Models for Crop Insurance Indemnities

Yuma Mizushima

April 20, 2020

Background

Introduction

- From 2007 to 2016, the federal crop insurance title had the second-largest outlays in the farm bill, after nutrition
- The total net cost of the program for crop years 2007-2016 was about $72 billion (Rosa, Isabel)

Why Crop Insurance?

- Financially protects farmers from loss of crop and revenue
- All consumers benefit from a secure agriculture industry
- Ratemaking and reserving are important problems in actuarial science
- Dependence models may influence the reserving of insurance companies

Causes of Loss

- Weather: rain, temperature, length of growing season
- Bacteria, viruses, pests
- Implies that insurance amounts may differ between regions and peril types
- Different causes of losses may influence the dependence of policyholders

What is a Liability?

- A measure of the insurer's exposure to loss
- Represents the total insured value of the crop

What is an Indemnity?

- A safeguard against loss
- In terms of US crop insurance: a payment made when crop yields under-perform
- USDA RMA authorizes 15 private companies to provide insurance

The Goal

- To predict indemnity amounts for specific farms based on region, commodity grown, and peril types
- To illustrate that spatial dependence matters, in terms of the aggregate loss, and should influence the operation of an insurance company
- To show that the risk capital that insurance companies should prepare will be influenced by this dependence

Exploratory Data Analysis

Importing Data

```
tpuin <- read.csv("tpuin.csv", stringsAsFactors=FALSE)
tpuout <- read.csv("tpuout.csv", stringsAsFactors=FALSE)
tpuinall <- tpuin[tpuin$LiabilityAmount>0 & tpuin$StateAbbr=="MI" & !(is.na(tpuin$LiabilityAmount)), ]
som18 <- read.delim("colsom18.txt", header=FALSE, sep="|", stringsAsFactors=FALSE)
som19 <- read.delim("colsom19.txt", header=FALSE, sep="|", stringsAsFactors=FALSE)
names(som18) <- names(som19) <- c("Year","StateCode","State","CountyCode","County","CropCode","Crop",
    "InsurancePlanCode","InsurancePlan","CoverageCategory","StageCode","CauseCode","Cause","Month",
    "MonthAbbr","PoliciesEarningPremium","PoliciesIndemnified","NetPlantedAcres","NetEndorsedAcres",
    "Liability","Premium","Subsidy","DeterminedAcres","Indemnity","LossRatio")
```

Overview of Data Aggregated by State

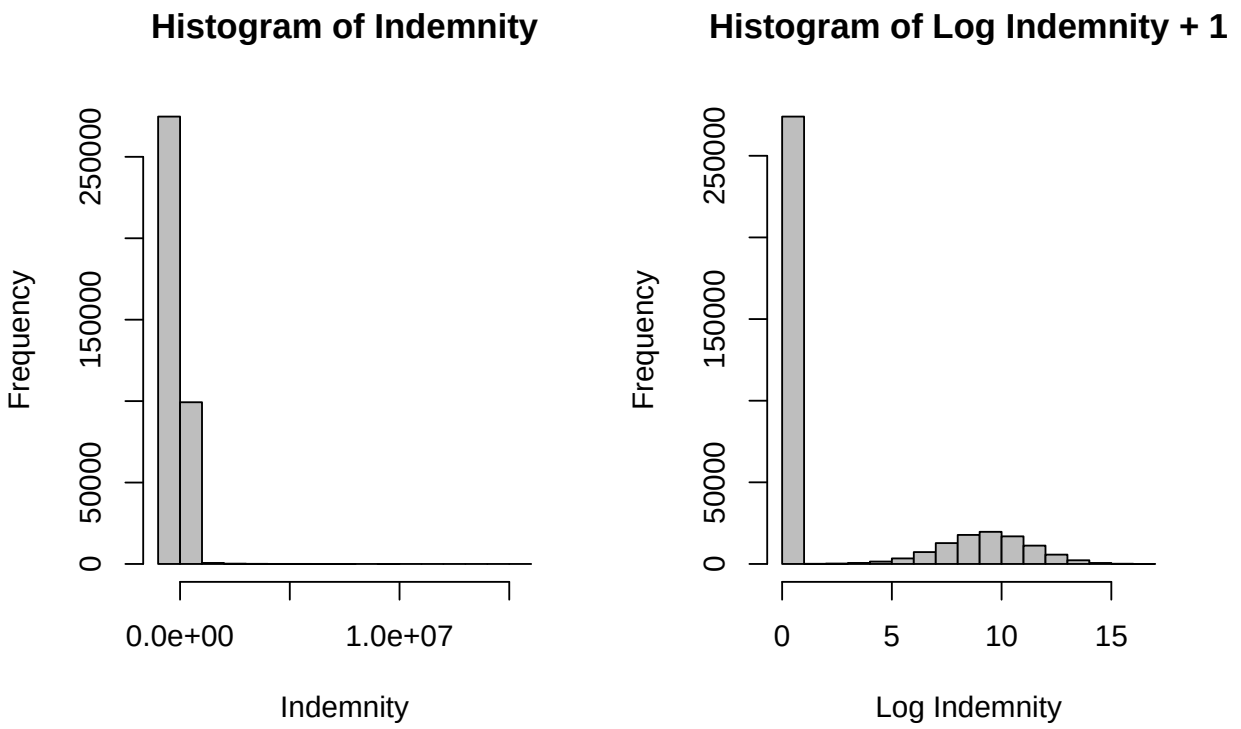
```
library(dplyr)
tpuin %>% filter(StateAbbr %in% c("MI", "OH", "FL", "CA", "NY")) %>% group_by(StateAbbr) %>%
    summarize(Farms = length(LiabilityAmount[!is.na(LiabilityAmount)]), Indemnity.Amount =
    sum(IndemnityAmount), Liability.Amount = sum(LiabilityAmount, na.rm = TRUE), Total.Zeros =
    sum(1*(IndemnityAmount==0)))
```

```
## # A tibble: 5 x 5
##   StateAbbr Farms Indemnity.Amount Liability.Amount Total.Zeros
##   <chr>     <int>            <dbl>            <dbl>       <dbl>
## 1 CA        10468        303183746       8492481313        8143
## 2 FL         5913        347854824       2824855303        3847
## 3 MI         8804         90308988       1932773671        6506
## 4 NY         3927         24691936        622665816        3063
## 5 OH         8436         56665925       3289548279        6318
```

Using the crop insurance dataset, I aggregated the indemnity amounts as well as the liability amounts based on the state. The table displays statistics for a few individual states: the number of farms, total indemnity amounts, and total liability amounts. It's important to note that each of these states have several zeros as observations, implying that a significant proportion of the farms don't receive anything from their insurance companies.

Histogram of Indemnity Amount

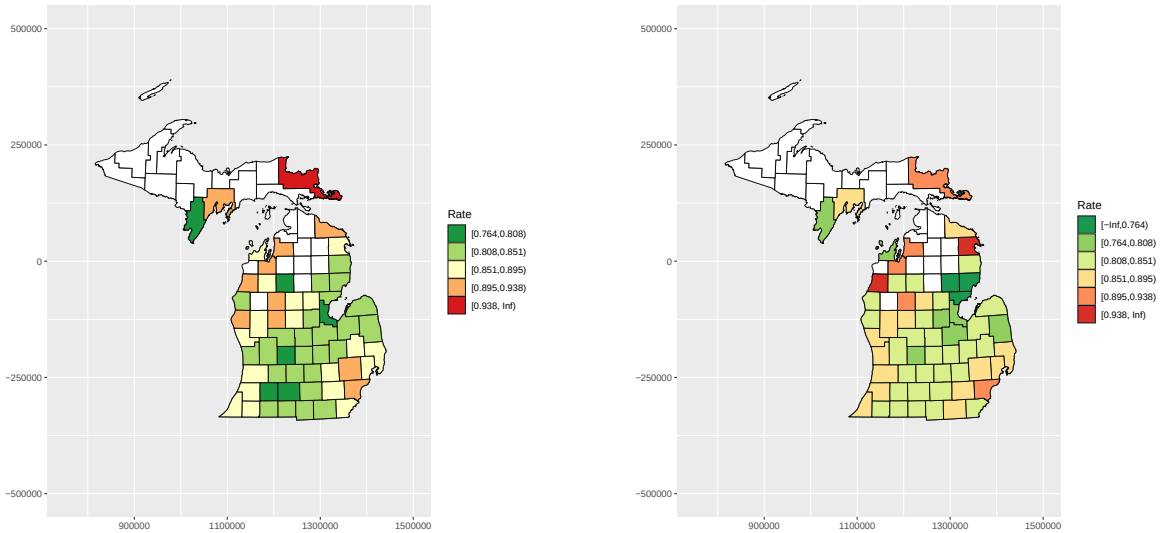
```
par(mfrow=c(1,2))
hist(tpuin$IndemnityAmount, xlab = "Indemnity", main = "Histogram of Indemnity", col = "gray")
hist(log(tpuin$IndemnityAmount+1), xlab = "Log Indemnity", main = "Histogram of Log Indemnity + 1",
     col = "gray", breaks = 20)
```



Unsurprisingly, a histogram of the indemnity amounts is heavily skewed right and has a lot of mass at zero. To better capture the distribution of the data, I did a log transformation of the indemnity amounts `+ 1`. The `+ 1` is necessary to consider the zero values in our modeling. In the right figure, we see a spike at zero and probability mass at other positive numbers.

The Poisson sum of the positive indemnities will follow a Tweedie distribution, where the aggregate loss could either be zero or a positive number. These figures motivate the use of the Tweedie distribution to transform the observations into residuals.

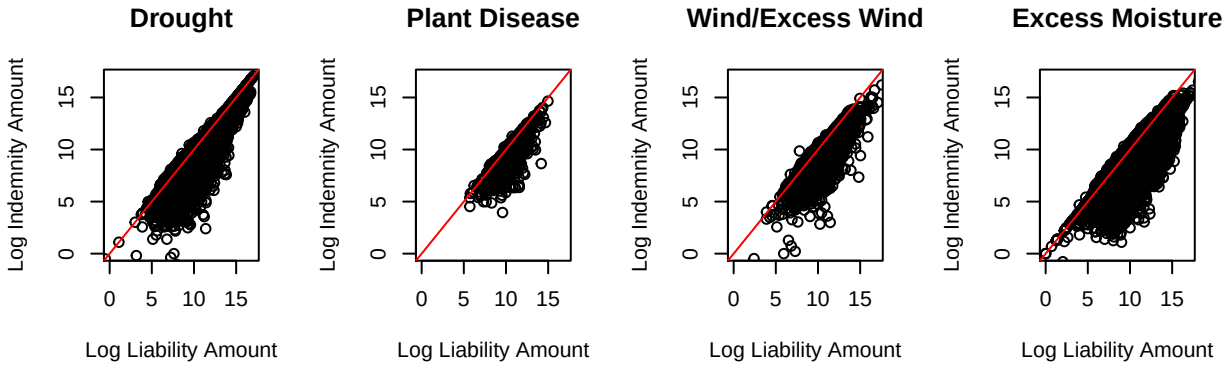
Relativity Map in Michigan (Training and Validation Sample)



The indemnity amounts are aggregated based on the state and county. Using Michigan as an example, the rate map on the left looks at the in-sample data, while the right figure looks at the out-of-sample data. We can see that the counties in which indemnity amounts were large, tend to have high indemnities in the next year as well. We see roughly the same areas highlighted in the validation sample, so the correlation is visually clear. These figures justify the use of the in-sample data as our model for the prediction task.

Indemnities and Liabilities by Peril Type

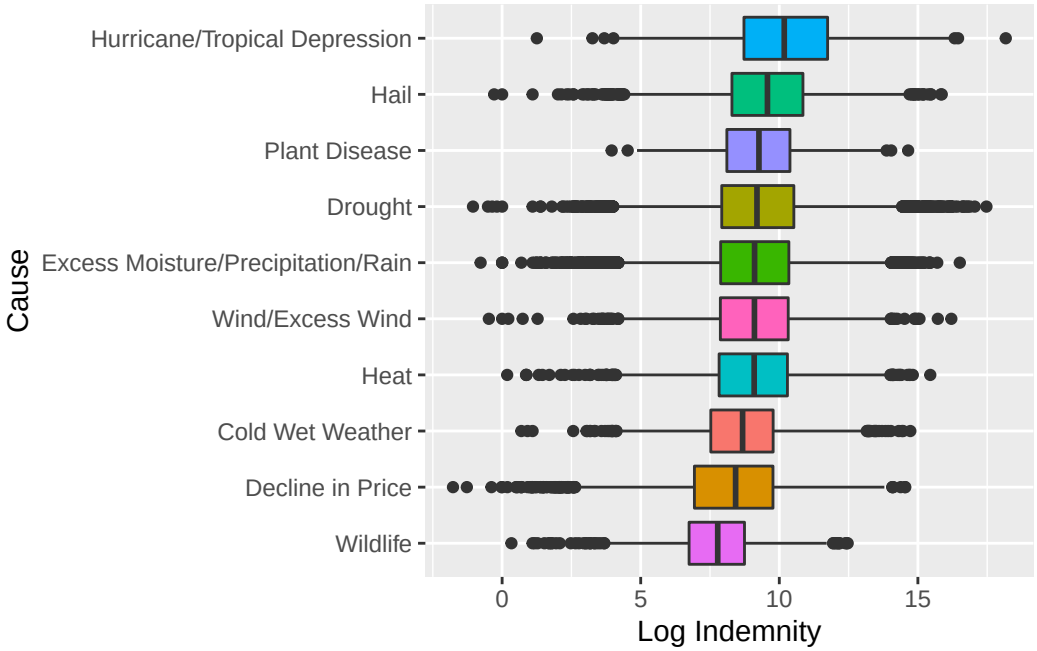
```
par(mfrow=c(1,4))
plot(log(subset(som18,Indemnity>0&Cause=="Drought")$Liability),
     log(subset(som18,Indemnity>0&Cause=="Drought")$Indemnity), xlim=c(0,17), ylim=c(0,17),
     xlab="Log Liability Amount", ylab="Log Indemnity Amount", main="Drought"); abline(0,1,col="red")
plot(log(subset(som18,Indemnity>0&Cause=="Plant Disease")$Liability),
     log(subset(som18,Indemnity>0&Cause=="Plant Disease")$Indemnity), xlim=c(0,17), ylim=c(0,17),
     xlab="Log Liability Amount", ylab="Log Indemnity Amount", main="Plant Disease"); abline(0,1,col="r
plot(log(subset(som18,Indemnity>0&Cause=="Wind/Excess Wind")$Liability),
     log(subset(som18,Indemnity>0&Cause=="Wind/Excess Wind")$Indemnity), xlim=c(0,17), ylim=c(0,17),
     xlab="Log Liability Amount", ylab="Log Indemnity Amount", main="Wind/Excess Wind"); abline(0,1,col
plot(log(subset(som18,Indemnity>0&Cause=="Excess Moisture/Precipitation/Rain")$Liability),
     log(subset(som18,Indemnity>0&Cause=="Excess Moisture/Precipitation/Rain")$Indemnity), xlim=c(0,17)
     xlab="Log Liability Amount", ylab="Log Indemnity Amount", main="Excess Moisture"); abline(0,1,col=
```



Here are some scatterplots of the log indemnity amount over the log liability amount for different causes of loss, such as drought, plant disease, wind, and excess moisture. There is very good correlation for all four peril types.

Log Indemnity by Peril Type

```
library(ggplot2)
ggplot(subset(som18,Indemnity>0&Cause %in% c("Drought", "Plant Disease", "Wind/Excess Wind",
      "Excess Moisture/Precipitation/Rain", "Hurricane/Tropical Depression", "Drought",
      "Decline in Price", "Hail", "Heat", "Cold Wet Weather", "Wind/Excess Wind", "Wildlife")),
      aes(x=reorder(Cause,Indemnity,median), y=log(Indemnity),fill=Cause)) + geom_boxplot() +
      coord_flip() + ylab("Log Indemnity") + xlab("Cause") + theme(legend.position = "none")
```



This is a boxplot of log indemnities by peril type, where we have 10 of the most common causes of insurance claims in our dataset. Hurricane/tropical depression cases had significantly different indemnity amounts compared to wildlife cases. This leads us to believe that each peril type has a different loss profile and therefore a different dependence structure.

Models

Tweedie Model

Introduction

A random variable Y follows a Tweedie distribution if Y follows exponential dispersion models with mean and variance:

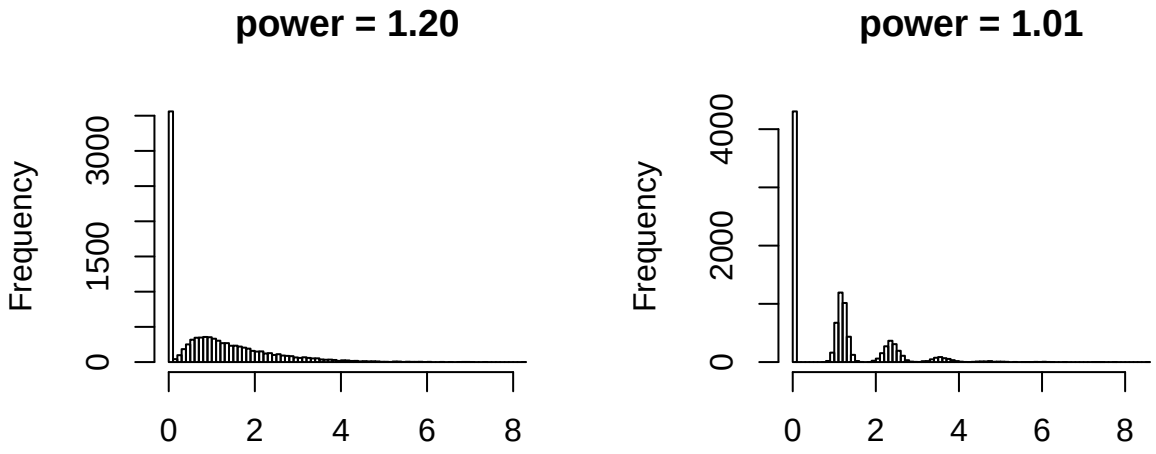
$$E(Y) = \mu \quad Var(Y) = \sigma^2 \mu^p$$

where p is called the Tweedie power parameter

If $p = 1$, our Tweedie distribution is exactly a Poisson distribution
If $p = 2$, it is exactly a Gamma distribution
If $1 < p < 2$, it will become a compound Poisson/Gamma distribution

Shape of Tweedie Distribution

```
library(tweedie)
par(mfrow=c(1,2))
hist(rtweedie(10000, mu = 1, phi = 1.2, xi = 1.2), breaks = 100, main = "power = 1.20", xlab = NA)
hist(rtweedie(10000, mu = 1, phi = 1.2, xi = 1.01), breaks = 100, main = "power = 1.01", xlab = NA)
```



We plot two examples of the Tweedie distribution. When power parameter $p = 1.20$, we have a huge zero mass followed by positive values. When $p = 1.01$, we still have zero mass but we can see the residuals change significantly. As p approaches exactly 1, we can observe a roughly discrete Poisson distribution, having point masses at the natural numbers.

We can look for a p value within the region $1 < p < 2$ such that the Tweedie distribution fits our data well.

Expression of Y

$$T \sim Poisson(\lambda)$$
$$X_i \sim Gamma(\alpha, \beta)$$
$$Y = \sum_{i=1}^{T} X_i$$

The frequency of indemnity $\sim Poisson(\lambda)$
The severity of indemnity $\sim Gamma(\alpha, \beta)$

A Tweedie distribution results when we assume the claim frequencies follow a Poisson distribution and the severities follow a Gamma distribution. We are combining frequency and severity for simplicity's sake, as it's easier to transform the observations into residuals with just one distribution. Here, we can use Tweedie directly modeling the aggregate of the indemnity amounts.

Copula Model

Sklar's Theorem

Sklar's theorem states that an m-dimensional copula is a function $\boldsymbol{C}$ from the unit $\boldsymbol{m}$-cube $[0,1]^{\boldsymbol{m}}$ to the unit interval $[0,1]$ which satisfies the following conditions:

1. $C(1, ..., 1, a_n, 1, .., 1) = a_n$ for every $n \leq m$ and all a_n in $[0,1]$;
2. $C(a_1, ..., a_m) = 0$ if $a_n = 0$ for any $n \leq m$;
3. C is m-increasing

Dependence Parameter

Copulas allow researchers to study the dependence between two separate but related issues.

$$\boldsymbol{F}(\boldsymbol{y}_1, ..., \boldsymbol{y}_m) = C(\boldsymbol{F}_1(y_1), ...,\boldsymbol{F}_m(y_m); \boldsymbol{\theta})$$

where θ is a parameter which measures dependence between the marginals.

Parameter Matrix

Assume the response vector is $\boldsymbol{y}$. Then we imagine there is a $d \times d$ dimensional parameter matrix $\boldsymbol{\Sigma}$, such that

$$\sigma_{i,j} = \begin{cases} 1 & \text{if } i = j \\ \rho & \text{if } s_i = s_j \\ 0 & \text{otherwise} \end{cases}$$

where s_i is the state code for the ith county, and s_j is the state code for the jth county.

If we have 1000 counties, there would be a 1000 $\times$ 1000 matrix, with the diagonals being 1s, as the same counties are perfectly correlated. Otherwise, if there's a different county, then you're either within the state or not within the state. All the pairs that are in the same state will have parameter p and the other cells will have 0. The p here is very close to the dependence parameter when estimating the copula function.

Gaussian Copula

The multivariate Gaussian copula of dimension d is defined by

$$C(\boldsymbol{u}) = \Phi_d(\Phi^{-1}(u_1), ..., \Phi^{-1}(u_d); \boldsymbol{\Sigma})$$

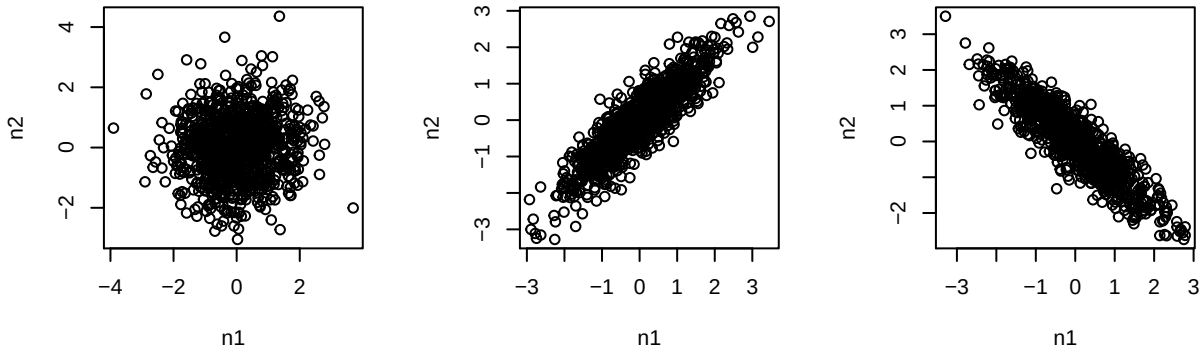
where Φ^{-1} is the inverse cumulative distribution function of the standard normal, and Φ_d is the joint cumulative distribution function of a multivariate normal distribution with mean vector zero, and covariance matrix $\boldsymbol{\Sigma}$.

Essentially, a Uniform variable can be transformed into any random variable by inverse transforming it using its cumulative distribution function (CDF). A copula is simply a joint distribution of Uniform random variables that gives the dependence among every pair defined by the correlation matrix $\boldsymbol{\Sigma}$.

Copula Function Demonstration

We use the Copula function to construct three groups of data. Within each group of data, there are two normally distributed variables. The covariances in each group are 0.1, 0.9, and -0.9 respectively.

```
library(copula)
par(mfrow=c(1,3))
# Observations of pairs of Uniform variables
uu <- rCopula(1000, normalCopula(0.1)) # Dependence parameter of .1
n1 <- qnorm(uu[,1], mean=0, sd=1) # Transform each column to Normal
n2 <- qnorm(uu[,2], mean=0, sd=1)
plot(n1,n2) # Bivariate normally distributed pair
uu <- rCopula(1000, normalCopula(0.90)) # Dependence parameter of .9
n1 <- qnorm(uu[,1], mean=0, sd=1)
n2 <- qnorm(uu[,2], mean=0, sd=1)
plot(n1,n2)
uu <- rCopula(1000, normalCopula(-0.90)) # Dependence parameter of -.9
n1 <- qnorm(uu[,1], mean=0, sd=1)
n2 <- qnorm(uu[,2], mean=0, sd=1)
plot(n1,n2)
```



$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y)$$

Let's just imagine that X is the price of a car, while Y is the price of a house. When there's low covariance, both prices are distributed randomly, like you see in the leftmost graph. If they have high dependence, when the price of the car goes up, the price of a house goes up as well. Fitting a Copula model would help attain the standard errors of the parameters of the Copula function to indicate dependence.

Analysis

Process

1. Aggregate the information we need to generate a table
2. Use the `tweedie.profile()` function to generate the **MLE** of the Tweedie index power parameter
3. Construct the exact PDF of Tweedie distribution
4. Obtain the Cox-Snell residuals
5. Obtain the covariance

```
# Pre-processing
tpuin$StateCountyCodeStr <- ifelse(tpuin$StateCountyCodeStr<=9999,paste("0",tpuin$StateCountyCodeStr,
    sep=""),tpuin$StateCountyCodeStr) # Appending the County Code after State Code
datin <- aggregate(IndemnityAmount ~ StateCountyCodeStr, data=tpuin, FUN=sum)
datin <- merge(datin, aggregate(LiabilityAmount ~ StateCountyCodeStr, data=tpuin, FUN=sum),
    by="StateCountyCodeStr")
nrow(datin) # 2360 unique counties
str(datin)
datin$StateCode <- substr(datin$StateCountyCodeStr,1,2) # Separate out the state code and county code
datin$CountyCode <- substr(datin$StateCountyCodeStr,3,5)
head(datin) # Now we have the dataframe we're gonna use for the model
summary(datin$IndemnityAmount) # We can see what the data looks like
plot(log(datin$LiabilityAmount+1),log(datin$IndemnityAmount+1)) # Check for positive relationship

# Maximum likelihood of Tweedie
library(tweedie)
out <- tweedie.profile(IndemnityAmount~log(LiabilityAmount), data=datin, xi.vec=seq(1.8, 1.99,
    length=10), do.plot=TRUE) # Find the shape parameter for the Tweedie
fittw <- glm(IndemnityAmount~log(LiabilityAmount), data=datin, family=tweedie(var.power=out$xi.max,
    link.power=0)) # Fits a Tweedie model to the data, and figure out mean of Tweedie distribution
summary(fittw) # Coefficient for log liability is .746

# Transform observations into Uniform
datin$uu <- ptweedie(datin$IndemnityAmount, xi=out$xi.max, mu=fittw$fitted.values,
    phi=summary(fittw)$dis) # Transform observations using CDF of Tweedie to get Uniform r.v. uu
hist(datin$uu,100) # Plot should be uniform if the fit is good enough
hist(log(datout$IndemnityAmount),100) # See if the fit is good
lines(density(rtweedie(1000, xi=out$xi.max, mu=fittw$fitted.values)))

# Take the transformed uniform observations and take the pairwise observations within counties
uupair <- NULL
for (sc in unique(datin$StateCode)) {
  m <- nrow(datin[datin$StateCode==sc,])
  if(m>1) {
    u1 <- datin[datin$StateCode==sc,][t(combn(1:m,2))[,1],"uu"]
    u2 <- datin[datin$StateCode==sc,][t(combn(1:m,2))[,2],"uu"]
    uupair <- rbind(uupair,cbind(u1,u2))
  }
}
uupair # Pairwise observations
cor(uupair,method="spearman") # 0.493118
```

Results

Covariance

2 policyholders in the same state have correlations of 0.4931.

For specific peril types, 2 policyholders in the same state have correlations of:

Drought	Plant Disease	Wind	Precipitation	Hurricane
0.4581	0.3214	0.3116	0.3406	0.0472

Covariances indicate dependence for those policyholders within the same state. The Spearman correlation has enough evidence to show that there is significant dependence.

Conclusion

Implications

If there are multiple farms in the same state, it may be safer for an insurance company to not insure them all since the risk will be correlated. An in-state correlation implies that it would be advantageous to diversify by scattering a portfolio over different states. Insurance companies located in one state may consider re-insuring their risk by paying a premium to other companies for coverage.

The risk capital held by an insurance company should be different depending on the peril types being covered, as each type has a different dependence structure. Underwriting strategies and loss-reserving practices may also be influenced by this.

References

Josephson, G.R., Lord, R., & Mitchell, C.W. (2000). Actuarial Documentation of Multiple Peril Crop Insurance Ratemaking Procedures.

Risk Management Agency. (n.d.). State/County/Crop Summary of Business. Retrieved from https://www.rma.usda.gov/Information-Tools/Summary-of-Business/State-County-Crop-Summary-of-Business

Risk Management Agency. (n.d.). Cause of Loss Historical Data Files. Retrieved from https://www.rma.usda.gov/Information-Tools/Summary-of-Business/Cause-of-Loss

Rosa, I. (2018, May 10). Federal Crop Insurance: Program Overview for the 115th Congress. Retrieved from https://www.everycrsreport.com/reports/R45193.html

Trivedi, P. K., & Zimmer, D. M. (2005). Copula modeling: An Introduction For Practitioners. Foundations and Trends in Econometrics.

Acknowledgements

I would like to thank my supervisor Dr. Gee Lee for his patient guidance, and my group members Wenhao Wu, Yikang Li, and Yue Zhang for their collaboration. Special thanks to Dr. Jeanne Wald for allowing me to participate in a research project as part of the Exchange Program at MSU's Department of Mathematics.