Modeling with Mixture of the Popular Models

Min Deng
Department of Actuarial Science and Mathematics
Maryville University at St. Louis

Abstract:
Modeling based the data information is one of the most important tasks in actuarial science. The purpose of building mathematical models is to predict and forecast the future loss or cost in insurance industry. In this paper we are going to develop some models by use the mixture of the popular distributions. Some suggested use of the models to the loss data or other actual data will be discussed and examples to illustrate the models will be given.

Key words: Modeling, Mathematical Models, Probability Distribution, Mixture Distribution,

1. Introduction

It is very important to construct the mathematical model base on the real data from industry especially in the finance and insurance industry. Since from the construct model, we will gain and improve the understanding of the real-world system of interest. We will able to predict or simulate the real-world system in order to avoid risk or transfer risk or reduce the impact of the risk, or increase the competitive advantage.

There are two types of the mathematical probability distributions (models).

- One is a data-dependent distribution. The distribution depend on data, the number of parameters are not fix. The larger the data set, the larger the number of parameters. The most popular data-dependent distributions in the insurance industry are the empirical distribution, Nelson Aalen estimator, and Kaplan-Meier product-limit estimator.
- Another type is a parameter distribution. The distribution is determined by one or more parameters. The number of parameters is fix and finite. There are many parameter distributions, such as binomial, normal, exponential, lognormal, pareto etc.

Among all continuous parametric distributions, the normal is the most popular and widely used distribution. Because of according to the Central Limit Theorem, when sample size is large enough, the sample mean will follow the normal probability distribution. In general insurance industry, the loss amount usually consists of the data with small amount with high frequencies and larger losses with low frequencies. In other word insurance losses follow the skewed and heavy tailed distribution. Insurance industry needs to pay attention to the small losses due to the large frequencies and also need to pay attention to larger losses even those losses with a very small probability due to large impact. Therefore normal distribution will be not good choice for the insurance losses due to quick fade to zero in the tail. Therefore, besides to use the heavy tailed distributions to model the insurance losses, such as LogNormal or Pareto, many researches consider the composite distribution. For example, Sandra Teodorescu and Raluca Vernic (2006), (2009), Klugman (2008), David Scollnik and ChenchenSun (2012), and etc.

In this paper we are going to introduce the composite models which consist of the data dependent distribution and parametric distribution. The proposed composite distributions

have the advantages of both a data-dependent distribution and parametric distribution. Some examples will be given to illustrate the composite distribution.

2. Background of the Data-Dependent Distribution and the Composite Distribution

2.1 Data-Dependent Distribution

Klugman, Panjer, and Willmot (2012) give the detail discuss for the data-dependent distributions as follows,

- The Empirical Distribution

The empirical distribution is obtained by assigning probability of $\frac{1}{n}$ to each data point. Therefore for the complete individual data, the empirical distribution function is as follows,

$$F(x) = \frac{number\ of\ obervations\ \leq x}{n}$$

Where n is the total number of the observations.

For the complete group data, let the group boundaries be $c_0 < c_1 < \cdots < c_k$, where often $c_0 = 0\ and\ c_k = \infty$. The number of observations falling between $c_{j-1}\ and\ c_j$ is denoted as n_j with $\sum_{j=1}^{k} n_j = n$. By the definition of the empirical distribution function, we are able to easily determine $F(c_j) = \frac{\sum_{i=1}^{j} n_i}{n}$. Let x be the loss random variable, the interpolating method is used to determine the empirical distribution function for x between $c_{j-1} < x < c_j$. Therefore, the empirical distribution function is given by

$$F(x) = \frac{c_j - x}{c_j - c_{j-1}} F(c_{j-1}) + \frac{x - c_{j-1}}{c_j - c_{j-1}} F(c_j),\ \ for\ c_{j-1} \leq x \leq c_j$$

The empirical density function is as follows,

$$f(x) = \frac{n_j}{n(c_j - c_{j-1})},\ \ for\ c_{j-1} \leq x < c_j$$

The advantage of the empirical distribution is that it is very simple to construct and it will change to follow the changes of the data set. The disadvantage is that it is not fit the small losses with high frequencies very well due to the group and loss many information data provided and it is undefined when the loss random variable x larger than the largest piece of the observation. This distribution is not able to describe the tail behavior of the loss random variable.

- Risk Set

For the incomplete data (either left truncated or right censored, for example: insurance deductible and insurance maximum benefit), in order to construct the probability distribution, we need to define the risk set first. Let $y_1 < y_2 < \cdots < y_k$ are k unique not censored values appear in the not censored observations. Let s_j be the number of time the uncensored observation y_j appears in the sample. Let r_j be the number of observations at risk. Then the risk set consists of $\{y_j, s_j, r_j\}$, where j from 1 to k.

- Nelson Aalen Estimator

Start from the either complete or incomplete data set, we are able to construct the risk set as above. Then the Nelson Aalen estimator of the cumulative hazard rate function is defined as follows,

$$H(x)=\begin{cases}0, & x<y_1\\ \sum_{i=1}^{j-1}\frac{s_i}{r_i}, & y_{j-1}\le x<y_j, j=2,\cdots,k\\ \sum_{i=1}^{k}\frac{s_i}{r_i}, & x\ge y_k\end{cases}$$

The distribution function can be obtained from

$$F(x)=1-S(x)=1-e^{-H(x)}$$

Where $S(x)$ is the survival function of the loss random variable.

- Kaplan-Meier product-limit estimator

Similar to Nelson Aalen estimator, we are going to use the same risk set to define the survival distribution function of the loss random variable as follows,

$$S(x)=\begin{cases}1, & x<y_1\\ \prod_{i=1}^{j-1}\left(\frac{r_i-s_i}{r_i}\right), & y_{j-1}\le x<y_j, j=2,\cdots,k\\ \prod_{i=1}^{k}\left(\frac{r_i-s_i}{r_i}\right) or\ 0, & x\ge y_k\end{cases}$$

It is very clear that if $s_k=r_k,$ then $S(t)=0, for\ t>y_k$. Otherwise, the survival function will be not equal to zero at infinity. This is the disadvantage of the product-limit distribution.

Both Nelson Aalen and Kaplan-Meier estimators have the difficulty to deal with the small losses with high frequencies and large losses with small probabilities.

2.2 Composite Distributions

There are many proposed composite distributions to model the insurance losses; we are going to review some of them.

- The Composite Exponential-Pareto distribution

Thodorescu and Vernic (2006) proposed the composite Exponential-Pareto as follows; Let X be the random variable with density

$$f(x)=\begin{Bmatrix}cf_1(x) & if\ 0<x<\theta\\ cf_2(x) & if\ \theta<x<\infty\end{Bmatrix}$$

Where $f_1(x)$ is an exponential density, $f_2(x)$ is a two parameters Pareto density, and c the normalizing constant. More specifically, the density functions are given as follows,

$$f_1(x)=\frac{1}{\mu}e^{-x/\mu}, \qquad x>0$$

$$f_2(x)=\frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}, \qquad x>\theta$$

Require continuity and differentiability condition at θ

$$f_1(\theta)=f_2(\theta), \qquad f_1'(\theta)=f_2'(\theta)$$

- The Composite Weibull-Pareto Model

Scollnik and Sun (2012) discussed many composite Weibull-Pareto models. One of them as follows;

Let X be the random variable with density

$$f(x) = \begin{Bmatrix} cf_1(x) & if\ 0 < x < \theta \\ cf_2(x) & if\ \theta < x < \infty \end{Bmatrix}$$

Where $f_1(x)$ is a Weibull density, $f_2(x)$ is a two parameters Pareto density, and c the normalizing constant. More specifically, the density functions are given as follows,

$$f_1(x) = \left(\frac{\tau}{x}\right)\left(\frac{x}{\varphi}\right)^{\tau} e^{-\left(x/\varphi\right)^{\tau}}, \qquad x > 0$$

$$f_2(x) = \frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}, \qquad x > \theta$$

Require continuity and differentiability condition at θ

$$f_1(\theta) = f_2(\theta), \qquad f_1'(\theta) = f_2'(\theta)$$

- The Composite Lognormal–Pareto Model

Cooray and Ananda (2005) introduced the composite Lognormal-Pareto models as follows;

Let X be the random variable with density

$$f(x) = \begin{Bmatrix} rf_1(x) & if\ 0 < x < \theta \\ (1-r)f_2(x) & if\ \theta < x < \infty \end{Bmatrix}$$

Where $f_1(x)$ is a Lognormal density, $f_2(x)$ is a two parameters Pareto density; More specifically, the density functions are given as follows,

$$f_1(x) = \frac{1}{\sqrt{2\pi}x\sigma} e^{-\left(\frac{1}{2}\left(\frac{lnx-\mu}{\sigma}\right)^2\right)}, \qquad x > 0$$

$$f_2(x) = \frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}, \qquad x > \theta$$

Require continuity and differentiability condition at θ

$$f_1(\theta) = f_2(\theta), \qquad f_1'(\theta) = f_2'(\theta)$$

3. Composite Models for Small Losses with Large Frequencies

3.1 The Composite Exponential-Empirical Model

Definition: The composite exponential-empirical model is defined and constructed as follows; Let X be the random variable with density

$$f(x) = \begin{Bmatrix} cf_1(x) & if\ 0 < x < \theta \\ f_2(x) & if\ \theta < x < \infty \end{Bmatrix}$$

Where c the normalizing constant, $f_1(x)$ follows the exponential distribution and $f_2(x)$ follows the empirical distribution which is data dependent distribution. More specifically, the density functions are given as follows,

$$f_1(x) = \frac{1}{\mu} e^{-x/\mu}, \qquad x > 0$$

$$f_2(x) == \frac{n_j}{n(c_j - c_{j-1})},\ for\ c_{j-1} \le x < c_j\ and\ j = 1, 2, \cdots k$$

Where μ is unknown parameter and we require the following condition θ

$$cF_1(\theta) = F_2(\theta)$$

This condition will ensure the total frequencies for losses below the threshold θ are same as the total probability below the threshold θ.

Density and Value of c: It is not hard to show that $f(x)$ is the density function. It satisfies the two conditions;

- $f(x) \geq 0 \; for \; all \; x$
- $\int_0^\infty f(x)dx = \int_0^\theta cf_1(x)dx + \int_\theta^\infty f_2(x)dx =$
$= cF_1(\theta) + S_2(\theta) = cF_1(\theta) + \big(1 - F_2(\theta)\big) = 1$

Now, we are going to determine the constant c. For the complete group data, let the group boundaries be $c_0 < c_1 < \cdots < c_k$, where often $c_0 = 0 \; and \; c_k = \infty$. Assume that the threshold θ is in one of the interval, for example, $c_{j-1} \leq \theta < c_j$, due to the required condition

$$cF_1(\theta) = F_2(\theta)$$

So we have

$$c\left(1 - e^{-\theta/\mu}\right) = \frac{c_j - \theta}{c_j - c_{j-1}} F\big(c_{j-1}\big) + \frac{\theta - c_{j-1}}{c_j - c_{j-1}} F\big(c_j\big)$$

Therefore,

$$c = \frac{e^{\theta/\mu}}{e^{\theta/\mu} - 1}\left(\frac{c_j - \theta}{c_j - c_{j-1}} F\big(c_{j-1}\big) + \frac{\theta - c_{j-1}}{c_j - c_{j-1}} F\big(c_j\big)\right)$$

Where $F\big(c_{j-1}\big) = \frac{\sum_{i=1}^{j-1} n_i}{n}$ and $F\big(c_j\big) = \frac{\sum_{i=1}^{j} n_i}{n}$.

Example 3.1: The following observations represent payments on 422 claims from an insurance policy. Data are artificial.

Payment Range	Number of Payments
0-7,500	299
7,500-17,500	45
17,500-32,500	26
32,500-67,500	24
67,500-125,000	16
125,000-300,000	9
Over 300,000	3

Select Let $\theta = 7{,}500$ and let x be the payment random variable follows the composite exponential-empirical model. Then the density function of x is given by

$$f(x) = \begin{cases} \frac{299}{422}\left(\frac{e^{7{,}500/\mu}}{e^{7{,}500/\mu} - 1}\right)\frac{1}{\mu}e^{-x/\mu}, & 0 < x \leq 7{,}500 \\ \frac{45}{422(10{,}000)}, & 7{,}500 \leq x < 17{,}500 \\ \frac{26}{422(15{,}000)}, & 17{,}500 \leq x < 32{,}500 \\ \frac{24}{422(35{,}000)}, & 32{,}500 \leq x < 67{,}500 \\ \frac{16}{422(57{,}500)}, & 67{,}500 \leq x < 125{,}000 \\ \frac{9}{422(175{,}000)}, & 125{,}000 \leq x < 300{,}000 \\ Not\ Define, & x \geq 300{,}000 \end{cases}$$

Even the data is artificial, but it has the properties of the real data. Use the exponential to model the losses small than 7,500 is better than to assume the uniform. We still need to consider the large loss with small probability part. By using the empirical distribution, the density is not defined. We are going to discuss this situation late.

3.2 The Composite Exponential-Nelson Aalen Model

Definition: The composite exponential-Nelson Aalen model is defined and constructed as follows; Let

Let X be the random variable with density

$$f(x) = \begin{cases} cf_1(x) & if\ 0 < x < \theta \\ f_2(x) & if\ \theta < x < \infty \end{cases}$$

Where c the normalizing constant, $f_1(x)$ follows the exponential distribution and $f_2(x)$ follows the Nelson Aalen estimator which is data dependent distribution. More specifically, the density functions are given as follows,

$$f_1(x) = \frac{1}{\mu} e^{-x/\mu}, \qquad x > 0$$

And $f_2(x)$ with the distribution function $F_2(x)$ explicitly as follows;
For a given insurance data set (complete or modified data), the risk set can be defined as the $\{y_j, s_j, r_j\}$, where j from 1 to k. Then the Nelson Aalen estimator of the cumulative hazard rate function is defined as follows,

$$H(x) = \begin{cases} 0, & x < y_1 \\ \sum_{i=1}^{j-1} \frac{s_i}{r_i}, & y_{j-1} \le x < y_j, j = 2, \cdots, k \\ \sum_{i=1}^{k} \frac{s_i}{r_i}, & x \ge y_k \end{cases}$$

The distribution function can be obtained from

$$F_2(x) = 1 - S_2(x) = 1 - e^{-H(x)}$$

Where $S_2(x)$ is the survival function of the loss random variable and μ is unknown parameter and we require the following condition θ

$$cF_1(\theta) = F_2(\theta)$$

This condition will ensure the total frequencies for losses below the threshold θ are same as the total probability below the threshold θ.

Density and Value of c: It is not hard to show that $f(x)$ is the density function. It satisfies the two conditions;

- $f(x) \ge 0\ for\ all\ x$
- $\int_0^\infty f(x)dx = \int_0^\theta cf_1(x)dx + \int_\theta^\infty f_2(x)dx =$
  $= cF_1(\theta) + S_2(\theta) = cF_1(\theta) + \big(1 - F_2(\theta)\big) = 1$

Now, we are going to determine the constant c; for a complete or an incomplete data, Let $y_1 < y_2 < \cdots < y_k$ are k unique not censored values appear in the not censored observations. Assume that the threshold θ is in one of the interval, for example, $y_{j-1} \le \theta < y_j$, due to the required condition

$$cF_1(\theta) = F_2(\theta)$$

So we have

$$c\left(1-e^{-\theta/\mu}\right)=1-e^{-\sum_{i=1}^{j-1}\frac{s_i}{r_i}}$$

Therefore,

$$c=\frac{e^{\theta/\mu}}{e^{\theta/\mu}-1}\left(1-e^{-\sum_{i=1}^{j-1}\frac{s_i}{r_i}}\right)$$

Example 3.2: The claim payments on sample of 15 policies are, 2, 2, 2, 2, 3, 3, 5, 5+, 7, 7+, 9, 9+, 10, 10+, The + indicates that the losses exceeded the policy limit. This data set is incomplete data which reflect the real insurance industry cases. In order to construct model, we first need to construct the risk set as follows;

y_j	s_j	r_j
2	4	15
3	2	11
5	1	9
7	1	7
9	1	4
10	1	2

Then the Nelson Aalen estimator of the cumulative hazard rate function is given by;

$$H(x)=\begin{cases}0, & x<2\\ \frac{4}{5}=0.267, & 2\le x<3\\ 0.267+\frac{2}{11}=0.448, & 3\le x<5\\ 0.448+\frac{1}{9}=0.560, & 5\le x<7\\ 0.560+\frac{1}{7}=0.702, & 7\le x<9\\ 0.702+\frac{1}{4}=0.952, & 9\le x<10\\ 0.952+\frac{1}{2}=1.452, & x\ge 10\end{cases}$$

Select $\theta=2$ and let x be the payment random variable follows the composite exponential-Nelson Aalen model. Then the cumulative function of $x, F(x)$ is given by

$$F(x)=\begin{cases}\left(\frac{0.234e^{2/\mu}}{e^{2/\mu}-1}\right)\left(1-e^{-x/\mu}\right), & 0<x\le 2\\ 1-e^{-0.267}=0.234, & 2\le x<3\\ 1-e^{-0.448}=0.361, & 3\le x<5\\ 1-e^{-0.560}=0.429, & 5\le x<7\\ 1-e^{-0.702}=0.504, & 7\le x<9\\ 1-e^{-0.952}=0.614, & 9\le x<10\\ 1-e^{-1.452}=0.766, & x\ge 10\end{cases}$$

Even the data is artificial, but it has the properties of the real data. Use the exponential to model the losses small than 2 is better than to assume the uniform. We still need to consider the large loss with small probability part. By using the Nelson Aalen estimator,

the cumulative function at ∞ is not equal to 1 which comes from the drawback of the Nelson Aalen estimator itself. We are going to discuss this situation late.

3.3 The Composite Exponential- Kaplan-Meier Product Limit Model

Definition: The composite exponential-Kaplan-Meier product limit model is defined and constructed as follows; Let X be the random variable with density

$$f(x) = \begin{Bmatrix} cf_1(x) & if\ 0 < x < \theta \\ f_2(x) & if\ \theta < x < \infty \end{Bmatrix}$$

Where c the normalizing constant, $f_1(x)$ follows the exponential distribution and $f_2(x)$ follows the Kaplan-Meier product limit estimator which is data dependent distribution. More specifically, the density functions are given as follows,

$$f_1(x) = \frac{1}{\mu} e^{-x/\mu}, \qquad x > 0$$

And $f_2(x)$ with the distribution function $F_2(x)$ explicitly as follows;
For a given insurance data set (complete or modified data), the risk set can be defined as the $\{y_j, s_j, r_j\}$, where j from 1 to k. Then the survival distribution function of the loss random variable $S_2(x)$ as follows,

$$S_2(x) = \begin{cases} 1, & x < y_1 \\ \prod_{i=1}^{j-1} \left(\frac{r_i - s_i}{r_i}\right), & y_{j-1} \le x < y_j, j = 2, \cdots, k \\ \prod_{i=1}^{k} \left(\frac{r_i - s_i}{r_i}\right) or\ 0, & x \ge y_k \end{cases}$$

The cumulative distribution function can be obtained from

$$F_2(x) = 1 - S_2(x)$$

Where μ is unknown parameter and we require the following condition θ

$$cF_1(\theta) = F_2(\theta)$$

This condition will ensure the total frequencies for losses below the threshold θ are same as the total probability below the threshold θ.

Density and Value of c: It is not hard to show that $f(x)$ is the density function. It satisfies the two conditions;

- $f(x) \ge 0\ for\ all\ x$
- $\int_0^\infty f(x)dx = \int_0^\theta cf_1(x)dx + \int_\theta^\infty f_2(x)dx =$

$$= cF_1(\theta) + S_2(\theta) = cF_1(\theta) + \left(1 - F_2(\theta)\right) = 1$$

Now, we are going to determine the constant c; for a complete or an incomplete data, Let $y_1 < y_2 < \cdots < y_k$ are k unique not censored values appear in the not censored observations. Assume that the threshold θ is in one of the interval, for example, $y_{j-1} \le \theta < y_j$, due to the required condition

$$cF_1(\theta) = F_2(\theta)$$

So we have

$$c\left(1 - e^{-\theta/\mu}\right) = 1 - \prod_{i=1}^{j-1} \left(\frac{r_i - s_i}{r_i}\right)$$

Therefore,

$$c = \frac{e^{\theta/\mu}}{e^{\theta/\mu} - 1}\left(1 - \prod_{i=1}^{j-1}\left(\frac{r_i - s_i}{r_i}\right)\right)$$

Example 3.3: We use the same information as example 3.2. The claim payments on sample of 15 policies are, 2, 2, 2, 2, 3, 3, 5, 5+, 7, 7+, 9, 9+, 10, 10+, The + indicates that the losses exceeded the policy limit. The risk set was constructed as follows;

y_j	s_j	r_j
2	4	15
3	2	11
5	1	9
7	1	7
9	1	4
10	1	2

Select $\theta = 2$ and let x be the payment random variable follows the composite exponential-Kaplan Meier product limit model. Then the cumulative function of x, $F(x)$ is given by

$$F(x) = \begin{cases} \left(\dfrac{4e^{2/\mu}}{15\left(e^{2/\mu} - 1\right)}\right)\left(1 - e^{-x/\mu}\right), & 0 < x \le 2 \\ 1 - \dfrac{11}{15} = \dfrac{4}{15} = 0.267, & 2 \le x < 3 \\ 1 - \dfrac{11}{15} \times \dfrac{9}{11} = 0.400, & 3 \le x < 5 \\ 1 - \dfrac{11}{15} \times \dfrac{9}{11} \times \dfrac{8}{9} = 0.467, & 5 \le x < 7 \\ 1 - \dfrac{11}{15} \times \dfrac{9}{11} \times \dfrac{8}{9} \times \dfrac{6}{7} = 0.543, & 7 \le x < 9 \\ 1 - \dfrac{11}{15} \times \dfrac{9}{11} \times \dfrac{8}{9} \times \dfrac{6}{7} \times \dfrac{3}{4} = 0.657, & 9 \le x < 10 \\ 1 - \dfrac{11}{15} \times \dfrac{9}{11} \times \dfrac{8}{9} \times \dfrac{6}{7} \times \dfrac{3}{4} \times \dfrac{1}{2} = 0.829, & x \ge 10 \end{cases}$$

Even the data is artificial, but it has the properties of the real data. Use the exponential to model the losses small than 2 is better than to assume the uniform. We still need to consider the large loss with small probability part. By using the Kaplan-Meier product limit, the cumulative function at ∞ is not equal to 1 due to $r_k \neq s_k$; which comes from the drawback of the Kaplan Meier product limit itself. We are going to discuss this situation late.

4. Composite Models for Large Losses with Small Probabilities

All three models in the section 4 have the drawback for the large losses with small probabilities due to their cumulative distribution functions either are not defined at ∞ or are not equal to zero. But the large losses with small probabilities are very important cases in the insurance industry. Because they represent the occasional occurred events with the major impact. Therefore to model the tail distribution is very important. As we discussed before, normal is not the good distribution to model tail due to it goes fade very fast. The exponential is also not a good model because the thin tails. Therefore the Pareto distribution has been considered by many researchers to model the large losses due to it has slower fade rate than the normal and exponential distribution.

4.1 The Composite Empirical-Pareto Model

Definition: The composite empirical-Pareto model is defined and constructed as follows; Let X be the random variable with density

$$f(x) = \begin{cases} f_1(x) & if\ 0 < x < \theta \\ cf_2(x) & if\ \theta < x < \infty \end{cases}$$

Where c the normalizing constant $f_1(x)$ follows the empirical distribution which is data dependent distribution and $f_2(x)$ follows the Pareto distribution and More specifically, the density functions are given as follows,

$$f_1(x) == \frac{n_j}{n(c_j - c_{j-1})},\ for\ c_{j-1} \le x < c_j\ and\ j = 1, 2, \cdots k$$

$$f_2(x) = \frac{\alpha\theta^\alpha}{x^{\alpha+1}}, \qquad x > \theta$$

Where α is unknown parameter and we require the following condition θ

$$S_1(\theta) = cS_2(\theta)$$

This condition will ensure the total frequencies for losses above the threshold θ are same as the total probability above the threshold θ.

Density and Value of c: It is not hard to show that $f(x)$ is the density function. It satisfies the two conditions;

- $f(x) \ge 0\ for\ all\ x$
- $\int_0^\infty f(x)dx = \int_0^\theta f_1(x)dx + \int_\theta^\infty cf_2(x)dx =$

$$= F_1(\theta) + cS_2(\theta) = 1 - S_1(\theta) + cS_2(\theta) = 1$$

Now, we are going to determine the constant c. For the complete group data, let the group boundaries be $c_0 < c_1 < \cdots < c_k$, where often $c_0 = 0\ and\ c_k = \infty$. Assume that the threshold θ is in one of the interval, for example, $c_{j-1} \le \theta < c_j$, due to the required condition

$$S_1(\theta) = cS_2(\theta)$$

So we have

$$\frac{c_j - \theta}{c_j - c_{j-1}} S_1(c_{j-1}) + \frac{\theta - c_{j-1}}{c_j - c_{j-1}} S_1(c_j) = c\left(\frac{\theta}{\theta + \theta}\right)^\alpha$$

Therefore,

$$c = 2^\alpha \left(\frac{c_j - \theta}{c_j - c_{j-1}} S_1(c_{j-1}) + \frac{\theta - c_{j-1}}{c_j - c_{j-1}} S_1(c_j) \right)$$

Where $S_1(c_{j-1}) = \frac{\sum_{i=j+1}^k n_i}{n}$ and $S_1(c_j) = \frac{\sum_{i=j+2}^k n_i}{n}$.

Example 4.1: The information is same as the example 3.1. The following observations represent payments on 422 claims from an insurance policy. Data are artificial.

Payment Range	Number of Payments
0-7,500	299
7,500-17,500	45
17,500-32,500	26
32,500-67,500	24
67,500-125,000	16
125,000-300,000	9
Over 300,000	3

Select Let $\theta = 300{,}00$ and let x be the payment random variable follows the composite empirical-Pareto model. Then the density function of x is given by

$$f(x) = \begin{cases} \frac{299}{422(7{,}500)}, & 0 < x \leq 7{,}500 \\ \frac{45}{422(10{,}000)}, & 7{,}500 \leq x < 17{,}500 \\ \frac{26}{422(15{,}000)}, & 17{,}500 \leq x < 32{,}500 \\ \frac{24}{422(35{,}000)}, & 32{,}500 \leq x < 67{,}500 \\ \frac{16}{422(57{,}500)}, & 67{,}500 \leq x < 125{,}000 \\ \frac{9}{422(175{,}000)}, & 125{,}000 \leq x < 300{,}000 \\ 2^{\alpha}\frac{3\alpha(300{,}000)^{\alpha}}{422x^{\alpha+1}}, & x \geq 300{,}000 \end{cases}$$

Even the data is artificial, but it has the properties of the real data. Using the Pareto distribution to model the losses larger than 300,000 solve the probable, the distribution is undefined when the losses greater than 300,000. This is better model than the empirical model.

4.2 The Composite Nelson Aalen-Pareto Model

Definition: The composite Nelson Aalen-Pareto model is defined and constructed as follows; Let X be the random variable with density

$$f(x) = \begin{cases} f_1(x) & if\ 0 < x < \theta \\ cf_2(x) & if\ \theta < x < \infty \end{cases}$$

Where c the normalizing constant $f_1(x)$ follows the Nelson Aalen estimator which is data dependent distribution and $f_2(x)$ follows the Pareto distribution and more specifically, the density functions are given as follows,

$$f_2(x) = \frac{\alpha\theta^{\alpha}}{x^{\alpha+1}}, \qquad x > \theta$$

And $f_1(x)$ with the distribution function $F_1(x)$ explicitly as follows;

For a given insurance data set (complete or modified data), the risk set can be defined as the $\{y_j, s_j, r_j\}$, where j from 1 to k. Then the Nelson Aalen estimator of the cumulative hazard rate function is defined as follows,

$$H(x) = \begin{cases} 0, & x < y_1 \\ \sum_{i=1}^{j-1} \frac{s_i}{r_i}, & y_{j-1} \leq x < y_j, j = 2, \cdots, k \\ \sum_{i=1}^{k} \frac{s_i}{r_i}, & x \geq y_k \end{cases}$$

The distribution function can be obtained from

$$F_1(x) = 1 - S_1(x) = 1 - e^{-H(x)}$$

Where $S_1(x)$ is the survival function of the loss random variable and α is unknown parameter and we require the following condition θ

$$S_1(\theta) = cS_2(\theta)$$

This condition will ensure the total frequencies for losses above the threshold θ are same as the total probability above the threshold θ.

Density and Value of c: It is not hard to show that $f(x)$ is the density function. It satisfies the two conditions;

- $f(x) \geq 0 \; for \; all \; x$
- $\int_0^\infty f(x)dx = \int_0^\theta f_1(x)dx + \int_\theta^\infty cf_2(x)dx =$
$$= F_1(\theta) + cS_2(\theta) = 1 - S_1(\theta) + cS_2(\theta) = 1$$

Now, we are going to determine the constant c; for a complete or an incomplete data, Let $y_1 < y_2 < \cdots < y_k$ are k unique not censored values appear in the not censored observations. Assume that the threshold θ is in one of the interval, for example, $y_{j-1} \leq \theta < y_j$, due to the required condition

$$S_1(\theta) = cS_2(\theta)$$

So we have

$$e^{-\sum_{i=1}^{j-1}\frac{s_i}{r_i}} = c\left(\frac{\theta}{\theta+\theta}\right)^\alpha$$

Therefore,

$$c = 2^\alpha e^{-\sum_{i=1}^{j-1}\frac{s_i}{r_i}}$$

Example 4.2: The information is same as the example 3.2. The claim payments on sample of 15 policies are, 2, 2, 2, 2, 3, 3, 5, 5+, 7, 7+, 9, 9+, 10, 10+, The + indicates that the losses exceeded the policy limit. The risk set was found as follows;

y_j	s_j	r_j
2	4	15
3	2	11
5	1	9
7	1	7
9	1	4
10	1	2

Then the Nelson Aalen estimator of the cumulative hazard rate function is given by;

$$H(x) = \begin{cases} 0, & x < 2 \\ \frac{4}{5} = 0.267, & 2 \leq x < 3 \\ 0.267 + \frac{2}{11} = 0.448, & 3 \leq x < 5 \\ 0.448 + \frac{1}{9} = 0.560, & 5 \leq x < 7 \\ 0.560 + \frac{1}{7} = 0.702, & 7 \leq x < 9 \\ 0.702 + \frac{1}{4} = 0.952, & 9 \leq x < 10 \\ 0.952 + \frac{1}{2} = 1.452, & x \geq 10 \end{cases}$$

Select $\theta = 10$ and let x be the payment random variable follows the composite Nelson Aalen-Pareto model. Then the survival function of x, $S(x)$ is given by

$$S(x)=\begin{cases}1, & 0<x\le 2\\ e^{-0.267}=0.766, & 2\le x<3\\ e^{-0.448}=0.639, & 3\le x<5\\ e^{-0.560}=0.571, & 5\le x<7\\ e^{-0.702}=0.496, & 7\le x<9\\ e^{-0.952}=0.386, & 9\le x<10\\ 0.234\times 2^{\alpha}\left(\dfrac{10}{x+10}\right)^{\alpha} & x\ge 10\end{cases}$$

Even the data is artificial, but it has the properties of the real data. Use the Pareto to model the losses large than 10 solve the problem that the survival function never fade at ∞. This model is better than the Nelson Aalen estimator to fit the given data and improve the drawback of the Nelson Aalen estimator.

4.3 The Composite Kaplan-Meier Product Limit-Pareto Model

Definition: The composite Kaplan-Meier product limit-Pareto model is defined and constructed as follows; Let X be the random variable with density

$$f(x)=\begin{Bmatrix} f_1(x) & if\ 0<x<\theta \\ cf_2(x) & if\ \theta<x<\infty \end{Bmatrix}$$

Where c the normalizing constant $f_1(x)$ follows the Nelson Aalen estimator which is data dependent distribution and $f_2(x)$ follows the Pareto distribution and more specifically, the density functions are given as follows,

$$f_2(x)=\frac{\alpha\theta^{\alpha}}{x^{\alpha+1}},\qquad x>\theta$$

And $f_1(x)$ with the distribution function $S_1(x)$ explicitly as follows;

For a given insurance data set (complete or modified data), the risk set can be defined as the $\{y_j, s_j, r_j\}$, where j from 1 to k. Then the survival distribution function of the loss random variable $S_1(x)$ as follows,

$$S_1(x)=\begin{cases}1, & x<y_1\\ \prod_{i=1}^{j-1}\left(\dfrac{r_i-s_i}{r_i}\right), & y_{j-1}\le x<y_j, j=2,\cdots,k\\ \prod_{i=1}^{k}\left(\dfrac{r_i-s_i}{r_i}\right) or\ 0, & x\ge y_k\end{cases}$$

Where α is unknown parameter and we require the following condition θ

$$S_1(\theta)=cS_2(\theta)$$

This condition will ensure the total frequencies for losses above the threshold θ are same as the total probability above the threshold θ.

Density and Value of c: It is not hard to show that $f(x)$ is the density function. It satisfies the two conditions;

- $f(x)\ge 0\ for\ all\ x$
- $\int_0^{\infty} f(x)dx=\int_0^{\theta} f_1(x)dx+\int_{\theta}^{\infty} cf_2(x)dx=$

$$=F_1(\theta)+cS_2(\theta)=1-S_1(\theta)+cS_2(\theta)=1$$

Now, we are going to determine the constant c; for a complete or an incomplete data, Let $y_1<y_2<\cdots<y_k$ are k unique not censored values appear in the not censored observations. Assume that the threshold θ is in one of the interval, for example, $y_{j-1}\le$

$\theta < y_j$, due to the required condition

$$S_1(\theta) = cS_2(\theta)$$

So we have

$$\prod_{i=1}^{j-1}\left(\frac{r_i - s_i}{r_i}\right) = c\left(\frac{\theta}{\theta+\theta}\right)^{\alpha}$$

Therefore,

$$c = 2^{\alpha}\prod_{i=1}^{j-1}\left(\frac{r_i - s_i}{r_i}\right)$$

Example 4.3: We use the same information as example 3.2. The claim payments on sample of 15 policies are, 2, 2, 2, 2, 3, 3, 5, 5+, 7, 7+, 9, 9+, 10, 10+, The + indicates that the losses exceeded the policy limit. The risk set was constructed as follows;

y_j	s_j	r_j
2	4	15
3	2	11
5	1	9
7	1	7
9	1	4
10	1	2

Select $\theta = 10$ and let x be the payment random variable follows the composite Kaplan Meier product limit-Pareto model. Then the survival function of $x, S(x)$ is given by

$$S(x) = \begin{cases} 1, & 0 < x \le 2 \\ \frac{11}{15} = 0.733, & 2 \le x < 3 \\ \frac{11}{15}\times\frac{9}{11} = 0.600, & 3 \le x < 5 \\ \frac{11}{15}\times\frac{9}{11}\times\frac{8}{9} = 0.533, & 5 \le x < 7 \\ \frac{11}{15}\times\frac{9}{11}\times\frac{8}{9}\times\frac{6}{7} = 0.457, & 7 \le x < 9 \\ \frac{11}{15}\times\frac{9}{11}\times\frac{8}{9}\times\frac{6}{7}\times\frac{3}{4} = 0.343, & 9 \le x < 10 \\ 0.1714 \times 2^{\alpha}\left(\frac{10}{x+10}\right)^{\alpha}, & x \ge 10 \end{cases}$$

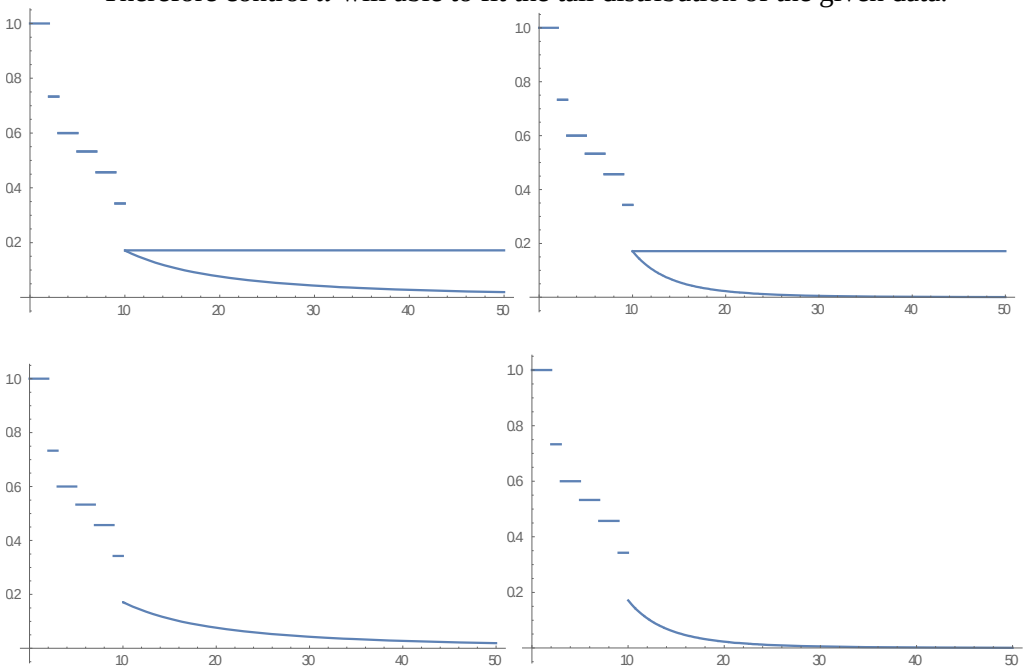
Even the data is artificial, but it has the properties of the real data. Use the Pareto to model the losses greater than 10 solve the problem that survival function is not fade at ∞. It doesn't matter whether $r_k = s_k \; or \; r_k \neq s_k$; the composite model is defined and it is a better fit to the given data especially for the large losses with small probabilities. This is the better model than the Kaplan Meier product limit.

5. Conclusions

The following picture shows the composite Kaplan-Meier product limit-Pareto model when $\alpha = 2$ and $\alpha = 5$ respectively. From the picture, we are able to draw the following conclusions

- The composite Kaplan-Meier product limit-Pareto model solves the drawback of the Kaplan-Meier product limit estimator. The survival function fades to zero at ∞.

- The composite Kaplan-Meier product limit-Pareto model is a one parameter model. α is unknown parameter. The larger the α, the quick the model fades. Therefore control α will able to fit the tail distribution of the given data.



References

1. Teodorescu, S. and Vernic, R. (2006) *A Composite Exponential-Pareto Distribution*, An. St. Univ. Ovidius Constanta Vol. 14(1), 99-108.
2. Vernic, R., Teodorescu, S. and Pelican, E. (2009) *Two Lognormal Models for Real Data*, An. St. Univ. Ovidius Constanta Vol. 17(3), 263-270.
3. Scollnik, D. and Sun, C. (2012) *Modeling with Weibull-Pareto Models*, North American Actuarial Journal, Vol. 16, Number 2, (260-272)
4. Deng, M. (2002) *The Total Claims Distributions Under The Different Conditions*. American Statistical Association 2002 Proceedings of the Section on Business and Economic Statistics Section. 741-746.
5. Deng, M. (2013) *Estimation and Evaluating of Right Tail Risk*, International Journal of Applied Science and Technology, Vol. 3, No. 7 (1-13)
6. Klugman, S., Panjer, H., and Willmot G. (2012) *Loss Models from Data to Decisions* A John Wiley and Sons, Inc.
7. Cooray, K. and Ananda, M.M.A. (2005) *Modeling Actuarial Data with Composite Lognormal-Pareto Model*, Scandinavian Actuarial Journal, 5, 321-334