UNIVERSIDAD SANTO TOMAS
PRIMER CLAUSTRO UNIVERSITARIO DE COLOMBIA

**Facultad de Estadística**
Proyecto de Grado
Noviembre 2019

# Modelamiento de la severidad para automóviles asegurados en Bélgica por medio de mixturas finitas de distribuciones para el año 1997

**Severity modeling for cars insured in Belgium through finite mixtures of distributions for the year 1997**

Jessica Alexandra Cardona Rodriguez.[a]
jessicacardona@usantotomas.edu.co

Wilmer Pineda Ríos.[b]
wilmerpineda@usantotomas.edu.co

## Resumen

En este trabajo se realiza un análisis de la severidad para automóviles asegurados, se pretende observar la relación con el tipo de automóvil, cobertura de seguro, edad del automóvil, edad de la persona, genero, seguro grupal, periodo de pago y tipo de combustible, por medio de una mixtura finita de distribuciones condicionada a la media. Se considerarán diferentes distribuciones continuas que permitan ajustar valores donde la severidad es alta, debido al tipo de automóvil, cobertura y demás factores que influyan al aumento de esta, se identifican las distribuciones o distribución que se ajuste a los datos obteniendo así los diferentes parámetros y estimaciones de la mixtura con sus respectivos pesos y número de componentes. El algoritmo hope-maximization (EM) es un método iterativo que se encarga de encontrar máximos locales de la verosimilitud, en donde su función es hacer la estimación de máxima verosimilitud de los parámetros de la mixtura.

## Abstract

In this work an analysis of the severity for insured cars will be carried out, it is intended to observe the relationship with the type of car, insurance coverage, age of the car, age of the person, gender, group insurance, payment period and type of fuel , by means of a finite mixture of distributions conditioned to the average. Different continuous distributions that allow adjusting values where the severit y is high, due to the type of car, coverage and other factors that influence the increase of this, will be considered distributions or distribution that fits the data thus obtaining the different parameters and estimates of the mixture with their respective weights and number of components. The hope-maximization algorithm (EM) is an iterative method which is responsible for finding local maximums of likelihood, where its function is to estimate the maximum likelihood of the parameters of the mixture.

[a]Estudiante pregrado Estadística, U. Santo Tomás, sede Bogotá
[b]Docente Facultad de Estadística, U. Santo Tomás, sede Bogotá

# 1. Introducción

Existen poblaciones que por su naturaleza se encuentran valores que son igual de frecuentes al punto de que se observa una bimodalidad o multimodalidad, por lo tanto, se obtienen múltiples picos en su distribución, así que la población se puede fraccionar por cada moda para obtener subpoblaciones y hallar la distribución que se ajusta a cada subpoblación.

Se han realizado modelos para determinar el número de siniestro para el seguro automovilístico, Hiraldo y Guerrero (2015) realizaron un análisis econométrico realizando un modelo binomial negativo inflado de ceros. La severidad es un factor importante para la industria aseguradora, en Barranquilla Díaz (2017) realizó un estudio de los factores de riesgo que afectan la severidad aplicado a la ciudad de Cartagena, para este caso se realizaron modelos multinomiales, logit anidado y ordenados.

Por otra parte, los modelos de mixtura proporcionan un marco semiparamétrico conveniente en el cual modelar formas de distribución desconocidas, independientemente del objetivo, ya sea puede ser, por ejemplo, la estimación de densidad (Mc Lachlan & Peel, 2000)

Además, las mixturas finitas de distribuciones han proporcionado un enfoque matemático para el modelado estadístico de una amplia variedad de fenómenos aleatorios. Debido a su utilidad como método extremadamente flexible de modelado, los modelos de mixturas finitas han seguido recibiendo una atención creciente a lo largo de los años, tanto desde el punto de vista práctico como teórico. (Mc Lachlan & Peel, 2000)

Por otro lado, se han realizado estudios de mixturas de distribuciones normales condicionado al modelado de la media y varianza (Roldan,2014), donde se evalúan diversos escenarios para la varianza y la media. Autores como Molina y Jiménez (2014), realizaron mixturas de distribuciones Weibull para la valoración de derivados europeos, donde se toma esta distribución por su asimetría ajustándose así a los datos que se tienen en este estudio.

Del mismo modo, la empresa MAPFRE realizó un estudio en el cual se aplican mixturas de distribuciones para la siniestralidad, como alternativa a las aproximaciones clásicas que consideran, en el caso del número de siniestros y coste de cada uno de ellos. una única función de distribución, se pueden utilizar mixturas de distribuciones que consisten básicamente en construir una distribución de probabilidad que es combinación lineal de distintas componentes. (Cid, 1999).

En esa misma línea, los modelos de mixtura finita sustentan una variedad de técnicas en áreas principales de estadística, incluidos análisis de clase y de clase latente, análisis discriminante, análisis de imágenes y análisis de supervivencia, además de su papel más directo en el análisis de datos e inferencia de proporcionar descripciones modelos para distribuciones. (Mc Lachlan & Peel, 2000)

Por consiguiente, este proyecto pretende identificar el comportamiento de los datos, para así elegir la mixtura adecuada, obteniendo un buen ajuste , se busca identificar el número de modas que se presenten en los datos para asi poder determinar el número de componentes del modelo.

La organización de este documento es la siguiente, en la sección 2 se abarca el marco teórico de las mixturas de distribuciones y el algoritmo EM;en la tercera sección se desarrolla el marco metodológico donde se explica el tipo de estudios, base de datos y descriptivos;en la cuarta sección se muestran los resultados obtenidos en el proceso;en la quinta sección se muestran las conclusiones y trabajos futuros.

# 2. Marco Teórico

## 2.1. Mixtura de distribuciones

Una mixtura se encarga de descomponer una función densidad de probabilidad en la suma de varias funciones de densidad ponderadas por un peso.

Las distribuciones mixtas se utilizan para la modelización de datos heterogéneos en multitud de situaciones experimentales, en donde aquellos pueden interpretarse como procedentes de dos o más subpoblaciones (componentes). La obtención de estas componentes conduce a la estimación de los parámetros de la mixtura (Gómez, 2014).

En primero lugar, los modelos de mixturas finitas no condicionadas han sido aplicados y generado aproximaciones importantes en eventos considerados aleatorios. De manera que las mixturas finitas se analizan como densidades que se componen de poblaciones subyacentes m que conforman una población estadística. Estas poblaciones subyacentes se denominaran como componentes de la mixtura $f_i$, donde $i = 1, 2, ..., m$, entonces cada población que compone la mixtura son proporciones de la mixtura, es decir, es una fracción $\alpha_i$, por lo que $i = 1, 2, ..., m$.

Por lo tanto, el modelo en su forma matemática se denota de la siguiente manera:

$$f(y) = \sum_{j=1}^{m} p_j f_j(y; \theta_j)$$

Dada una variable aleatoria Y, los modelos de mixtura finita descomponen una función de densidad de probabilidad $f(y)$ en la suma de $m$ funciones de densidad de probabilidad. Si $f_j(y)$ es la j-ésima función de densidad de probabilidad que compone la mixtura finita con $m$ componentes y donde $p_j$ es la proporción de la mixtura o peso de la j-ésima componente en la mixtura con la restricción de que $0 \leqslant p_j \leqslant 1$ y $\sum_{j=1}^{m} p_j = 1$ para $j = 1$, ..., $m$. La proporción $p_j$ se puede interpretar como la probabilidad a priori de observar una muestra de la componente $j$ (Roldan,2014).

Por otra parte, los modelos de mixturas finitas condicionadas, hace referencia a que están condicionadas a la media, y estos modelos de mixturas de regresiones lineales por medio del modelaje de distintas distribuciones (Ding, 2006).

De manera que, se tiene el siguiente modelo:

$$\mu_j = \beta_{0j} + \beta_{1j} x_1 + \hat{a} + \beta_{kj} x_k \mu_k = \beta_j' x$$

Donde $x' = (x_{1,}\hat{a}, x_k)$ es un vector de k variables independientes explicativas para la media de la j-ésima componente de la mixtura, y $\beta_j' = (\beta_{0j}, \beta_{1j}, \hat{a}, \beta_{kj})$ son los coeficientes del modelo de regresión de la media de la j-ésima componente de la mixtura.

Para cada $\theta_1$ ,..., $\theta_g$ es un elemento del mismo espacio de parámetros $\Theta$. Entonces se puede pensar que:

$$\pi_1 = (\pi_1, ..., \pi_g)^T$$

La función H se llama distribución de mixtura y está dada por:

$f(y_j; H) = \int f(y_j; \theta) dH(\theta)$

Con g puntos se obtiene que la tasa óptima de convergencia en la estimación de H es $n^{-\frac{1}{4}}$.

A menudo, en la práctica, donde el interés principal en ajustar un modelo de mixtura es estimar las proporciones de mixtura, las densidades de los componentes se especifican o se pueden estimar por separado de los datos clasificados disponibles.

Se evalúa una función de gradiente simple que tan cerca esta de un estimador candidato $H^*$ de una solución ML $\hat{H}$, esta función puede ser la base para utilizar el algoritmos EM.

## 2.2. Algoritmo EM

Los algoritmos de maximización de la esperanza (expectation-maximization, EM) son procedimientos que permiten la maximización de una función LL cuando los procedimientos estándar son numéricamente difíciles o inviables.

El algoritmo EM se puede utilizar para estimar distribuciones de preferencias muy flexibles, incluidas especificaciones no paramétricas que pueden aproximar asintóticamente cualquier distribución verdadera subyacente(Train,K).

Para la estimación de los parámetro se denotan como $\theta$, la probabilidad condicionada sobre la densidad esta dada por:

$$p(y|\theta) = \int p(y|z, \theta) f(z|\theta) dz.$$

La función LL que se busca maximizar es:

$$LL(\theta) = log\ P(y|\theta) = log\left(\int P(y|z, \theta)\ f(z|\theta) dz\right)$$

El procedimiento alternativo es iterativo, comenzando con un valor inicial de los parámetros y actualizándolos de una manera que se describirá a continuación. Denotemos el valor de prueba de los parámetros en una iteración dada como $\theta_t$ Definamos una nueva función en $\theta_t$ que se relacione con LL pero que utilice la distribución condicionada $h$. Esta nueva función es:

$$\varepsilon(\theta|\theta^t) = \int h(z|y, \theta^t)\ log(P(y|z, \theta) f(z|\theta)) dz$$

El procedimiento EM consiste en maximizar $\varepsilon$ repetidamente. Empezando con un cierto valor inicial, los parámetros se actualizan en cada iteración a través de la siguiente fórmula:

$$\theta^{t+1} = argmax_\theta \varepsilon(\theta|\theta^t)$$

En cada iteración, los valores actuales de los parámetros,$\theta_t$ , se utilizan para calcular los pesos $h$, y a continuación se maximiza la LL conjunta ponderada. El nombre EM proviene del hecho de que el procedimiento utiliza una esperanza que es maximizada.

la siguiente tabla podemos observarla expansión del uso de datos en M-step del algoritmo EM para ajustar el modelo de mixtura de parámetros comunes

Tabla 1: Algoritmo EM

| i | Mass | $y_e$ | $x_e$ | $\hat{p}^{(r+1)}$ |
|---|---|---|---|---|
| 1 | 1 | | | |
| 2 | 1 | | | |
| . | . | y | x | $\hat{p}_1^{(r+1)}$ |
| . | . | | | |
| . | . | | | |
| n | 1 | | | |
| 1 | 2 | | | |
| 2 | 2 | | | |
| . | . | y | x | $\hat{p}_2^{(r+1)}$ |
| . | . | | | |
| . | . | | | |
| n | 2 | | | |
| . | . | . | . | . |
| . | . | . | . | . |
| . | . | . | . | . |
| 1 | k | | | |
| 2 | k | | | |
| . | . | y | x | $\hat{p}_k^{(r+1)}$ |
| . | . | | | |
| . | . | | | |
| n | k | | | |

de los $\hat{p}_k^{(r+1)}$ contiene los pesos (calculados internamente) en la iteración (r + 1), la columna encabezada MASS identifica los componentes de la mixtura K.

# 3. Marco Metodológico

## 3.1. Tipo de estudio

El tipo de estudio es descriptivo, se busca descubrir y analizar la relación de las variables seleccionadas con la severidad, además es correlacionado donde se evalúa el grado de relación de las variables proporcionando una explicación a los resultados obtenidos.

## 3.2. Base de datos

La base de datos que se utilizó es este estudio es de Bélgica del año 1997, la cual contaba con 163.660 registros de seguros de automóviles de gama alta, media-baja, se toman para el estudio las observaciones que contengan valores mayores a cero en la severidad debido a que este estudio esta enfocado en los seguros que presenten al menos un reclamo a la aseguradora, se toma una muestra de 6000 datos para realizar el modelo las variables a trabajar son las siguientes:

- Amount: Severidad.
- Ageph: Edad de la persona.
- Agec: Antigüedad del vehículo.
- Fuel: Tipo de combustible:
  - 1 Gasolina.
  - 2 Diésel.
- Coverage: Cobertura del seguro:
  - 1 TPL (Solo daños de terceras partes).
  - 2 PO (TPL + daños parciales materiales propios).
  - 3 FO (TPL + daños totales materiales propios).
- Sport: Modelo del vehículo sport:
  - 1 Modelo sport.
  - 2 Modelo no sport.
- Sex: Género de la persona donde:
  - 1 Mujer.
  - 2 Hombre.
- Period : Pago en partes de la prima.
- Fleet : Vehículo que es parte de una flota.
  - 1 Es parte de flota.
  - 2 NO es parte de una flota.
- Nclaims: Número de reclamos

### 3.2.1. AMOUNT

La variable amount es la variable que mide la severidad la cual representa el costo del daño ocasionado por el asegurado, se presentaran algunos descriptivos de la variable para así identificar el comportamiento.

Tabla 2: Elaboración propia

| | |
|---|---|
| Minimo | 60 |
| máximo | 20'166.618 |
| media | 76.564,32 |

En la tabla 2 se puede observar un panorama de los datos donde el valor máximo se encuentra alejado de la media, estos valores son los que se quieren ajustar en la segunda componente del modelo.

se puede observar su comportamiento en la siguiente gráfica:

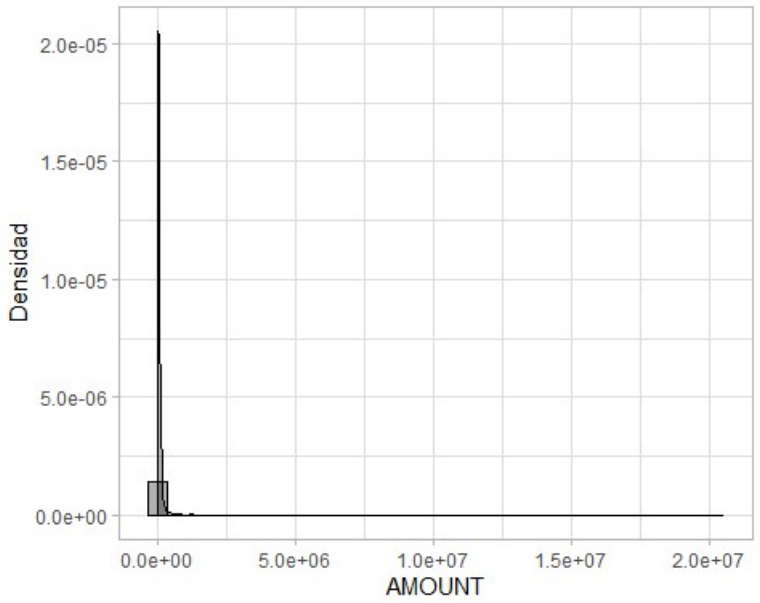


Figura 1: Elaboración propia

En la figura 1 podemos observar los valores de la severidad, no se logra identificar el comportamiento debido a los valores que se tienen para esto se va a realizar un acercamiento donde se filtra la variable con los valores menores a 1'000.000

Menos de un millón Más de un millón

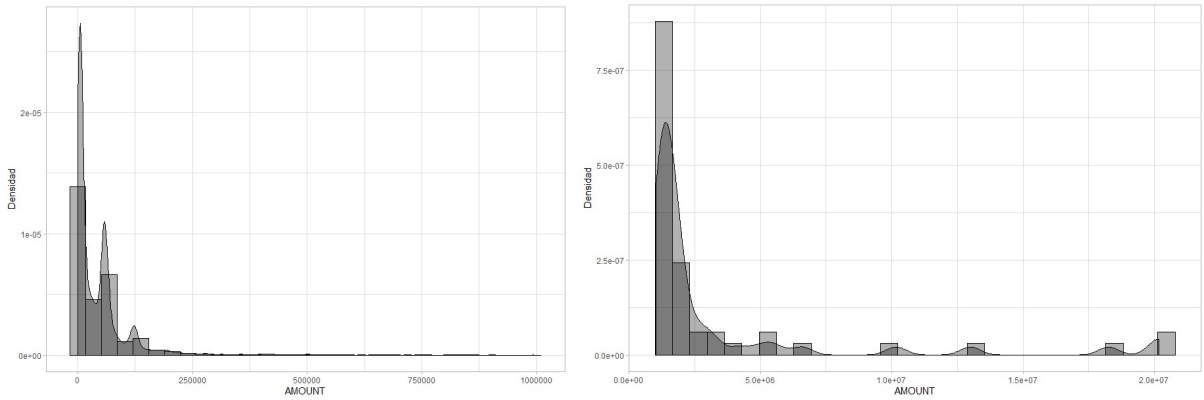


Figura 2: Elaboración propia

Al realizar el filtro se puede observar mejor la distribución y el posible número de componentes que se tendrán en la mixtura.

### 3.2.2. AGEPH

La variable Ageph representa la edad de las personas que adquieren el seguro para su automóvil, esta variable se describe en la siguiente tabla:

| Edad asegurado | |
|---|---|
| Minimo | 18 |
| máximo | 91 |
| media | 44 |
| Cuartil 1 | 32 |
| Cuartil 3 | 54 |

Tabla 3: Elaboración propia

En la tabla 3 podemos observar que el máximo de la muestra tiene personas de todas las edades con una media de 44 años y un máximo de 91 años.

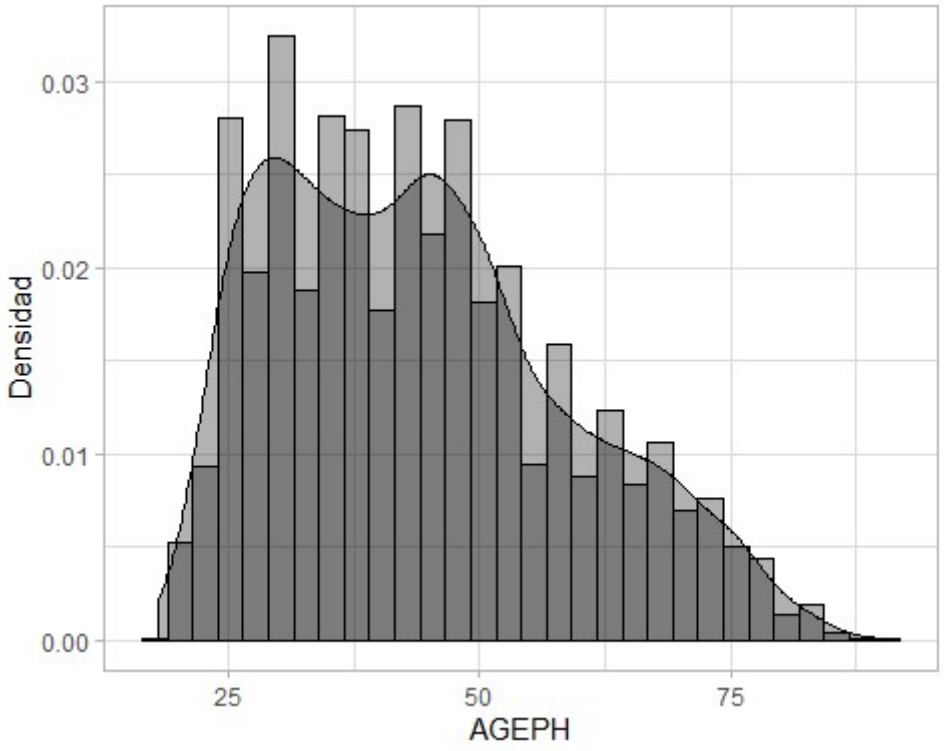


Figura 3: Elaboración propia

En la Figura 3 se observa una asimetría positiva la cual se concentra en el rango de 25 a 50 años, se presentan frecuencias con picos altos lo cual puede ser un factor importante que afecte la severidad.

### 3.2.3. AGEC

La variable Agec nos indica la antigüedad de los vehiculos, observamos que hay autos que son nuevos, el auto con mayor antigüedad es de 27 años.

| Antigüedad | |
|---|---|
| Minimo | 0 |
| máximo | 27 |
| media | 7,36 |

Tabla 4: Elaboración propia

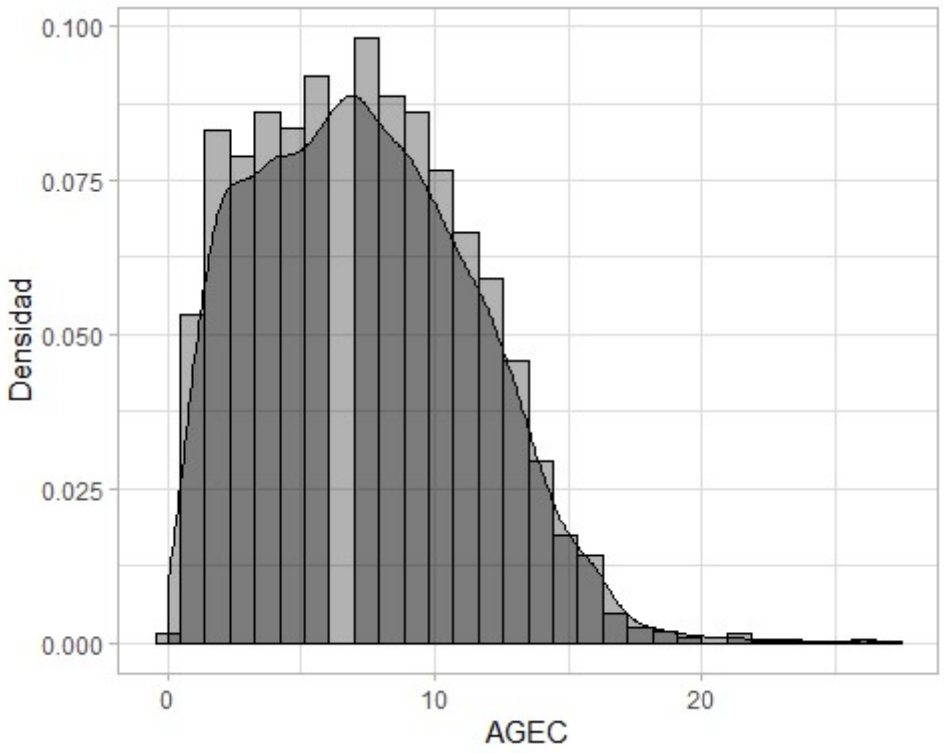


Figura 4: Elaboración propia

En el anterior gráfico se observa que la edad del vehículo tiene una asimétria a la derecha, concentrado así sus valores entre 0 y 10 apróximadamente. Este gráfico nos indica que los vehículos que se van a modelar son en su mayoria menores de 15 años.

### 3.2.4. COVERAGE

La variable Coverage nos indica la cobertura del seguro, podemos observar que en su mayoría la muestra tiene automóviles con cobertura 1 y 2. La cobertura Tipo 1 cuenta con más del 50 % de la muestra.

| Cobertura del seguro | |
|---|---|
| Tipo | Frecuencia |
| TPL | 3.607 |
| PO | 1.580 |
| FO | 813 |

Tabla 5: Elaboración propia

### 3.2.5. SEX

La variable Sex nos indica el género, observamos que en la muestra la mayoria de personas pertenecen al género masculino con un 71 %.

| TABLA DE GÉNERO | | | |
|---|---|---|---|
| ID | GÉNERO | FRECUENCIA | PORCENTAJE |
| 1 | Mujer | 1.714 | 28,57 |
| 2 | Hombre | 4.286 | 71,43 |

Tabla 6: Elaboración propia

### 3.2.6. SPORT

La variable Sport nos indica si el vehículo es Sport o no, se observa una frecuencia mayor en los vehículos que no son Sport contando con un 98,9 % de la muestra.

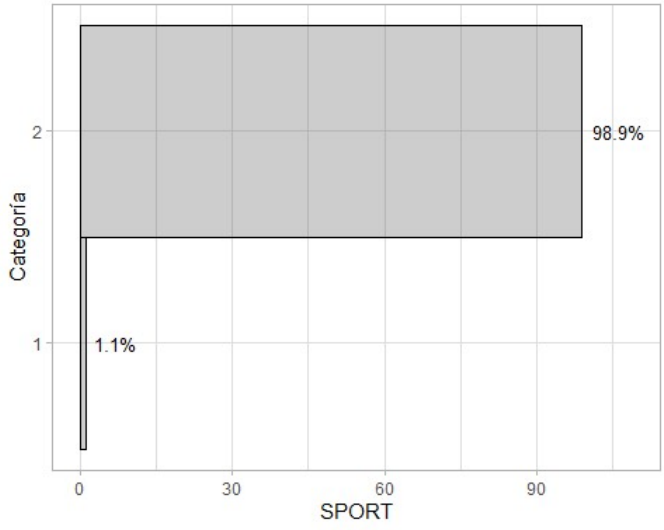


Figura 5: Elaboración propia

### 3.2.7. NCLAIMS

La variable Nclaims contiene la información del número de reclamos que han hecho los propietarios de los seguros, donde podemos observar que la mayoria ha realizado solo un reclamo con un 89,9 %

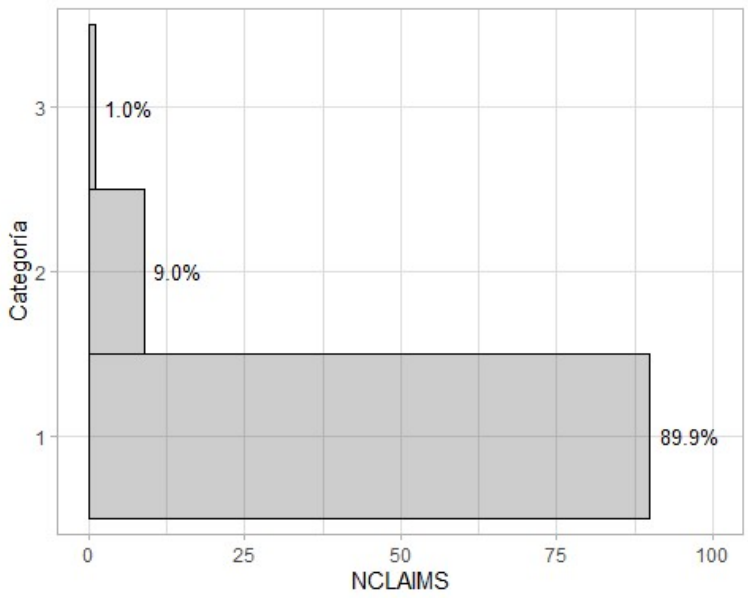


Figura 6: Elaboración propia

# 4. Resultados

## 4.1. Prueba de normalidad

Se realiza una prueba de normalidad para comprobar que la muestra no proviene de una distribución normal, se evaluaran 4 pruebas la hipotesis a contrastar es la siguiente:

$H_0 = La\ muestra\ proviene\ de\ una\ distribución\ normal.$
$H_1 = La\ muestra\ no\ proviene\ de\ una\ distribución\ normal.$

| PRUEBAS DE NORMALIDAD | | |
|---|---|---|
| PRUEBA | P-VALOR | R |
| Anderson-Darling | 2,2e-16 | ad.test() |
| Pearson chi-square | 2,2e-16 | pearson.test() |
| Jarque-Bera | 2,2e-16 | jb.norm.test() |
| Lilliefors (Kolmogorov-Smirnov) | 2,2e-16 | lillie.test() |

Tabla 7: Elaboración propia

En la tabla ocho podemos observar los valores de cada prueba, las cuales nos indican que los datos no siguen una distribución normal por ello se van a probar diferentes distribuciones que cumplan con el rango requerido.

## 4.2. Modelo

### 4.2.1. Distribuciones

estas distribuciones cumplen con el rango son continuas positivas donde $x > 0$ las distribuciones que se van a probar sonn las siguientes:

- GAMMA

  $f(x) = \frac{1}{\alpha^p \Gamma(p)} . e^{-\frac{x}{\alpha}} x^{p-1}$

- LOGNORMAL

  $f(x) = \frac{1}{\sqrt{2\pi\sigma}} \frac{1}{x} . e^{-\frac{(ln(x)-\mu)^2}{2\sigma^2}}$

- PARETO

  $f(x) = \frac{\alpha \ x_m^{\alpha}}{x^{\alpha+1}}$

- WEIBULL

  $f(x) = \frac{k}{\lambda} \left(\frac{x}{\lambda}\right)^{k-1} e^{-(\frac{x}{\lambda})^k}$

A lo largo de los años, se han utilizado una variedad de enfoques para estimar las distribuciones de mixturas, uno de los problemas de las distribuciones con dos parámetros es la asimetría y curtosis de la distribución, estos son fijos; el modelo GAMLSS(modelo aditivo generalizado para ubicación, escala y forma.) permite tener mayor flexibilidad por lo tanto se utiliza este modelo en el estudio.

El paquete que se va a utilizar para realizar la mixtura finita de distribuciones es *gamlss.mx*

Los parámetros y enlaces de estas distribuciones son los siguientes:

| DISTRIBUCION | RANGO | $\mu$ | $\sigma$ |
|---|---|---|---|
| GAMMA | (0,∞) | LOG | LOG |
| LOGNORMAL | (0,∞) | LOG | LOG |
| PARETO | (0,∞) | LOG | LOG |
| WEIBULL | (0,∞) | LOG | LOG |

Tabla 8: Elaboración propia

### 4.2.2. Resultado modelos de las distribuciones

Al aplicar el método se tienen los siguientes criterios de información:

De manera que en la tabla 10 podemos observar los criterios de información y la probabilidad estimada para cada componente del modelo, los AIC y deviance no tienen una diferencia significativa, por esto se analizaran los modelos que presenten una probabilidad estimada que logre capturar el comportamiento de los datos, debido a que los valores de severidad altos son pocos el componente que estime esta parte del modelo debe tener una probabilidad estimada pequeña, por esto se analizaran los modelos Gamma, Lognormal y Pareto Normal.

| DISTRIBUCIÓN | AIC | DEVIANCE | MODELO1 | MODELO 2 |
|---|---|---|---|---|
| Gamma | 141259 | 141209 | 0,03322167 | 0,9667783 |
| Lognormal | 140769 | 140719 | 0,966191 | 0,3380897 |
| Pareto | 140892 | 140842 | 0,3247549 | 0,6752451 |
| Weibull | 142449 | 142399 | 0,02562316 | 0,9743768 |
| Pareto lognormal | 140768 | 140718 | 0,03187405 | 0,968126 |
| Pareto Weibull | 140833 | 140783 | 0,5562663 | 0,4437337 |
| Pareto Gamma | 140893 | 140843 | 0,7193181 | 0,2806819 |

Tabla 9: Elaboración propia

### 4.2.3. MODELO GAMMA

$$\hat{f}(y_i) = \hat{\pi}_1 f_1(\hat{\mu}_{1i}, \hat{\sigma}_1) + \hat{\pi}_2 f_2(\hat{\mu}_{2i}, \hat{\sigma}_2)$$

$componente\ 1 : \hat{\mu}_1\ ln(\hat{\mu}_1) = \hat{\beta_0} + \hat{\beta_1}\ Ageph_i + \hat{\beta_2}\ Agec_i + \hat{\beta_3}(si\ Fuel = 2) + \hat{\beta_4}(si\ Period = 1) - \hat{\beta_5}(si\ Coverage = 2) + \hat{\beta_6}(si\ Coverage = 3) + \hat{\beta_7}(si\ Sport = 2) + \hat{\beta_8}(si\ Fleet = 2) + \hat{\beta_9}\ Nclaim_i + \hat{\beta_{10}}(si\ Sex = 2)$

$ln(\hat{\mu}_1) = -0,0452 + 0,0402 + 0,2624 + 3,6592(si\ Fuel = 2) + 2,1777(si\ Period = 1) - 0,2892(si\ Coverage = 2) + 0,3372(si\ Coverage = 3) + 4,4575(si\ Sport = 2) + 5,1514(si\ Fleet = 2) + 0,4676 + 0,3349(si\ Sex = 2)$

$\hat{\sigma}_1 = e^{0,6681} = 1,950528$

$\hat{\pi}_1 = 0,03322167$

$componente\ 2 : \hat{\mu}_1\ ln(\hat{\mu}_1) = \hat{\gamma_0} + \hat{\gamma_1}\ Ageph_i + \hat{\gamma_2}\ Agec_i + \hat{\gamma_3}(si\ Fuel = 2) + \hat{\gamma_4}(si\ Period = 1) - \hat{\gamma_5}(si\ Coverage = 2) + \hat{\gamma_6}(si\ Coverage = 3) + \hat{\gamma_7}(si\ Sport = 2) + \hat{\gamma_8}(si\ Fleet = 2) + \hat{\gamma_9}\ Nclaim_i + \hat{\gamma_{10}}(si\ Sex = 2)$

$ln(\hat{\mu}_2) = 9,3970 + 0,0031 + 0,0099 + 0,0549(si\ Fuel = 2) + 0,0425(si\ Period = 1) - 0,1235(si\ Coverage = 2) + 0,2454(si\ Coverage = 3) + 0,0055(si\ Sport = 2) + 0,0188(si\ Fleet = 2) + 0,8108 - 0,0223(si\ Sex = 2)$

$\hat{\sigma}_2 = e^{0,1348} = 1,144308$

$\hat{\pi}_2 = 0,9667783$

En la primera componente del modelo Gamma se describe los automóviles que presentan valores de severidad altos, la edad del automóvil tiene un efecto positivo, indica que a mayor edad de la persona el valor de la severidad aumenta 1,04 veces, por otro lado se observa un efecto negativo en la cobertura 2 en comparación con la cobertura 1.

La segunda componente tiene efectos similares a la primera, se observa una diferencia entre estas la cual esta dada por el genero, donde los hombres tienen un efecto positivo en la componente 1 a diferencia de la componente 2.

### 4.2.4. MODELO LOGNORMAL

$$\hat{f}(y_i) = \hat{\pi}_1 f_1(\hat{\mu}_{1i}, \hat{\sigma}_1) + \hat{\pi}_2 f_2(\hat{\mu}_{2i}, \hat{\sigma}_2)$$

$componente\ 1 : \hat{\mu}_1\ ln(\hat{\mu}_1) = \hat{\beta_0} + \hat{\beta_1}\ Ageph_i + \hat{\beta_2}\ Agec_i + \hat{\beta_3}(si\ Fuel = 2) + \hat{\beta_4}(si\ Period = 1) - \hat{\beta_5}(si\ Coverage = 2) + \hat{\beta_6}(si\ Coverage = 3) + \hat{\beta_7}(si\ Sport = 2) + \hat{\beta_8}(si\ Fleet = 2) + \hat{\beta_9}\ Nclaim_i + \hat{\beta_{10}}(si\ Sex = 2)$

$ln(\hat{\mu}_1) = 9,1035 + 0,0025 + 0,0066 + 0,0029(si\ Fuel = 2) + 0,0071(si\ Period = 1) - 0,0692(si\ Coverage = 2) + 0,0889(si\ Coverage = 3) + 0,0723(si\ Sport = 2) - 0,0327(si\ Fleet = 2) + 0,5195 - 0,0168(si\ Sex = 2)$

$\hat{\sigma}_1 = e^{0,3492} = 1,4179$

$\hat{\pi}_1 = 0,966191$
*componente* $2 : \hat{\mu}_1\ ln(\hat{\mu}_1) = \hat{\gamma_0} + \hat{\gamma_1}\ Ageph_i + \hat{\gamma_2}\ Agec_i + \hat{\gamma_3}(si\ Fuel = 2) + \hat{\gamma_4}(si\ Period = 1) - \hat{\gamma_5}0.28928(si\ Coverage = 2) + \hat{\gamma_6}(si\ Coverage = 3) + \hat{\gamma_7}(si\ Sport = 2) + \hat{\gamma_8}(si\ Fleet = 2) + \hat{\gamma_9}\ Nclaim_i + \hat{\gamma_{10}}(si\ Sex = 2)$

$ln(\hat{\mu}_2) = 7,7275 + 0,0072 + 0,0328 + 0,1056(si\ Fuel = 2) + 0,2679(si\ Period = 1) - 0,2681(si\ Coverage = 2) - 0,2738(si\ Coverage = 3) + 0,7672(si\ Sport = 2) + 0,7463(si\ Fleet = 2)\ + 0,5968 + 0,0256(si\ Sex = 2)$

$\hat{\sigma}_2 = e^{1,219} = 3,3838$

$\hat{\pi}_2 = 0,03380897$

En la componente 2 se observa la cobertura de tipo 3 la cual tiene un efecto menor que la componente 1, esto no describe bien el comportamiento debido a que la cobertura con mayor valor promedio en severidad es la cobertura 3.

| Cobertura del seguro vs severidad | |
|---|---|
| Tipo | Promedio severidad |
| TPL | 79.965 |
| PO | 58.791 |
| FO | 96.013 |

### 4.2.5. MODELO PARETO-LOG NORMAL

$$\hat{f}(y_i) = \hat{\pi}_1 f_1(\hat{\mu}_{1i}, \hat{\sigma}_1) + \hat{\pi}_2 f_2(\hat{\mu}_{2i}, \hat{\sigma}_2)$$

*componente* $1 : \hat{\mu}_1\ ln(\hat{\mu}_1) = \hat{\beta_0} + \hat{\beta_1}\ Ageph_i + \hat{\beta_2}\ Agec_i + \hat{\beta_3}(si\ Fuel = 2) + \hat{\beta_4}(si\ Period = 1) - \hat{\beta_5}(si\ Coverage = 2) + \hat{\beta_6}(si\ Coverage = 3) + \hat{\beta_7}(si\ Sport = 2) + \hat{\beta_8}(si\ Fleet = 2) + \hat{\beta_9}\ Nclaim_i + \hat{\beta_{10}}(si\ Sex = 2)\ ln(\hat{\mu}_1) =$

$7,6549 + 0,0076 + 0,0344 + 0,1159(si\ Fuel = 2) + 0,2789(si\ Period = 1) - 0,2774(si\ Coverage = 2) - 0,2812(si\ Coverage = 3) + 0,8030(si\ Sport = 2) + 0,7870(si\ Fleet = 2) + 0,5990 + 0,0257(si\ Sex = 2)$

$\hat{\sigma}_1 = e^{1,235} = 3,4383$

$\hat{\pi}_1 = 0,03187405$

*componente* $2 : \hat{\mu}_1\ ln(\hat{\mu}_1) = \hat{\gamma_0} + \hat{\gamma_1}\ Ageph_i + \hat{\gamma_2}\ Agec_i + \hat{\gamma_3}(si\ Fuel = 2) + \hat{\gamma_4}(si\ Period = 1) - \hat{\gamma_5}0.28928(si\ Coverage = 2) + \hat{\gamma_6}(si\ Coverage = 3) + \hat{\gamma_7}(si\ Sport = 2) + \hat{\gamma_8}(si\ Fleet = 2) + \hat{\gamma_9}\ Nclaim_i + \hat{\gamma_{10}}(si\ Sex = 2)$

$ln(\hat{\mu}_2) = 9,1031 + 0,0025 + 0,0066 + 0,0028(si\ Fuel = 2) + 0,0073(si\ Period = 1) - 0,0692(si\ Coverage = 2) + 0,0881(si\ Coverage = 3) + 0,0727(si\ Sport = 2) - 0,0324(si\ Fleet = 2) + 0,5197 - 0,0166(si\ Sex = 2)$

$\hat{\sigma}_2 = e^{0,3508} = 1,4202$

$\hat{\pi}_2 = 0,968126$

Al realizar el análisis de los tres modelos propuestos, el modelo que da una mejor interpretabilidad es el modelo gamma donde se observa para la componente 1 que entre más edad de la persona la severidad aumenta 1,04 veces, al igual que la antigüedad del vehículo, esto debido a que entre más años tenga un vehículo puede tener mayor siniestralidad generando que la severidad aumente. Los automóviles que utilizan gasolina Diesel también tienen un efecto positivo esto se debe a que en el año 1997 lo automóviles que utilizaban este tipo de combustible eran autos grandes de carga, por otro lado las personas que pagan en más de una cuota su seguro aumenta el valor de la severidad 8,82 veces, la cobertura de tipo 2 tiene un efecto negativo en comparación con la cobertura tipo 1, debido a que la muestra cuenta con más del 50 % de los datos con cobertura 1, a diferencia de la cobertura 3 que por el tipo de cobertura aumenta el valor, si el vehículo no es sport aumenta el valor de la severidad debido a que los automóviles sport se

encuentran en un 1 % de la muestra, los hombres para este caso tiene un efecto positivo a comparación de las mujeres.

# 5. Conclusiones y trabajos futuros

- La mixtura de distribuciones gamma es el modelo que mejor interpretabilidad da, este modelo nos muestra el comportamiento de las dos componentes donde podemos observar que para la muestra el cambio entre conclusiones es de género. Las otras distribuciones que se ajustaban a estos datos no dan una buena interpretación, el modelo Lognormal y Pareto-Lognormal nos dice que la cobertura 1 es la cobertura que más afecta en la componente de los automóviles con severidad alta, por esto se descartan y se toma el modelo gamma.
- En la muestra de estos datos se observa que en Bélgica las personas en su mayoria prefieren pagar una cobertura TPL, el 71 % de la muestra es del género masculino, solo el 1 % de los automoviles son Sport.
- Para futuros trabajos se puede proponer otro algoritmo de maximización diferente al EM, para así poder comparar si se obtienen mejores resultados debido a que este algoritmo tiene un alto costo computacional.

# 6. Agradecimientos

A mis padres Carlos y Nancy por su apoyo incondicional, a mi tutor Wilmer Pineda por su disposición y ayuda durante este proceso, a todos mis profesores y en especial a Deisy Camargo por su orientación y apoyo en este trabajo; a todas las personas que me acompañaron en esta etapa, aportando a mi crecimiento tanto profesional y como personal.

# Referencias

[1] Cid,C(1999). *Estudio de la siniestralidad y aplicaciones económicas en las entidades aseguradoras* Resumen del texto de la beca concedida por la Fundación MAPFRE[PDF],Recuperado de https://www.fundacionmapfre.org/documentacion/publico/i18n/catalogoimagenes

[2] Díaz,C(2017). *Factores de riesgo que afectan la severidad de los accidentes de trafico en áreas urbanas. El caso de cartagena, Colombia.* Trabajo Master[PDF],Recuperado de http://manglar.uninorte.edu.co/bitstream/handle/10584/8207/130424.pdf?sequence=1isAllowed=y

[3] Ding,C.S(2006).*Practical Assessment,Research & evaluation*Vol. 11.

[4] Gómez,L(2014). *Modelos de mixturas finitas para la caracterización y mejora de la redes de monitorización de la calidad del aire* Trabajo Master[PDF],Recuperado de https://www.masteres.ugr.es/moea/pages/curso201314/tfm1314/memoriagomezlosadaalvaroweb/!

[5] Guerrero,C & Hiraldo,M(2015). *Los siniestros en el seguro del automóvil: un análisis econométrico aplicado* vol.4,Recuperado de: https://www.redalyc.org/pdf/301/30123117.pdf

[6] Kenneth E. Train, *Discrete Choice Methods with Simulation*, segunda edicion, Cambridge University Press,Inglaterra,2009.

[7] McLachlan,G & Peel,D *Finite Mixture Models*,Primera edición, Wiley-Interscience publication,2000.

[8] Mikis D. Stasinopoulos, Robert A. Rigby, Gillian Z. Heller, Vlasios Voudouris & Fernanda De Bastiani, *Flexible Regression and Smoothing Using GAMLSS in R*, Primera edición, Taylor & Francis Group,Inglaterra,2017.

[9] Molina & Jimenéz (2014). *Valoración de derivados europeos con mixtura de distribuciones Weibull* Articulo[PDF],Recuperado de https://www.scielo.org.co/pdf/ceco/v34n65/v34n65a04.pdf

[10] Roldan,F(2014). *Mixtura de distribuciones normales incluyendo modelado conjunto de media y varianza desde un enfoque clásico* Trabajo Master[PDF],Recuperado de https://www.http://bdigital.unal.edu.co/61129/1/80112212.2014.pdf