summary of copula tail coefficients between X and Y

	Copula Class	Log-likelihood value	GoF p-value	Lower Tail Coefficient	Upper Tail Coefficient
Elliptical*	Gaussian	1.131	0.302	0	0
	Student-t	4.818	0.8861	0.179	0.179
Archimedean*	90^o R.Gumbel	2.749	0.9286	0	0.180
	90^o R.Clayton	0.443	0.1080	0.005	0
	270^o R.Gumbel	1.362	0.5050	0.144	0
	270^o R.Clayton	2.505	0.8056	0	0.094
Extreme Value**	90^o R.Gumbel	2.749	0.513	0	0.180
	90^o R.Galambos	2.219	0.493	0	0.165
	90^o R.HuelerReiss	0.967	0.4321	0	0.150

**For these copulas, GoF tests are based on Algorithm 2.1*
***For these copulas, GoF tests are based on Algorithm 2.2*

cedure (Remillard, 2010) (Genest et al., 2011), the highest p-value goes to 90^o rotated Gumbel copula; this large p-value indicates that we have very little evidence against the null hypothesis, therefore Gumbel copula shall not be rejected for our analysis. In addition, Student-t copula also has a large p-value. Overall, most copulas that can capture an upper tail dependence have a very significant goodness of fit test p-value.

Since student-t and 90^o rotated Gumbel copulas yield the larger log-likelihood values and p-values, they are chosen for our analysis. However, our empirical analysis showed that lower tail correlation is 0 when q is low enough. Hence, we will only use 90^o rotated Gumbel copula for our forecasting analysis.

Table 3.5 shows the parameter estimate and upper tail coefficient as well as the corresponding standard errors and p-values. Both p-values are significantly small, this leads to rejection of both the null hypothesis that

Table 3.5: Summary of fitted 90^o rotated Gumbel copula

	Parameter	Upper Tail Coefficient
Estimate	1.1577	0.18
Standard error	0.07306	0.07498
p-value	0.00001	0.00878

the parameter estimate and upper tail coefficient are 0.

Overall, compared to linear and rank correlations, tail coefficients of the copulas that pass the goodness of fit test are generally larger and significant at the 5% level. When there is no obvious linear or rank dependence, it is reasonable to seek for some tail dependence to explain the data for pricing or forecasting purposes.

Generally, 90^o rotated Gumbel copula beats all of the other copulas in terms of log-likelihood value and goodness of fit test. What is important is this copula also yields the highest tail coefficient 0.18 and it is significant at 0.05. Therefore, it is reasonable to choose Gumbel copula for our data series. Our results seem to imply that a very low mortality rate may be accompanied by a very high index price.

Chapter 4

Modified Gumbel Copula

This chapter reviews Gumbel copula from the Archimedean family and introduces a modified copula which is based on the Archimedean generator for Gumbel copula. The advantages and significances of using the modified Gumbel copula will also be discussed. Finally, this newly developed copula will also be fitted to our data series $\hat{U}$ and $\hat{V}$ and tail coefficients will be derived and calculated.

4.1 Motivation

Gumbel copula has been chosen for our forecasting analysis. However, for this copula, there is still room for improvement. In order to understand the modified Gumbel copula, both Gumbel copula and Archimedean generators shall first be understood. As noted in Chapter 3, all bivariate Archimedean copulas are in the form of $C(u, v) = \varphi^{-1}(\varphi(u), \varphi(v))$, where $\varphi(t)$ is the Archimedean generator mapping from $[0, 1]$ to $[0, \infty]$. A strict Archimedean generator satisfies the following characteristics:

1. $\varphi(1) = 0$
2. $\varphi(0) = \infty$
3. $\varphi(t)$ is a continuously decreasing function
4. $\varphi(t)$ is a convex function

Archimedean generator provides a useful platform for data simulation. Suppose for a bivariate Archimedean copula with generator $\varphi(t)$, the simulation procedure is as follows (Wu, Valdez, & Sherris, 2007):

Algorithm 4.1. Bivariate Archimedean copula simulation procedure

(i) Simulate two random numbers $r \sim U(0,1)$ and $s \sim U(0,1)$

(ii) Compute $t = k^{-1}(r)$ where $k(t) = t - \frac{\varphi(t)}{\varphi'(t)}$

(iii) Compute $u = \varphi^{-1}(s\,\varphi(t))$ and $v = \varphi^{-1}((1-s)\,\varphi(t))$

(iv) Repeat (i) to (iii) T times, receive two sets **u** and **v**

Note that **u** and **v** have a length of T, and their values are between 0 and 1. One can further compute the "raw data" by defining the marginal distributions of the copula F_X and F_Y such that $x = F_X^{-1}(u)$ and $y = F_Y^{-1}(v)$.

In addition to the simulation procedure, Archimedean generator can also be used to calculate theoretical Kendall's Tau rank correlation (Genest & MacKay, 1986).

$$\tau_C = 1 + 4 \int_0^1 \frac{\varphi(t)}{\varphi'(t)}\, dt \tag{4.1}$$

For Gumbel copula, the Archimedean generator is defined by

$$\varphi^{Gu}(t) = (ln(\frac{1}{t}))^{\theta} \tag{4.2}$$

where $\theta \geq 1$. Note that this generator satisfies all of the four characteristics as discussed above. And hence, as noted in the last chapter and by using the equation (3.9), the bivariate Gumbel copula is then defined by $C_{\theta}^{Gu}(u,v) = e^{-((-lnu)^{\theta}+(-lnv)^{\theta})^{1/\theta}}$ where $1 \leq \theta < \infty$. When $\theta = 1$, the upper tail coefficient of a bivariate Gumbel copula is 0, and when θ approaches ∞ the upper tail coefficient is 1. The lower tail coefficient of a Gumbel copula does not depend on θ and is always 0. In addition, by using equation (4.1), Kendall's Tau rank correlation is given by

$$\tau_{Gu} = 1 - \frac{1}{\theta} \tag{4.3}$$

This means that Kendall's Tau rank correlation also depends on θ. Since $\theta \geq 1$, τ_{Gu} is between 0 and 1.

In summary, both upper tail and overall dependence are linked to the only parameter θ. The larger the θ, the larger the upper tail and overall dependence. In other words, for a Gumbel copula, a higher upper tail coefficient will be accompanied by a higher rank correlation, and a lower upper tail coefficient will be accompanied by a lower rank correlation. We know that, however, this will probably not work for bivariate data series for which they are only correlated sometimes. For example, a coffee stock in Japan is not correlated to an oil stock in the US during normal times. However, during a global economy crisis which lasts for more than a year, most stocks including these two will go down as a result; hence, these two stocks are strongly correlated during these times. Therefore, there exist some bivariate data series for which the tail dependence is pretty strong but the overall dependence may be very small. In this case, if Gumbel copula is used, it will probably not be fitted well.

4.2 Specification

We consider the following innovative copula: a modified generator function of a bivariate Gumbel copula is given by

$$\varphi^{M.Gu}(t) = (ln(\frac{\delta}{t} - (\delta - 1)))^{\theta} \tag{4.4}$$

where $\theta \geq 1$ and $\delta > 0$ are the parameters. Note that this generator function satisfies all of the strict Archimedean generator conditions: it is continuously decreasing and convex; $\varphi(1) = 0$ and $\varphi(0) = \infty$. As such, a modified bivariate Gumbel copula can be constructed by using equation (3.9):

$$C^{M.Gu}(u, v) = \delta(exp((ln(\frac{\delta}{u} - (\delta - 1)))^{\theta} + (ln(\frac{\delta}{v} - (\delta - 1)))^{\theta})^{1/\theta} + (\delta - 1))^{-1} \tag{4.5}$$

Note that if δ equals 1, then it becomes the original Gumbel copula. In other words, Gumbel copula is nested under modified Gumbel copula. Also note that the upper and lower tail coefficients of this copula can be obtained by using (3.12) and (3.13). It is interesting that the lower tail coefficient of the modified Gumbel copula is derived to be 0, and the upper tail coefficient is $2 - 2^{1/\theta}$; these are the same as the upper and lower tail coefficients of the original Gumbel copula. Details of derivations of upper and lower tail coefficients for modified Gumbel copula are shown in Appendix A.

Furthermore, by using equation (4.1), Kendall's Tau rank correlation can also be calculated for modified Gumbel copula, and is given by

$$\tau_{M.Gu} = \begin{cases} 1 + 4\frac{-\delta^2 \; ln(\delta) \; + \; \delta \; - \; 1}{6(\delta-1)^2 \; \theta} & \delta \neq 1 \\ 1 - \frac{1}{\theta} & \delta = 1 \end{cases} \tag{4.6}$$

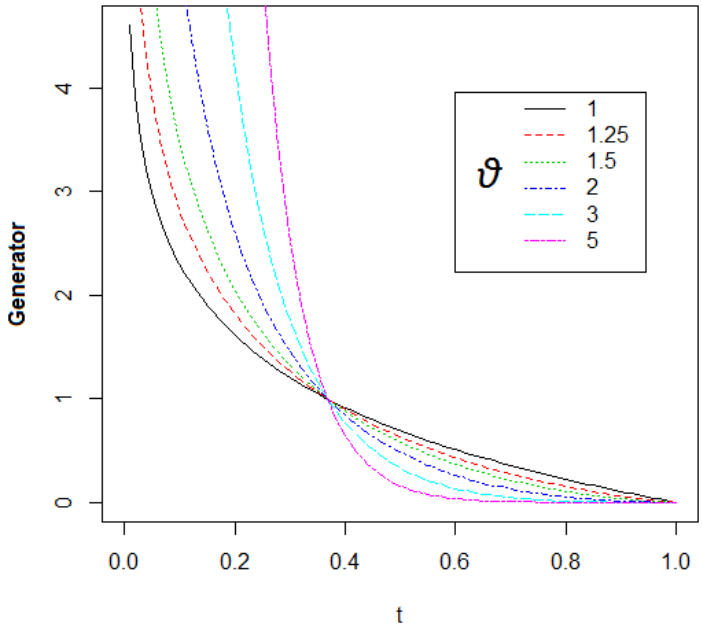
In other words, $\tau_{M.Gu}$ depends on both parameters θ and δ. We firstly look at the effect of θ. Similar to τ_{Gu}, the larger the θ, the larger the rank correlation. However, the effect of δ on the correlation is not clear; this will be discussed in section 3.3. Note that when $\delta = 1$, this will equal to $1 - \frac{1}{\theta}$ which is the just τ_{Gu}. Finally, the derivation of $\tau_{M.Gu}$ are shown in Appendix A.

4.3 Comparing Gumbel and Modified Gumbel Copulas

In order to understand the significance of modified Gumbel copula, one should firstly understand the differences between Gumbel and modified Gumbel copulas and the purpose of the second parameter δ in the modified Gumbel copula.

Firstly, we compare the generator of Gumbel and modified Gumbel copulas. Figure 4.1 displays the Gumbel copula generator function for different θ. Keep in mind that this is equivalent to a modified Gumbel copula where $\delta = 1$. Furthermore, Figure 4.2 displays the modified Gumbel generator

Figure 4.1: Gumbel copula generator function



function for different θ but δ is fixed at 2. Generally, although the patterns look the same (that is, the smaller the θ, the faster the generator function decreases from ∞ to 1, and the slower it decreases from 1 to 0), one can easily observe the differences between these two plots. In addition, Figure 4.3 displays the modified Gumbel generator function for different δ but θ is fixed at 2. This plot is obviously completely different from the above two. The smaller the δ, the faster the function decreases from ∞ to 0.

Secondly, we compare the simulated observations from these two copulas by using Algorithm 4.1. Figure 4.4 shows the 500 pairs of simulated observations **u** and **v** for different values of the parameter θ. When θ is 1, the plot displays no rank or tail dependence. However, when θ becomes larger, both upper tail and rank dependence increase, though this effect on upper tail dependence is more notable.

Figure 4.2: Modified Gumbel copula generator function, $\delta = 2$

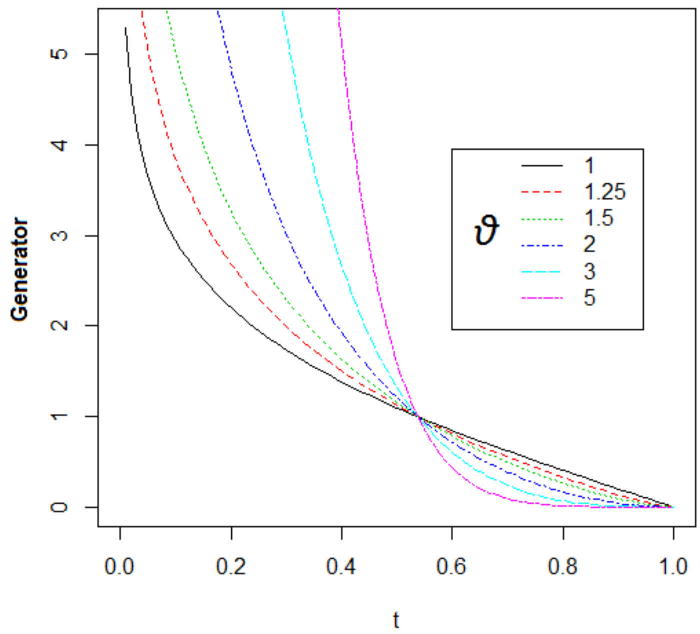


Figure 4.3: Modified Gumbel copula generator function, $\theta = 2$

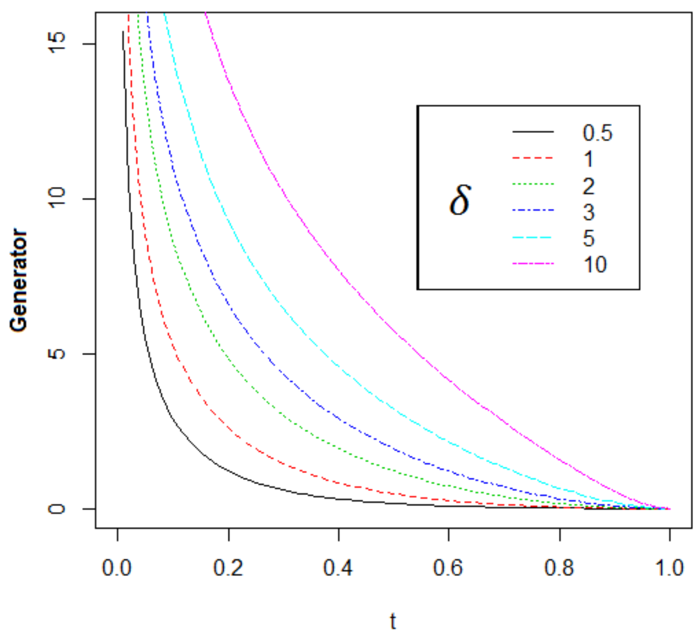


Table 4.1: Sample Spearman's Rho corresponding to Figure 4.4

θ	δ	Rank Correlation
1	1	-0.017
1.25	1	0.258
1.5	1	0.436
2	1	0.645
3	1	0.825
5	1	0.933

Figure 4.5 also shows the 500 pairs of simulated observations **u** and **v** for different values of θ, and $\delta = 2$. Note that the observations of both Figurers 4.4 and 4.5 are simulated from the same random sets **r** and **s** in Algorithm 4.1. Same as the original Gumbel copula, when θ is 1, the tail dependence is observed to be 0; and when θ becomes larger, both upper tail and rank dependence increase as shown in the plots.

The differences between Gumbel copula and modified Gumbel copula are also interesting. Figure 4.6 shows the 500 pairs of simulated observations **u** and **v** for $\theta = 2$ and different values δ. Always note that if δ is 1, then the modified Gumbel copula reduces to the original Gumbel copula. As shown in the plots, when θ is fixed at 2, no matter what value of δ is, the theoretical tail coefficient is always $2 - 2^{1/\theta}$. However, the overall dependence of the bivariate data does decrease when δ increases. This can be further justified in Tables 4.1, 4.2 and 4.3 where the sample Spearman's Rho rank correlations or $r(u, v)$ (overall dependence) corresponding to Figures 4.4, 4.5 and 4.6 are calculated and recorded.

Figure 4.4: Simulated bivariate observations from Gumbel copula

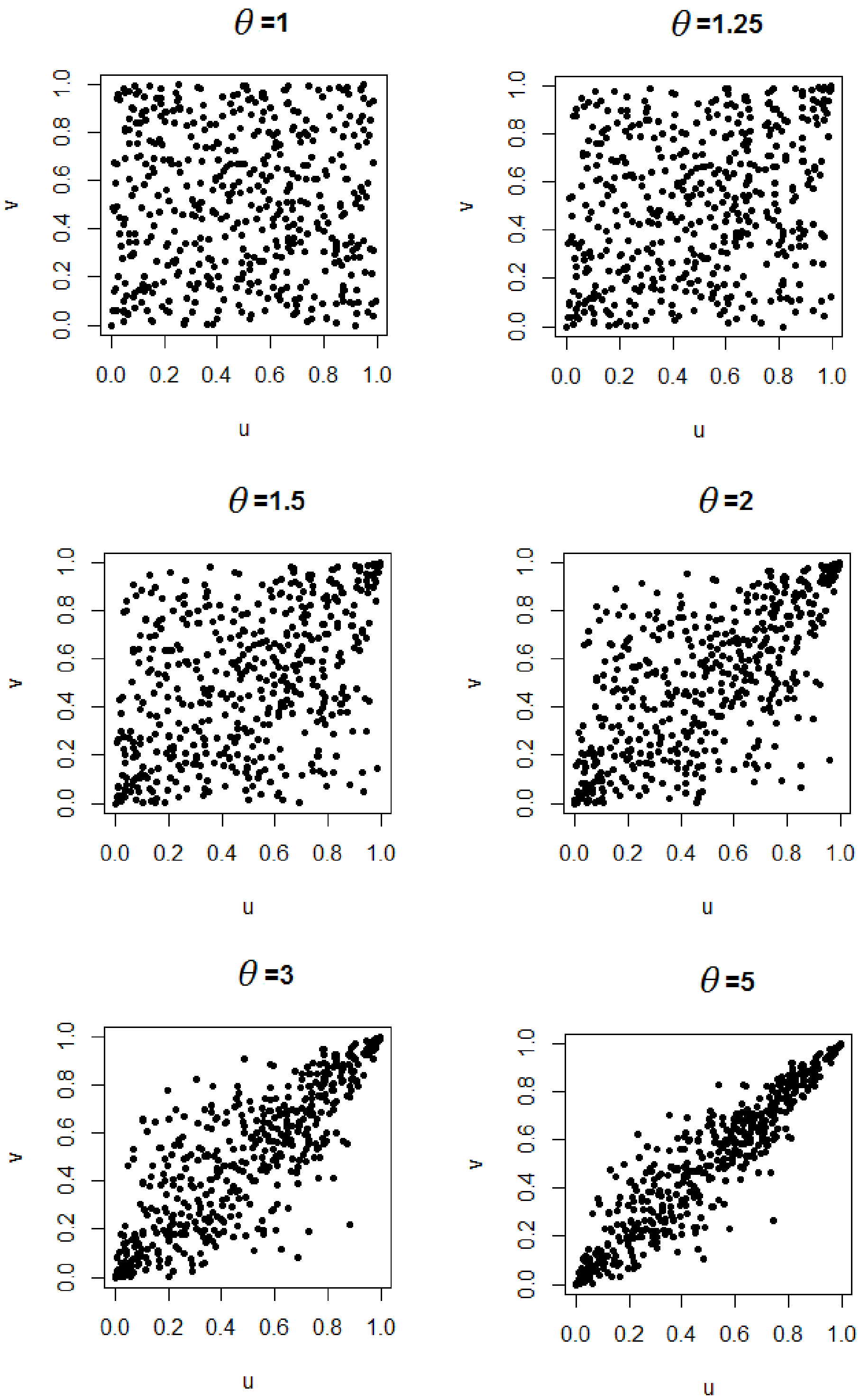


Figure 4.5: Simulated bivariate observations from Modified Gumbel copula with $\delta = 2$

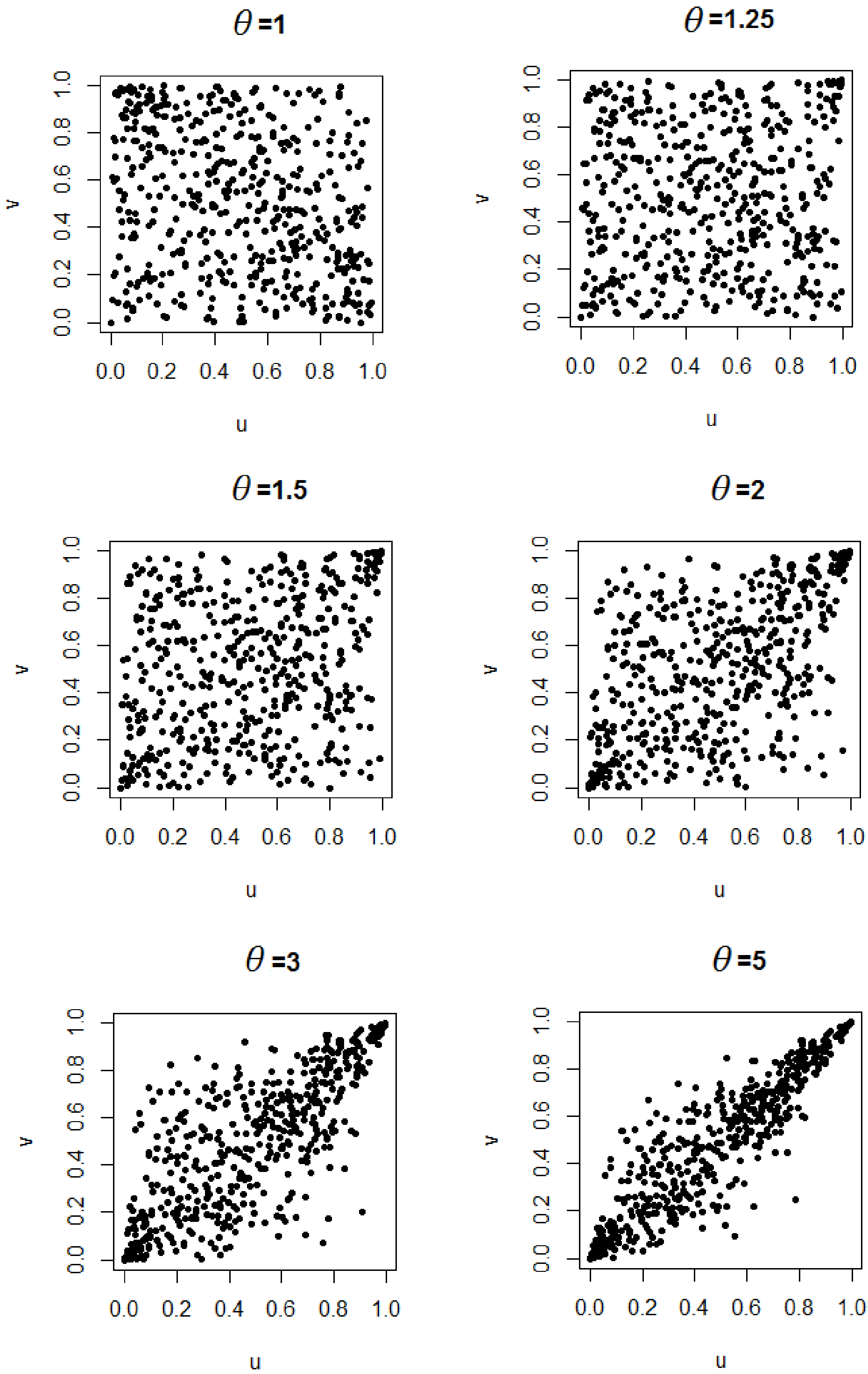


Figure 4.6: Simulated bivariate observations from Modified Gumbel copula with $\theta = 2$

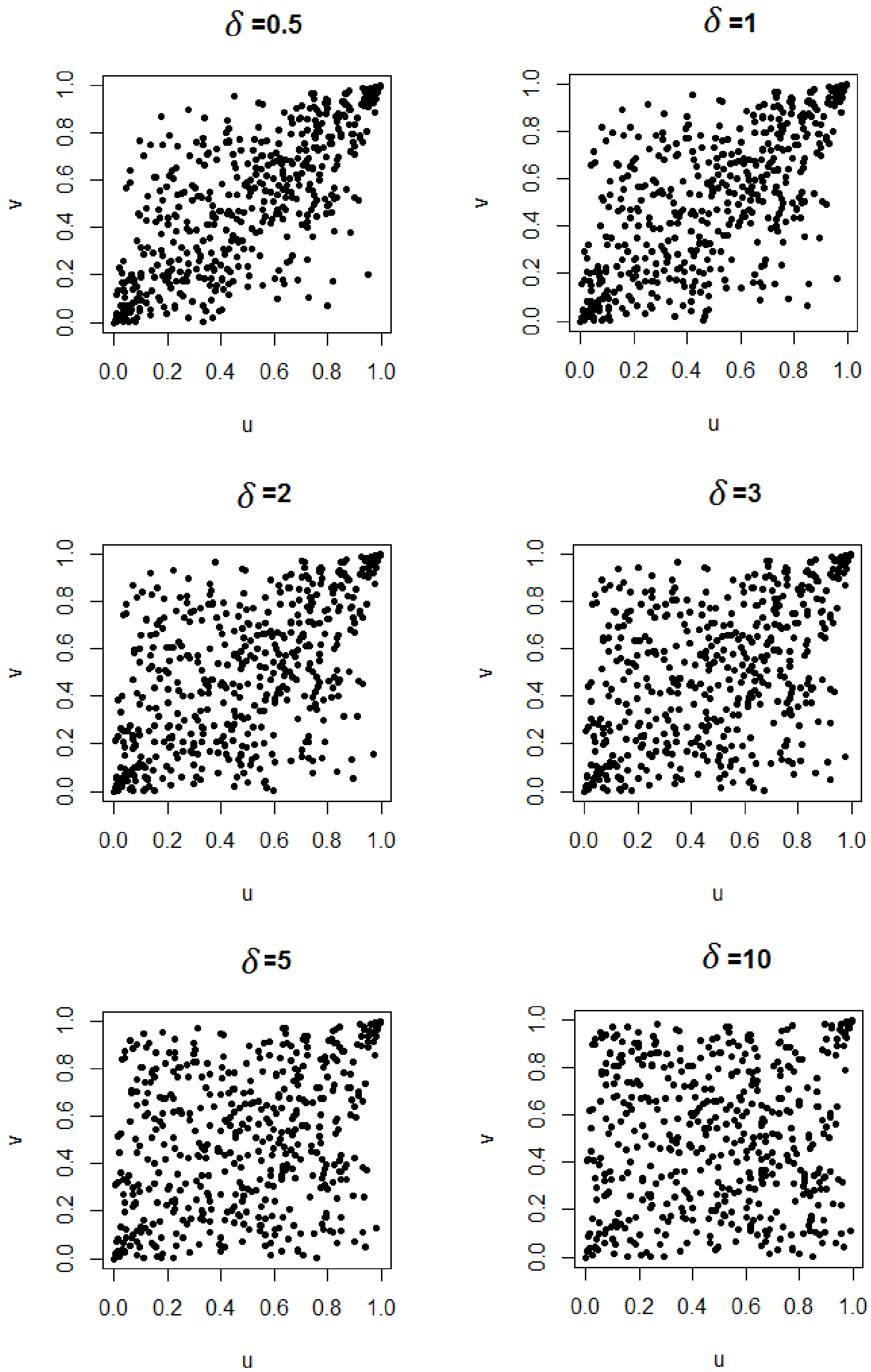


Table 4.2: Sample Spearman's Rho corresponding to Figure 4.5

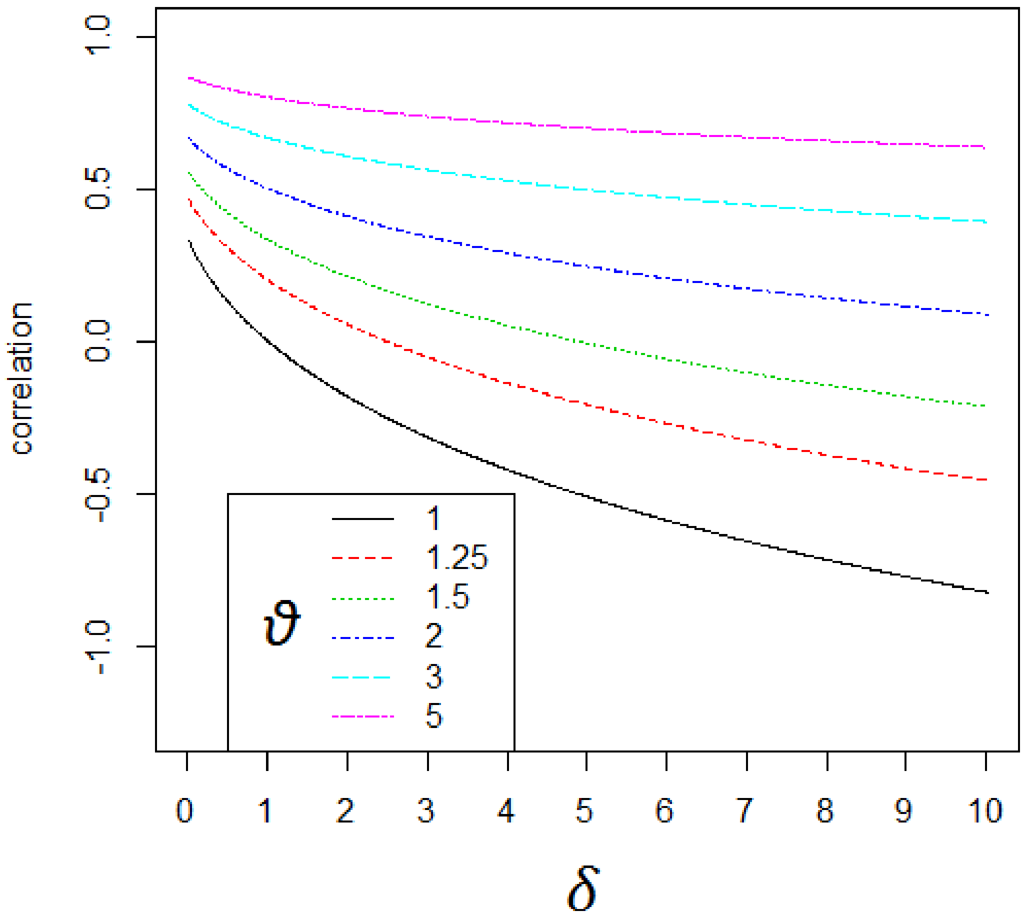
θ	δ	Rank Correlation
1	2	-0.268
1.25	2	0.059
1.5	2	0.274
2	2	0.532
3	2	0.762
5	2	0.906

Table 4.3: Sample Spearman's Rho corresponding to Figure 4.6

θ	δ	Rank Correlation
2	0.5	0.720
2	1	0.645
2	2	0.532
2	3	0.446
2	5	0.317
2	10	0.108

Since there exists evidence that when δ increases the overall dependence will decrease as a result, it is reasonable to also justify the theoretical Kendall's Tau rank correlation. Figure 4.7 displays the theoretical Kendall's Tau $\tau_{M.Gu}$ for different θ. It can be observed that $\tau_{M.Gu}$ is a decreasing function of δ. However, when θ increases, the whole function will be shifted upward. Similarly, Figure 4.8 also displays the theoretical Kendall's Tau $\tau_{M.Gu}$ for different δ; $\tau_{M.Gu}$ is an increasing function of θ. However, when δ increases, the whole function will be shifted downward.

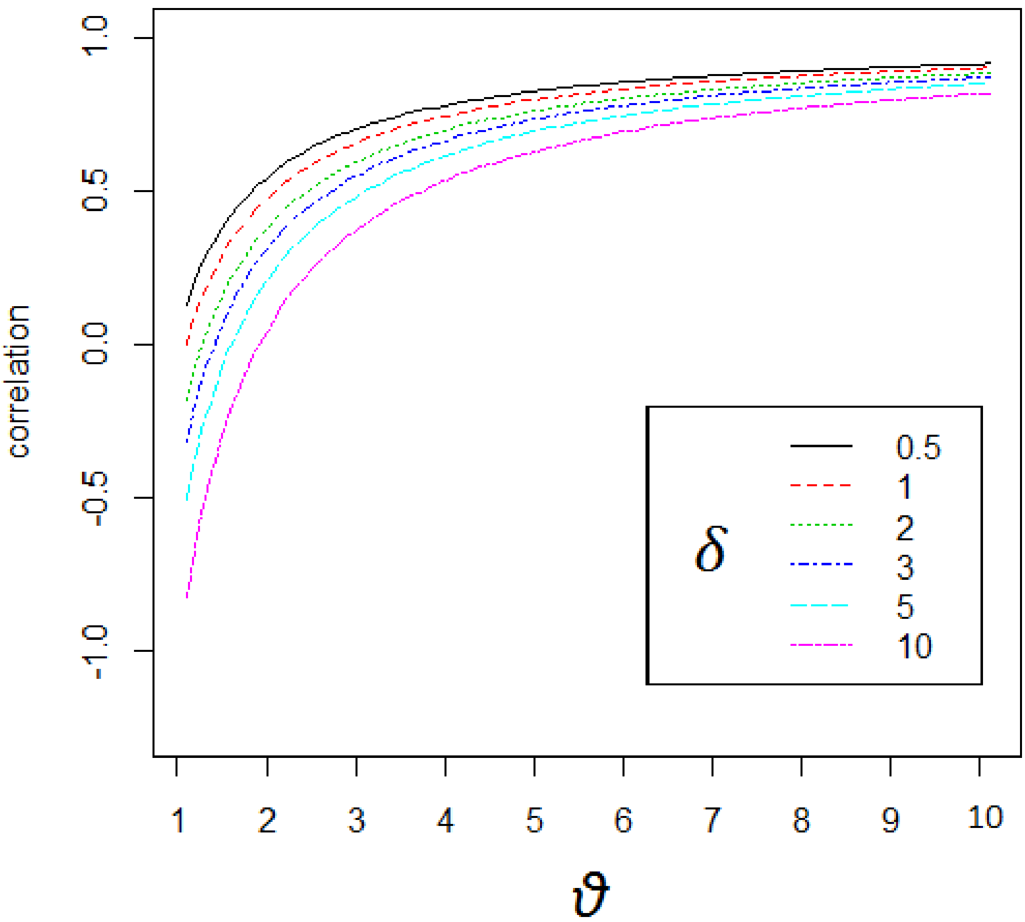
Figure 4.7: Theoretical Kendall's Tau for modified Gumbel copula



All in all, for Gumbel copula, when θ increases, both overall and upper tail dependence increase. On the other hand, both overall and tail dependence of a modified Gumbel copula depend on both the parameters θ and δ. As θ increases, both tail and rank dependence increase; however, as δ increases, tail dependence remains the same and overall dependence decreases. While θ controls both the upper tail and overall dependence, δ controls the overall dependence of the distribution.

One of the reasons we try to observe tail dependence is because there

Figure 4.8: Theoretical Kendall's Tau for modified Gumbel copula



exists no or smaller linear and rank correlations. When the rank correlation (and probably linear correlation) is larger than or about the same as the tail correlation, we probably do not need the tail correlation to explain the data.

With one more parameter in the modified Gumbel copula, one has a greater chance to determine the "true" distribution of the data with upper tail dependence. For example, if two stocks are only correlated during good or bad times (e.g. economic crisis), modified Gumbel copula will be

Table 4.4: Summary of modified Gumbel copula tail coefficients between X and Y

	Copula Class	Log-likelihood value	GoF p-value	Lower Tail Coefficient	Upper Tail Coefficient
Innovation*	90^o R.M.Gumbel	2.922	0.9995	0	0.257

**For this copulas, GoF test is based on Algorithm 2.1*

very useful in this case. However, since Gumbel copula is nested under modified Gumbel copula, one has to conduct some hypothesis tests, such as log-likelihood ratio test, to determine whether Gumbel copula should be rejected and hence modified Gumbel copula should be used.

4.4 Fitting Modified Gumbel Copula

For our data series, modified Gumbel copula will be fitted to $\hat{U}$ and $\hat{V}$ and comparisons between the results by using Gumbel and modified Gumbel copulas will be drawn. Table 4.4 shows the summary of the fitted 90^o rotated modified Gumbel copula. Surprisingly, both log-likelihood value and goodness of fit test p-value for modified Gumbel copula are larger than those for Gumbel copula. Moreover, modified Gumbel copula yields a larger upper tail coefficient.

Since Gumbel copula is nested under modified Gumbel copula, log-likelihood test shall be conducted to determine if Gumbel copula shall be rejected for our analysis. By using the log-likelihood value of both copulas fitted to our data, the log-likelihood ratio test yields a large p-value 0.73, and hence the null hypothesis that the data follows a Gumbel copula distribution shall not be rejected.

Nevertheless, in order to draw some comparisons, both Gumbel and modified Gumbel copulas will be used to forecast data series which will be discussed in the next chapter. Since modified Gumbel copula is not preferred to Gumbel copula based on the log-likelihood ratio test, we expect that these two copulas will yield very close results.

Chapter 5

Equity-linked Insurance

This chapter specifies the equity-linked insurance to be used for payoffs calculations. Our analysis show that by using the forecasted mortality and index price generated from copulas, the expected value of payoffs, variance and reserve are all smaller than those by using independent data series.

5.1 Forecast of index Price and mortality

Once the copulas are determined for the data series, bivariate observations can be generated and index price as well as mortality rate can be forecasted. Then, insurance payoffs can be calculated by using the forecasted data series. The simulation algorithm is as follows:

Algorithm 5.1. Forecast of index price and mortality rate

1. Simulate 300 pairs of bivariate observations $\hat{U}$ and $\hat{V}$ from the Gumbel and modified Gumbel copulas by using Algorithm 4.1.

In addition, simulate 300 pairs of bivariate independent observations $\hat{U}$ and $\hat{V}$ from $Unif(0,1)$.

2. Transform these simulated observations to $\hat{X}$ and $\hat{Y}$ by using empirical distribution functions of the original data X and Y.

3. Calculate both forecasted Log mortality and Log index recursively by using the time series equations.

4. Repeat Step 1 to Step 3 a total of 5000 times to obtain 5000 paths of forecasted data series.

Both Gumbel and modified Gumbel copulas will be used to simulate bivariate observations $(\hat{U}, \hat{V})$ by using Algorithm 4.1. In addition, in order to draw comparisons between dependent and independent data series, bivariate independent observations $(\hat{U}, \hat{V})$ will also be simulated from $Unif(0,1)$. Then, these observations are to be transformed to $(\hat{X}, \hat{Y})$ by using the empirical CDF of X and Y respectively. Finally, one can calculate the forecasted Log mortality as well as forecasted Log index recursively by using the time series equations.

Note that outlier models are not necessary in this case. As mentioned earlier, an outlier model is a combination of an ARIMA model and outliers, and the inputs X and Y for copulas are $\hat{e}_t$, the residuals inclusive of outlier effects of the outlier model. Since $\hat{e}_t$ is the residuals obtained from the ARIMA model fitted to the original data series (subject to the external iteration cycle), the ARIMA models with the corresponding parameters fitted to Log mortality and Log index shall be used for $\hat{X}$ and $\hat{Y}$ respectively.

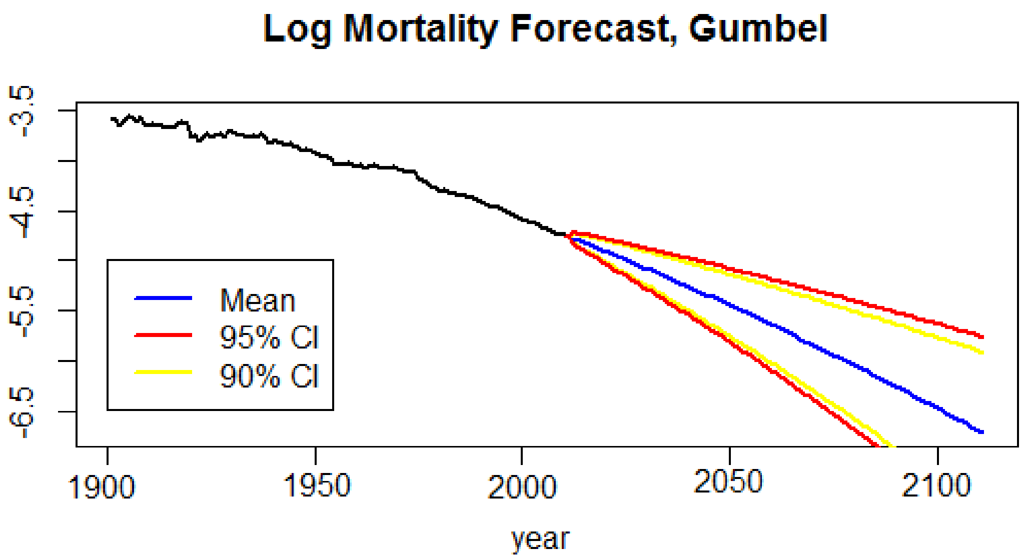
That is, ARIMA(1,1,2) will be used to calculate forecasted Log mortality, and ARIMA(0,1,0) will be used to calculate forecasted Log index.

Also, by taking exponential functions on forecasted Log mortality and Log index, forecasted mortality as well as index can then be drawn. These two forecasted data series will be used to calculate the payoffs of an equity-linked insurance.

For the case by using Gumbel copula, Figure 5.1 shows the mean, 90% and 95% confidence interval of the 5000 paths of forecasted Log mortality as well as Log index from year 2011 to 2110. Generally, it can be observed from the plots that the forecasted Log mortality exponentially decreases and Log index linearly increases throughout the time.

Similarly, for the case by using modified Gumbel copula, the forecasted data series are shown in Figures 5.2. In addition, by using bivariate independent observations, the forecasted data series are displayed in Figures 5.3. Generally, these three figures look very much the same and one can hardly tell the differences.

Figure 5.1: Forecasted Log mortality and Log index, with mean in blue, 90% CI in yellow, and 95% CI in red by using Gumbel copula



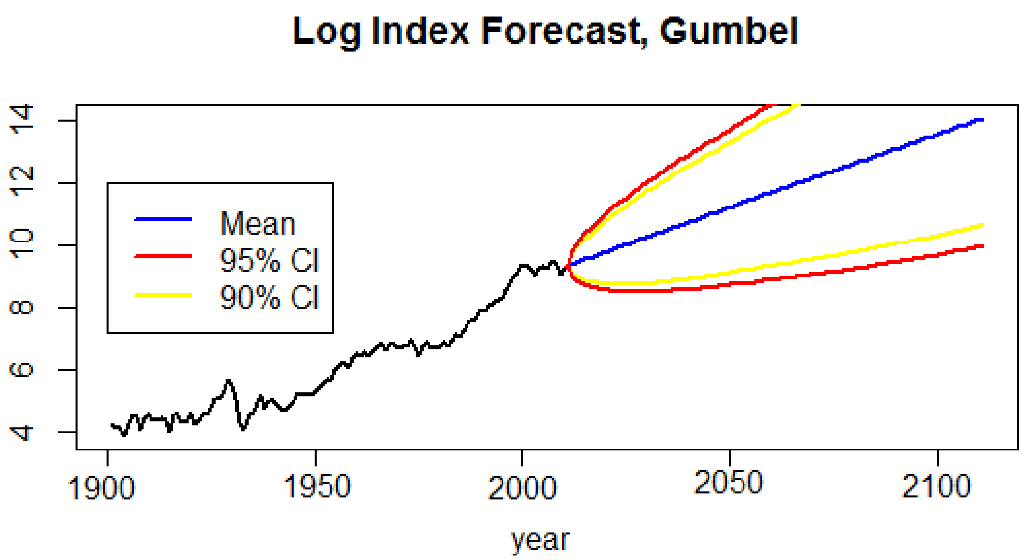
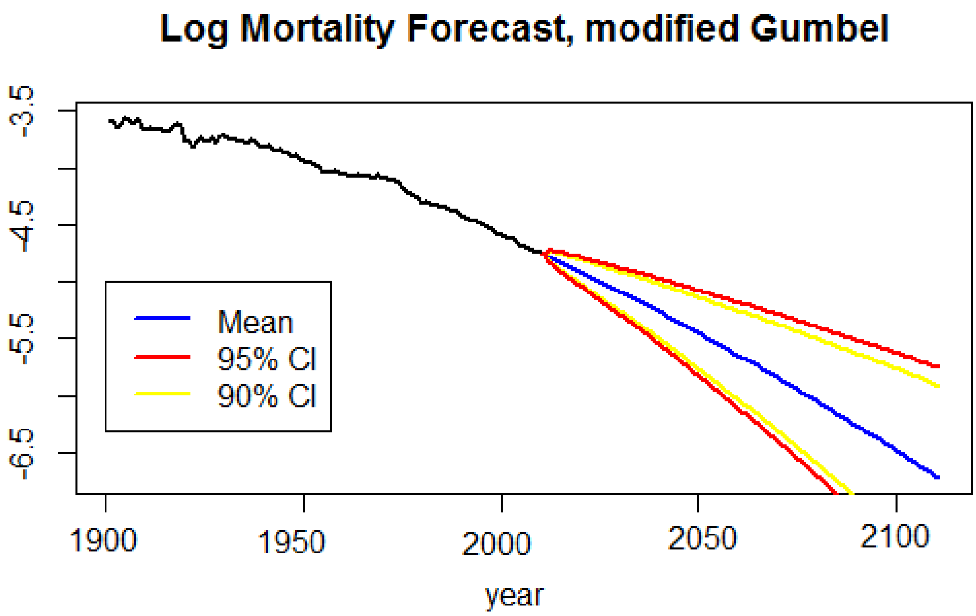


Figure 5.2: Forecasted Log mortality and Log index, with mean in blue, 90% CI in yellow, and 95% CI in red by using modified Gumbel copula



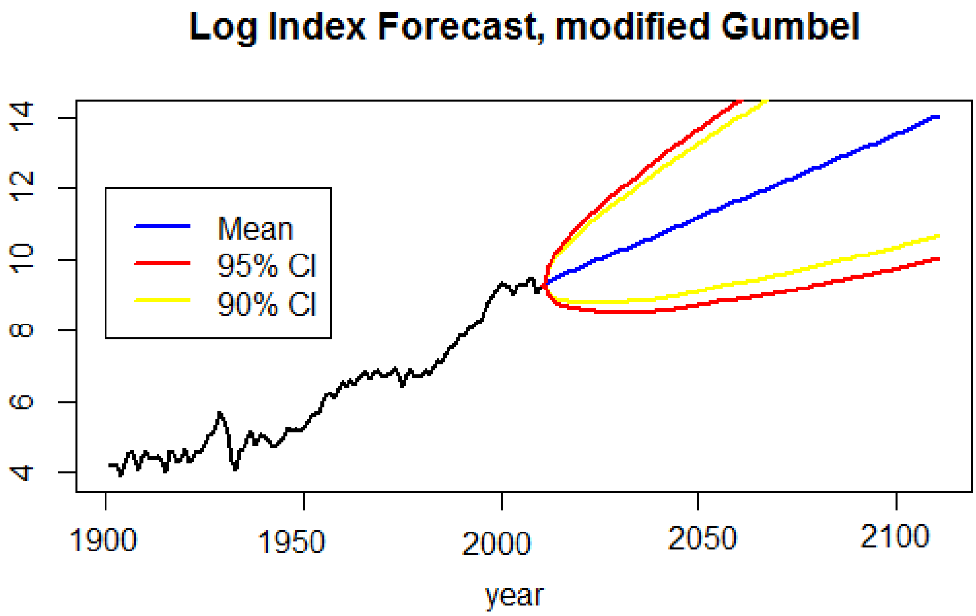
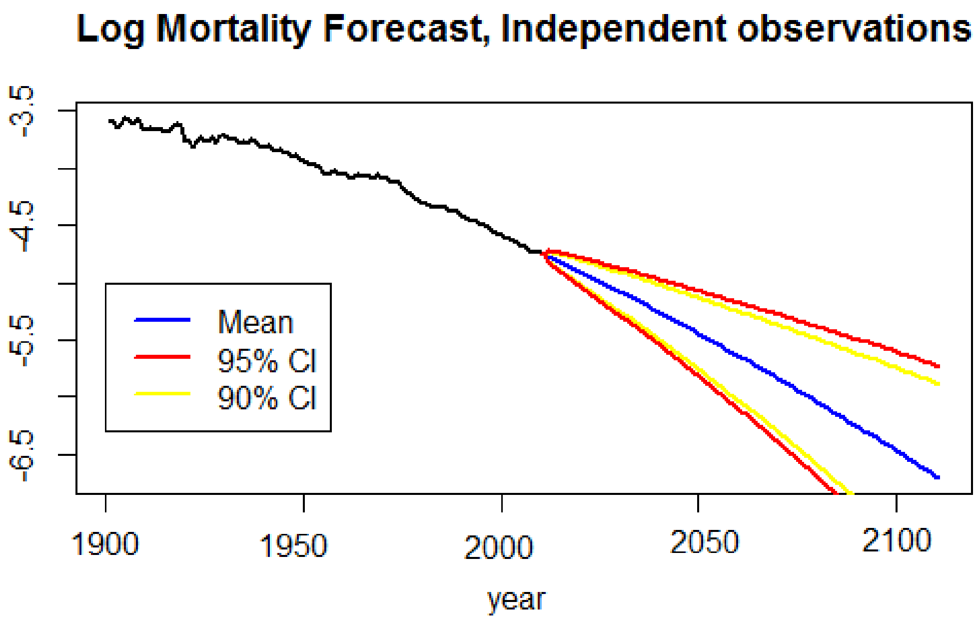
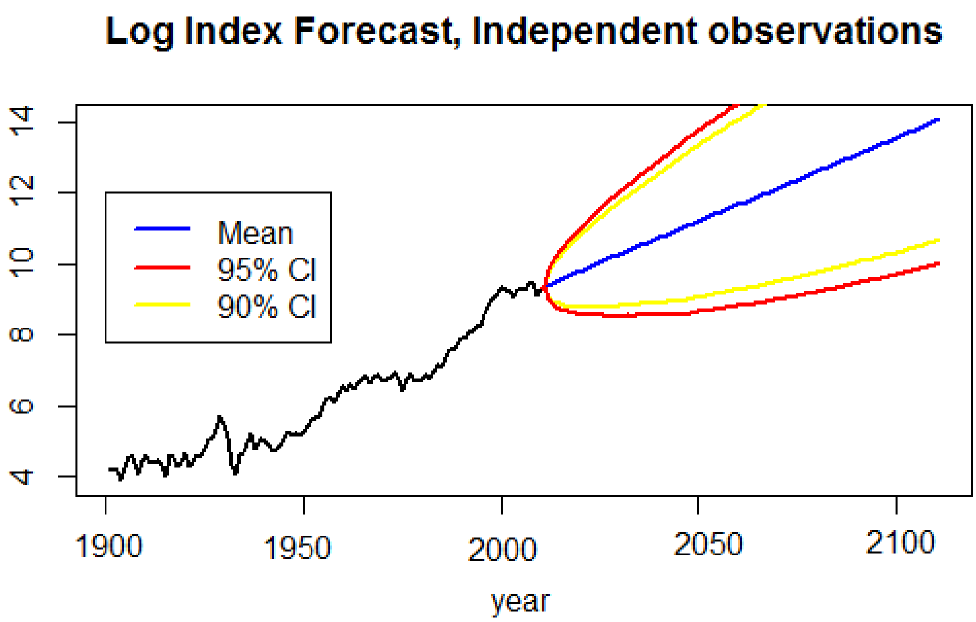


Figure 5.3: Forecasted Log mortality and Log index, with mean in blue, 90% CI in yellow, and 95% CI in red by using independent observations





5.2 An Illustrative Equity-linked Insurance

The price of an equity-linked insurance is determined by using at least two variables: portfolio or index returns, and mortality rate. For example, the payoff of an equity-linked insurance at time t (Bernard & Lemieux, 2008) is defined by

$$V_t = \alpha P\ max((1+g)^t, (\frac{S_t}{S_0})^k) \tag{5.1}$$

where α is a ratio regulated by law, P is the single premium or based rate, g is the minimum guaranteed rate, t is the time the insurance benefit is paid, k is the participation rate, and S_t and S_0 are the stock or index price at time t and 0 respectively.

We will use an equity-linked insurance for which the payoff is a death benefit payable at the end of the year of death. First, for our case, since we use mortality data of the age group "55 to 64 years" to forecast future mortality rates, these forecasted data should only be used for those who fall in the age group "55 to 64 years". For example, in 2021, a 55-year-old shall be subject to the forecasted mortality rate of 2021. Next year, she will turn 56 and hence will still be in the age group "55 to 64 years"; therefore, she will be subject to the forecasted mortality rate of 2022. This will continue until she reaches 65 for which she is no longer in the age group "55 to 64 years". Hence, the forecasted mortality rate is year specific instead of age specific. Due to this reason, we propose to calculate the payoffs of a 10-year term equity-linked insurance contract.

Second, if copulas that can capture upper and lower tail dependence are used to simulate the observations, then we assume that there exists some dependence between the two data series mortality and index. We define EPV the expected present value of insurance payoffs. We will use the following to compute EPV; we consider an insurance where the death benefit is linked to the equity return and payable at the end of the year of death:

Algorithm 5.2. EPV of insurance portfolio calculation.

1. Suppose at the beginning of year T', we have $N = 10000$ policyholders aged 55 in the portfolio, then the expected number of policyholders who will survive $t-1$ years and die in the following year is $N\ {}_{t-1}p_{55,i,T'}\ q_{55,i,T'+t-1}$

2. For each of the 5000 forecast paths, calculate the conditional expected present value of the payoffs. That is, for $i = 1, 2, ..., 5000$,

$$EPV_i = \sum_{t=1}^{10} e^{-rt}\, \alpha P\ max((1+g)^t, (\frac{S_{i,T'+t-1}}{S_{i,T'-1}})^k)\ N\ {}_{t-1}p_{55,i,T'}\ q_{55,i,T'+t-1} \tag{5.2}$$

3. Calculate the sample mean $\overline{EPV}$ by taking the average of the 5000 EPV's.

$$\overline{EPV} = \frac{1}{5000} \sum_{i=1}^{5000} EPV_i \tag{5.3}$$

Note that $S_{i,T'-1}$ denotes the forecasted index price of year $T'-1$ of the i^{th} forecast path, ${}_{t-1}p_{55,i,T'}\ q_{55,i,T'+t-1}$ denotes the probability that a 55-year-old in year T' will survive $t-1$ years and die in the following year, calculated by using the forecasted mortality rate of the i^{th} forecast path. According to Bernard and Lemieux (2008), the benchmark values are $P = 100$, $\alpha = 0.85$, $r = 0.04$, $g = 0.02$, and $k = 0.9$.

In order to calculate the required fund for the insurance portfolio, apart from calculating the mean of the 5000 EPV's, one can also obtain 95^{th} quantile ($EPV_{0.95}$) of the sample such that

$$\frac{1}{5000}[\sum_{i=1}^{5000} 1(EPV_i \leq EPV_{0.95})] = 0.95 \tag{5.4}$$

More specifically, $EPV_{0.95}$ will be the minimum fund to be reserved such that the probability that the future loss will be well covered is 95%. Alternatively, one can firstly calculate the variance of the sample mean. Then, by using normal approximation and for a 95% confidence level, the required fund to be reserved such that the future loss will be well covered can be obtained:

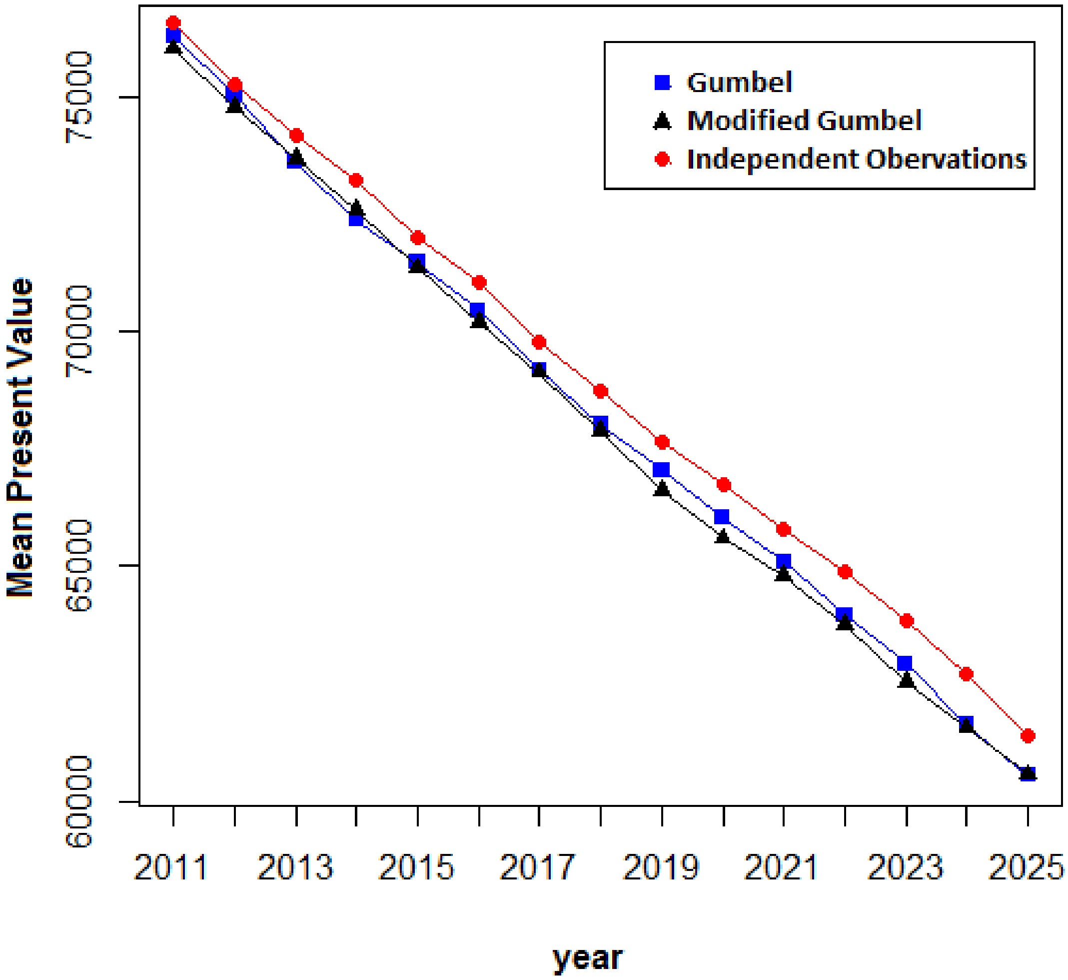
$$Var(\overline{EPV}) = \frac{1}{5000}[\frac{1}{5000}\sum_{i=1}^{5000}(EPV_i)^2 - \overline{EPV}^2] \tag{5.5}$$

$$Fund = \overline{EPV} + 1.96\sqrt{Var(\overline{EPV})} \tag{5.6}$$

5.3 Numerical Results

We will still use Algorithm 5.2 to calculate the $\overline{EPV}$ of insurance portfolio payoffs. By using Gumbel copula, modified Gumbel copula and independent observations, Figure 5.4 shows the $\overline{EPV}$ of payoffs of an equity-linked insurance portfolio being sold to a group of 10,000 55-year-olds in $T' = 2011, 2012, \ldots, 2025$. It can be observed, the $\overline{EPV}$ by using

Figure 5.4: $\overline{EPV}$ of an equity-linked insurance portfolio being sold to a group of 10,000 55-year-olds in $T' = 2011, 2012, ..., 2025$, by using copulas and independent observations



any of the copulas is always smaller, and the difference is increasing over time.

Both Gumbel and modified Gumbel copulas fitted to our data yield a large goodness of fit test p-value. Furthermore, the upper tail dependence of Gumbel copula is statistically significant at 0.05. This means that a very low mortality will be accompanied by a very high index price. However, these effects on the portfolio payoff is ambiguous. Firstly, if mortality decreases, then the $\overline{EPV}$ of portfolio payoffs will also decrease. Secondly, if index price (return) increases, then the $\overline{EPV}$ of portfolio will increase as well. Therefore, if a low mortality is accompanied by a high index price in a particular year, then the effects (decrease and increase in $\overline{EPV}$ of portfolio payoffs) may cancel out. But if the magnitude of increase in index price is larger than that of decrease in mortality, then the $\overline{EPV}$ of portfolio payoffs will go up, and vice versa.

Nevertheless, theoretically, the expected value of multiplications of two variables is equal to multiplication of the expected values plus a covariance. For our case, this covariance will be negative since index price and mortality rate are negatively dependent. Therefore, it is no surprise that the $\overline{EPV}$ by using Gumbel or modified Gumbel copula is smaller than that by using independent observations. However, such reasons may be too abstract.

To understand such effects in a more concrete way, mean of forecasted mortality and forecasted index price, as well as correlation between forecasted mortality and forecasted index price calculated by using the 5000 forecast paths for each of the years of $T' = 2011, 2012, \ldots, 2025$ are shown in Table 5.1. It can be easily observed that for every year, Gumbel copula, modified Gumbel copula and independent observations all yield the same mean forecasted mortality. In addition, Gumbel copula, modified Gumbel copula and independent observations all yield very close mean

Table 5.1: Mean of forecasted mortality, mean of forecasted index, correlation between forecasted mortality and forecasted index, calculated by using 5000 paths of forecasted data for each of the year of $T' = 2011, 2012, \ldots, 2025$

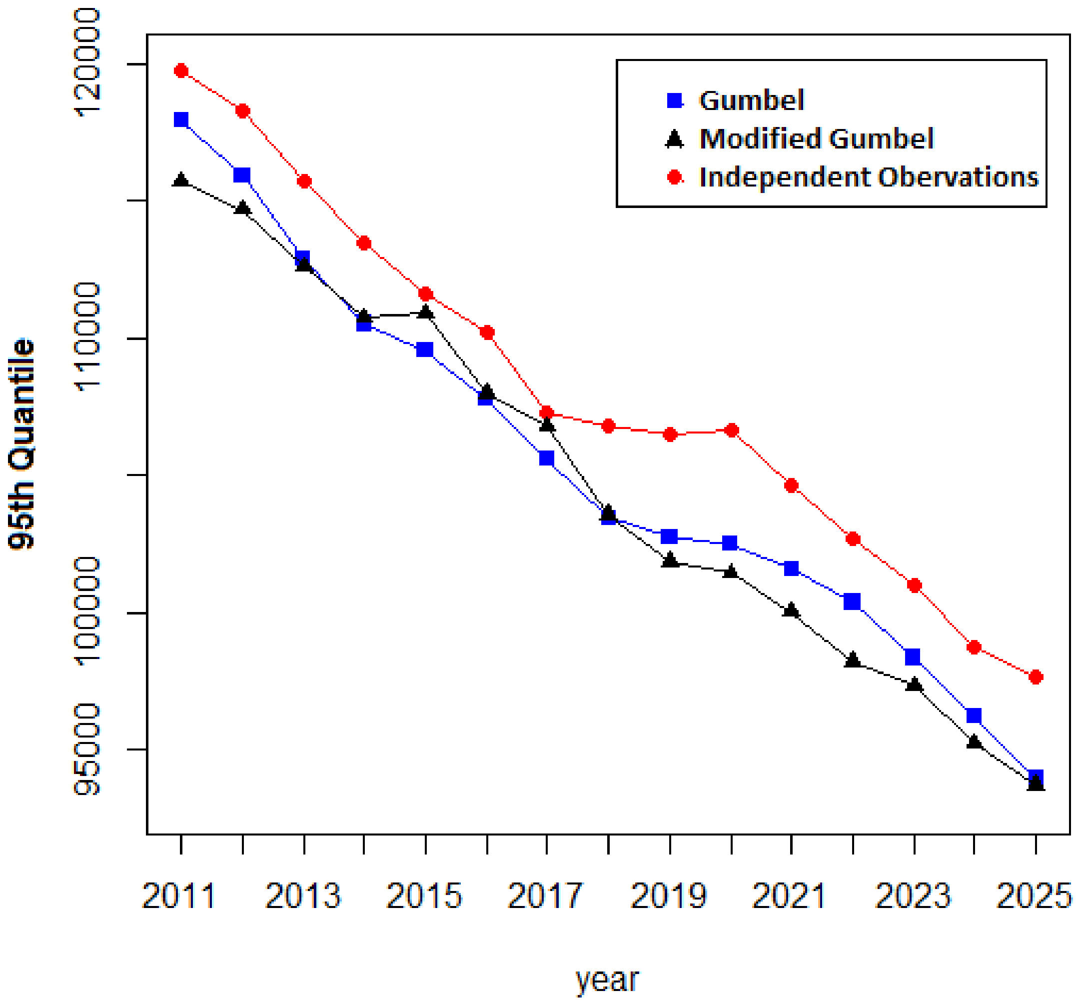
	Mean mortality			Mean index			Cor(mortality,index)		
Year	Indep	Gumbel	M.Gumbel	Indep	Gumbel	M.Gumbel	Indep	Gumbel	M.Gumbel
2011	0.00846	0.00846	0.00846	12395.27	12429.86	12383.88	-0.0255	-0.2288	-0.2219
2012	0.00832	0.00832	0.00833	13239.38	13336.00	13217.87	-0.0313	-0.2243	-0.1937
2013	0.00819	0.00819	0.00820	14088.89	14249.62	14110.12	-0.0160	-0.2237	-0.2057
2014	0.00806	0.00806	0.00806	15083.58	15185.50	15123.25	0.0004	-0.2014	-0.2276
2015	0.00793	0.00793	0.00792	16128.11	16203.51	16170.04	-0.0014	-0.2073	-0.2258
2016	0.00780	0.00780	0.00780	17325.55	17377.56	17262.87	-0.0059	-0.2150	-0.2255
2017	0.00767	0.00767	0.00767	18574.02	18651.38	18453.91	0.0055	-0.2181	-0.2201
2018	0.00755	0.00754	0.00754	19962.78	19891.54	19777.66	-0.0059	-0.2238	-0.2241
2019	0.00743	0.00742	0.00742	21301.93	21196.75	21071.84	0.0058	-0.2197	-0.2134
2020	0.00730	0.00729	0.00729	22778.56	22659.30	22435.82	0.0048	-0.2066	-0.2195
2021	0.00719	0.00717	0.00717	24328.17	24413.54	24072.86	0.0009	-0.2093	-0.2122
2022	0.00706	0.00705	0.00705	25945.16	25967.26	25804.33	-0.0082	-0.1953	-0.1967
2023	0.00694	0.00693	0.00694	27742.88	27801.46	27562.61	-0.0056	-0.1909	-0.2112
2024	0.00683	0.00682	0.00682	29904.77	29681.20	29551.28	-0.0103	-0.1927	-0.2101
2025	0.00671	0.00670	0.00670	32025.92	31695.83	31569.81	-0.0178	-0.1959	-0.1906

forecasted index price and the differences can be neglected. However, by using Gumbel or modified Gumbel copula, the correlation between the forecasted mortality and index is much smaller than that by using independent observations. This means that it is the negative covariance that makes up the largest portion of differences between the $\overline{EPV}$'s.

However, note that the future values of index price and mortality also depend on the historical values. That is, they are time series data instead of variables. In addition, also note that all $\overline{EPV}$'s are decreasing over time, this is due to the fact that forecasted mortality will continue to decrease over time but annual return of forecasted index will remain stationary.

Figures 5.5 shows the 95^{th} quantiles of the 5000 $\overline{EPV}$'s. It can be observed

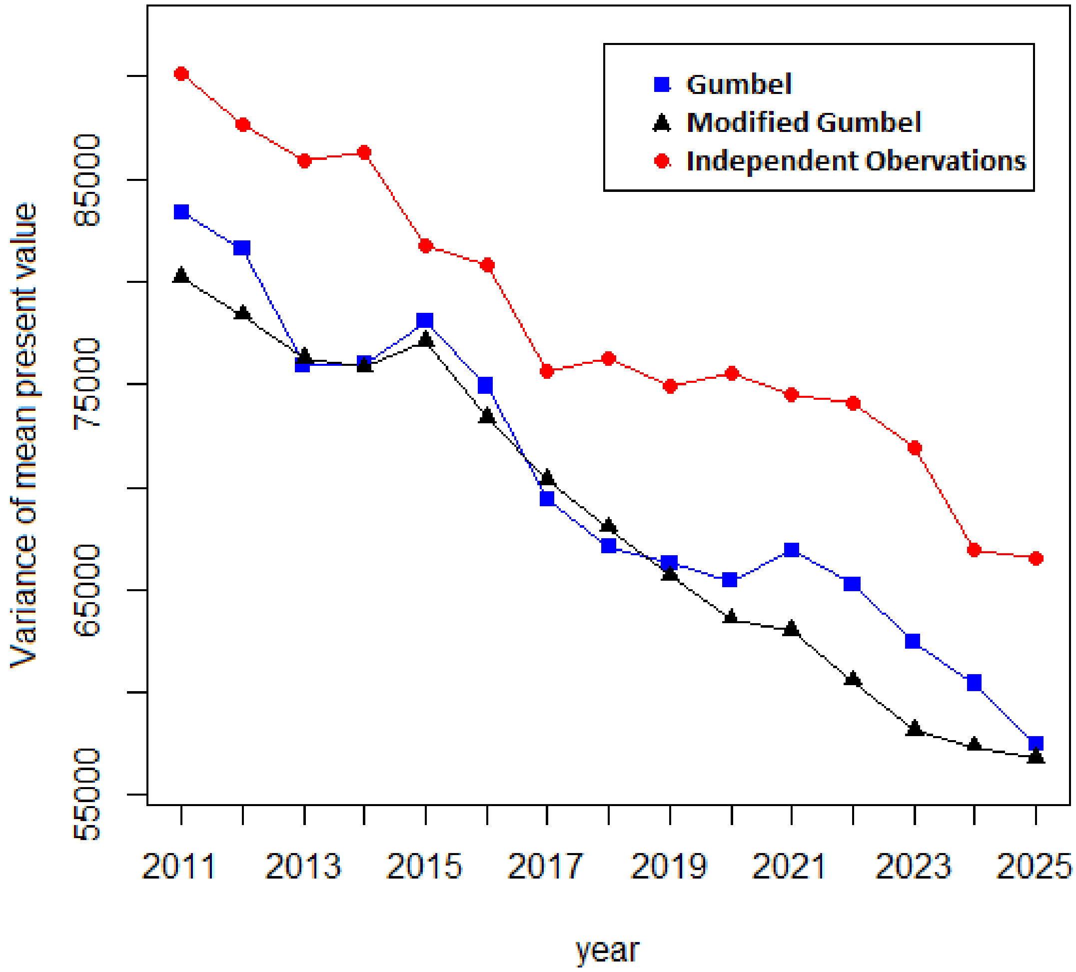
Figure 5.5: 95^{th} quantile of the sample corresponding to Figure 5.4



that the 95^{th} quantile by using any of the copula is always smaller than that by using independent observations, indicating that the required fund to be reserved is smaller.

It would be also interesting to compare the variances (or standard deviations) of the sample mean. Figure 5.6 shows the variance of $\overline{EPV}$ corresponding to Figure 5.4. It can be easily observed that the variance by using independent observations is always larger than that by using

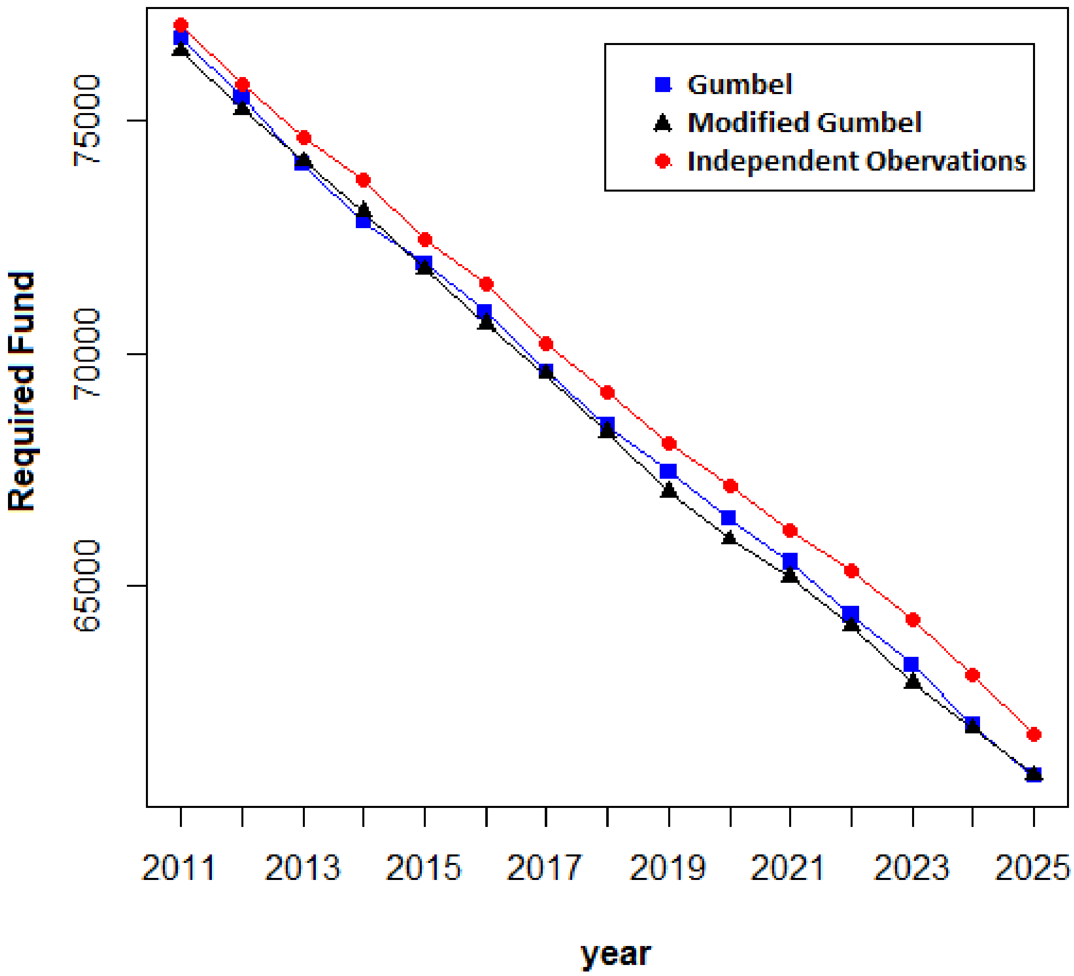
Figure 5.6: Variance of $\overline{EPV}$ corresponding to Figure 5.4



Gumbel or modified Gumbel copula. Such result implies that, by assuming negatively dependent mortality and index price, not only the expected loss is reduced, the corresponding risk shall be smaller.

Finally, by using normal approximation, the required fund to be reserved such that the insurer is 95% confident that the loss will be well covered is also shown in Figures 5.7. With a smaller expected value and variance of insurance payoffs by using Gumbel or modified Gumbel copula, it is no

Figure 5.7: Required fund corresponding to Figure 5.4, by using normal approximation for a 95% confidence level



doubt that the required fund is also smaller than that by using independent observations.

Chapter 6

Conclusion

Time series models that can be used to model human mortality include ARIMA/GARCH model, Lee-Carter model (Lee and Carter, 1992), and general time-series outlier model. The general time-series outlier model is shown to be preferred for our time series data Log American mortality for age group “55 to 64 years” and Log DJIA index price and hence shall be used to obtain the residuals.

In order to capture the tail dependence between the two data series, residuals inclusive of outlier effects should be obtained and used. Then, copulas can be fitted to the empirical CDF of residuals and correlations as well as tail coefficients can be obtained. It has been shown that 90^o rotated Gumbel copulas is fitted very well to the data. Furthermore, both the parameter estimate (1.1577) and upper tail correlation (0.18) of the fitted Gumbel copula are statistically significant. At this point, it can be assumed that there exists an upper tail dependence between the two data series.

In addition, a new copula, modified Gumbel copula, has been developed based on the Archimedean generator of Gumbel copula. 90^o rotated Modified Gumbel copula is also fitted very well to the data. The upper tail coefficient is determined to be 0.257, which is higher than that of Gumbel copula. However, Gumbel copula is nested under modified Gumbel copula; by using the log-likelihood ratio test, Gumbel copula was not rejected at 0.05. Nevertheless, both Gumbel and modified Gumbel copulas are used to forecast Log mortality and Log index.

Once the joint distributions are determined, bivariate observations can be simulated and time series data can be forecasted. Then, the payoffs of an equity-linked insurance can be calculated. In order to highlight the dependence between index price and mortality, one can compare the differences between the equity-linked insurance payoffs by using simulated observations from copulas and independent observations (see Algorithm 5.1).

Our analysis shows that the $\overline{EPV}$ by using Gumbel copula and the $\overline{EPV}$ by using bivariate independent observations do differ, but the difference is not really large. Similarly, the difference between the $\overline{EPV}$ by using modified Gumbel copula and the $\overline{EPV}$ by using bivariate independent observations is initially larger but gradually decreasing over time. Nevertheless, the $\overline{EPV}$ by using any of the copulas is always smaller than that by using independent observations, and this is due to the negative covariance between the forecasted index price and mortality when copulas are used.

Note that when mortality decreases, the payoff of the equity-linked insurance also decreases; and when index price (return) increases, the payoff of equity-linked insurance increases too. Finally, if a very low mortality is accompanied by a very high index price, then these effects (increase and decrease in the payoff) may eventually be canceled out. However, this will not be always the case. If the magnitude of mortality decrease is much larger than that of index price increase, then the payoffs may be smaller. On the other hand, if the magnitude of index price increase is much larger than that of mortality decrease, then the payoffs may be larger. The effects of low mortality accompanied by high index price (return) on the portfolio payoffs may be ambiguous.

Nonetheless, our results show that by using the 5000 forecasted data series for each of the year $T' = 2011, 2012, \dots 2025$, the mean forecasted mortality as well as mean forecasted index are very close by using Gumbel copula, modified Gumbel copula and independent observations. However, the correlation between the forecasted mortality and index is around 0.2 by using Gumbel or modified Gumbel copula, and 0 by using independent observations, indicating that the difference between the $\overline{EPV}$'s is due to the negative dependence between the two data series.

Moreover, both 95% quantile and variance of $\overline{EPV}$ are much smaller than those by using independent observations respectively, indicating that the risk of underwriting is also smaller. Finally, by either using 95% quantile or using normal approximation, the required funds to be reserved such that the future loss will be well covered are also smaller by using copulas. Though, these values tend to be very close to each other as we have shown that Gumbel copula was not rejected based on the log-likelihood test conducted earlier.

In conclusion, if mortality and index price were to follow modified Gumbel or Gumbel copula for which they are dependent, then one should be cautious of the pricing and reserving methods as the dependence between the two data series will reduce insurance payoffs as well as the corresponding risk.

Appendix A

Modified Gumbel Copula

This appendix shows the derivations of both upper and lower tail coefficient as well as Kendall's Tau of a modified Gumbel copula.

A.1 Upper and Lower Tail Coefficients

Following (4.5), we have the equation for modified Gumbel copula. We firstly replace both u and v by q, and then take first derivative on the copula with respect to q. Then, q shall be replaced by 1 and the upper tail coefficient can be derived. For lower tail coefficient, first derivative on the copula with respect to q is not needed as tail coefficient can be directly calculated by replacing q by 1.

$$C(u,v) \;=\; \delta(e^{([ln(\frac{\delta}{u}-(\delta-1))]^{\theta}+[ln(\frac{\delta}{v}-(\delta-1))]^{\theta})^{1/\theta}} + (\delta-1))^{-1}$$

$$\begin{aligned} C(q,q) &= \delta(e^{([ln(\frac{\delta}{q}-(\delta-1))]^{\theta}+[ln(\frac{\delta}{q}-(\delta-1))]^{\theta})^{1/\theta}} + (\delta-1))^{-1} \\ &= \delta(e^{(2[ln(\frac{\delta}{q}-(\delta-1))]^{\theta})^{1/\theta}} + (\delta-1))^{-1} \\ &= \delta(e^{(2^{1/\theta}[ln(\frac{\delta}{q}-(\delta-1))])} + (\delta-1))^{-1} \end{aligned}$$

$$
\begin{aligned}
&= \delta(e^{(2^{1/\theta} ln(\frac{\delta}{q}-(\delta-1)))} + (\delta-1))^{-1} \\
&= \delta(e^{(ln(\frac{\delta}{q}-(\delta-1))^{2^{1/\theta}})} + (\delta-1))^{-1} \\
&= \delta((\frac{\delta}{q}-(\delta-1))^{2^{1/\theta}} + (\delta-1))^{-1}
\end{aligned}
$$

$$
C'(q,q) = -\delta((\frac{\delta}{q}-(\delta-1))^{2^{1/\theta}} + (\delta-1))^{-2}\, 2^{1/\theta}\, (\frac{\delta}{q}-(\delta-1))^{2^{1/\theta}-1}\, (-\frac{\delta}{q^2})
$$

$$
\begin{aligned}
C'(1,1) &= -\delta((\frac{\delta}{1}-(\delta-1))^{2^{1/\theta}} + (\delta-1))^{-2}\, 2^{1/\theta}\, (\frac{\delta}{1}-(\delta-1))^{2^{1/\theta}-1}\, (-\frac{\delta}{1^2}) \\
&= -\delta(\delta)^{-2}\, 2^{1/\theta}\, (1)^{2^{1/\theta}-1}\, (-\delta) \\
&= 2^{1/\theta}
\end{aligned}
$$

$$
\begin{aligned}
\lambda_u &= \lim_{q\to 1^-} \frac{1-2q+C(q,q)}{1-q} \\
&= \lim_{q\to 1^-} \frac{-2+C'(q,q)}{-1} \\
&= 2-2^{1/\theta}
\end{aligned}
$$

$$
\begin{aligned}
\lambda_l &= \lim_{q\to 0^+} \frac{C(q,q)}{q} \\
&= \lim_{q\to 0^+} \frac{\delta}{(\frac{\delta}{q}-(\delta-1))^{2^{1/\theta}}-(\delta-1)}\,\frac{1}{q} \\
&= \lim_{q\to 0^+} \frac{\delta}{(\delta-q(\delta-1))^{2^{1/\theta}}-q^{2^{1/\theta}}(\delta-1)}\,\frac{1}{q^{1-2^{1/\theta}}} \\
&= \lim_{q\to 0^+} \frac{\delta\, q^{2^{1/\theta}-1}}{(\delta-q(\delta-1))^{2^{1/\theta}}-q^{2^{1/\theta}}(\delta-1)} \\
&= \frac{\delta\, 0^{2^{1/\theta}-1}}{(\delta-0(\delta-1))^{2^{1/\theta}}-0^{2^{1/\theta}}(\delta-1)} \\
&= \frac{\delta\, 0^{2^{1/\theta}-1}}{\delta^{2^{1/\theta}}} \\
&= 0^{**}
\end{aligned}
$$

$2^{1/\theta}-1>0$ since $1\le\theta<\infty$

A.2 Kendall's Tau Rank Correlation

By using equation (3.1), theoretical Kendall's Tau rank correlation can be calculated for every Archimedean copula. For modified Gumbel copula, $\varphi'(t)$ shall be calculated and then the integral can be solved.

$$\varphi(t) = (ln(\frac{\delta}{t} - (\delta - 1)))^{\theta}$$

$$\varphi'(t) = \theta(ln(\frac{\delta}{t} - (\delta - 1)))^{\theta-1} \frac{1}{\frac{\delta}{t} - (\delta - 1)} \frac{\delta}{-t^2}$$

$$\frac{\varphi(t)}{\varphi'(t)} = \frac{ln(\frac{\delta}{t} - (\delta - 1))\ (\frac{\delta}{t} - (\delta - 1))\ (-t^2)}{\theta\ \delta}$$

$$\begin{aligned}
\tau_C &= 1 + 4\int_0^1 \frac{\varphi(t)}{\varphi'(t)}\, dt \\
&= 1 + 4\int_0^1 \frac{ln(\frac{\delta}{t} - (\delta - 1))\ (\frac{\delta}{t} - (\delta - 1))\ (-t^2)}{\theta\ \delta}\, dt \\
&= 1 + 4\frac{-\delta^2\ ln(\delta)\ +\ \delta\ -\ 1}{6(\delta - 1)^2\ \theta} \\
&= \begin{cases} 1 + 4\frac{-\delta^2\ ln(\delta)\ +\ \delta\ -\ 1}{6(\delta-1)^2\ \theta} & \delta \neq 1 \\ 1 - \frac{1}{\theta}^{**} & \delta = 1 \end{cases}
\end{aligned}$$

**When $\delta = 1$, limit needs to be applied. That is,

$$\begin{aligned}
\lim_{\delta \to 1}(1 + 4\frac{-\delta^2\ ln(\delta)\ +\ \delta\ -\ 1}{6(\delta - 1)^2\ \theta}) &= 1 + \lim_{\delta \to 1}(4\frac{-\delta^2\ ln(\delta)\ +\ \delta\ -\ 1}{6(\delta - 1)^2\ \theta}) \\
&= 1 + \lim_{\delta \to 1}(4\frac{-2\delta\ ln(\delta) - \delta + 1}{12(\delta - 1)\ \theta}) \\
&= 1 + \lim_{\delta \to 1}(4\frac{-2 - 2\ ln(\delta) - 1}{12\ \theta}) \\
&= 1 - \frac{1}{\theta}
\end{aligned}$$

References

Abdi H (2007). The Kendall Rank Correlation Coefficient. in Salkind N, (2007) *Encyclopedia of Measurement and Statistics*. Sage Publications, Thousand Oaks, CA, USA. Available from: www.utdallas.edu/ herve/Abdi-KendallCorrelation2007-pretty.pdf

Adelman, I. (1963). An Econometric Analysis of Population Growth. *The American Economic Review*, 53(3), 314-339.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), pp.716-723.

Asteriou, D. and Hall, S. (2011). "ARIMA Models and the Box–Jenkins Methodology". *Applied Econometrics*. Palgrave Macmillan Ltd. pp. 265–286.

Bacinello, A. and Persson, S. (n.d.). Design and Pricing of Equity-Linked Life Insurance Under Stochastic Interest Rates. *SSRN Electronic Journal*.

Bernard, C. and Lemieux, C. (2008). Fast simulation of equity-linked life insurance contracts with a surrender option. *2008 Winter Simulation Conference*.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, 31(3), pp.307-327.

Box, G. and Jenkins, G. (1976). *Time series analysis*. San Francisco: Holden-Day.

Box, G. and Pierce, D. (1970). Distribution of Residual Autocorrelations in Autoregressive-Integrated Moving Average Time Series Models. *Journal of the American Statistical Association*, 65(332), pp.1509-1526.

Brenner, M. (2005). Commentary: Economic growth is the basis of mortality rate decline in the 20th century–experience of the United States 1901-2000. *International Journal of Epidemiology*, 34(6), pp.1214-1221.

Carter, L. (1996). FORECASTING U.S. MORTALITY:. A Comparison of Box-Jenkins ARIMA and Structural Time Series Models. *Sociological Quarterly*, 37(1), pp.127-144.

Carter, L. and Lee, R. (1992). Modeling and forecasting US sex differentials in mortality. *International Journal of Forecasting*, 8(3), pp.393-411.

Chen, C. and Liu, L. (1993). Joint Estimation of Model Parameters and Outlier Effects in Time Series. *Journal of the American Statistical Association*, 88(421), pp.284-297.

Chen, X. and Fan, Y. (2006). Estimation and model selection of semiparametric copula-based multivariate dynamic models under copula misspecification, *Journal of Econometrics*, 135, pp.125-154.

Daniel, W. (1990). "Spearman rank correlation coefficient". *Applied nonparametric statistics*. Boston: PWS-KENT. pp. 358–365

Demarta, S. and McNeil, A. (2005). The t Copula and Related Copulas. *International Statistical Review*, 73(1), pp.111-129.

Ettner, S. (1996). New evidence on the relationship between income and health. *Journal of Health Economics*, 15(1), pp.67-85.

Genest, C. and MacKay, J. (1986). The Joy of Copulas: Bivariate Distributions with Uniform Marginals. *The American Statistician*, 40(4), p.280.

Genest, C., Kojadinovic, I., Nešlehová, J. and Yan, J. (2011). A goodness-of-fit test for bivariate extreme-value copulas. *Bernoulli*, 17(1), pp.253-275.

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

Hoaglin, D., Mosteller, F. and Tukey, J. (1983). *Understanding robust and exploratory data analysis*. New York: Wiley. pp.404–414.

Jamison, E., Jamison, D. and Hanushek, E. (2007). The effects of education quality on income growth and mortality decline. *Economics of Education Review*, 26(6), pp.771-788.

Kalemli-Ozcan, S. (2002). Does the mortality Decline Promote Economic Growth?. *Journal of Economic Growth*, 7(4), pp.411-39.

Kendall, M. (1938). A New Measure of Rank Correlation. *Biometrika*, 30(1-2), pp.81-93.

Li , S. and Chan, W. (2005). Outlier analysis and mortality forecasting: The United Kingdom and Scandinavian countries. *Scandinavian Actuarial Journal*, 2005(3), pp.187-211.

Li, S. and Chan, W. (2007). The Lee-Carter Model for Forecasting mortality, Revisited. *North American Actuarial Journal*, 11(1), pp.68-89.

McNeil, A., Frey, R. and Embrechts, P. (2015). *Quantitative risk management*. Princeton: Princeton University Press. pp.184-234.

Oehlert, G. (1992). A Note on the Delta Method. *The American Statistician*, 46(1), p.27.

Patton, A. (2012). A review of copula models for economic time series. *Journal of Multivariate Analysis*, 110, pp.4-18.

Preston, S. (2007). The changing relation between mortality and level of economic development. *International Journal of Epidemiology*, 36(3), pp.484-490.

Pritchett, L. and Summers, L. (1996). Wealthier is Healthier. *The Journal of Human Resources*, 31(4), p.841.

Remillard, B. (2010). Goodness-of-Fit Tests for Copulas of Multivariate Time Series. *SSRN Electronic Journal*.

Ribeiro, R., & Pietro, V. d. (2009). Longevity risk and portfolio allocation. *Global Asset Allocation & Alternative Investments*.

Rummel, R. J. (n.d.). *Understanding correlation*. Retrieved from http://www.mega.nu:8080/ampp/rummel/uc.htm

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, 6(2), pp.461-464.

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

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15(1), p.72.

Wu, F., Valdez, E. and Sherris, M. (2007). Simulating from Exchangeable Archimedean Copulas. *Communications in Statistics - Simulation and Computation*, 36(5), pp.1019-1034.