Instituto Tecnológico Autónomo de México

CREDIBILIDAD BAYESIANA PARA EL CÁLCULO DE UNA PRIMA NIVELADA DE GASTOS MÉDICOS

Tesis

que para obtener el título de

Licenciado en Actuaría

Presenta

Daniel Maximiliano Guerrero Meneses

Asesora

Dra. Maria Mercedes Gregorio Domínguez

Ciudad de México 2024

≪Con fundamento en los artículos 21 y 27 de la Ley Federal del Derecho de Autor y como titular de los derechos moral y patrimonial de la obra titulada "**Credibilidad Bayesiana para el cálculo de una prima nivelada de gastos médicos**", otorgo de manera gratuita y permanente al Instituto Tecnológico Autónomo de México y a la Biblioteca Raúl Bailléres Jr., la autorización para que fijen la obra en cualquier medio, incluido el electrónico, y la divulguen entre sus usuarios, profesores, estudiantes o terceras personas, sin que pueda percibir por tal divulgación una contraprestación.≫

Fecha

Daniel Maximiliano Guerrero Meneses

A todos mis seres queridos.

Agradecimientos

A mis padres, por todo el amor, el cariño y el acompañamiento incondicional; por ser un gran ejemplo de personas y de pareja para mí y para mis hermanas; por guiarme siempre hacia lo mejor e inspirarme todos los días con su esfuerzo, dedicación y valores. Gracias por enseñarme el valor del trabajo duro y la perseverancia, y por ser mi pilar fundamental en cada paso de este camino académico.

A mis hermanas, por ser mis compañeras de vida desde nuestros primeros pasos; por todas las risas, las aventuras compartidas y las enseñanzas mutuas. Gracias por estar siempre ahí para mí, brindándome apoyo, comprensión y alegría en cada momento.

A mis abuelos, por su amor incondicional, su tiempo y su apoyo constante durante mis estudios. Gracias por sus sabios consejos, por las historias y enseñanzas que me han transmitido, y por ser un ejemplo de fortaleza y dedicación. Su presencia ha sido una fuente constante de inspiración y motivación.

A mi novia, Fer, por haber estado a mi lado no solo a lo largo de la licenciatura, sino también durante la redacción de esta tesis. Gracias

por tu calidez, tu comprensión y por inspirarme todos los días a ser mejor. Tu amor y apoyo han sido una fuente inagotable de fuerza y motivación, y no puedo expresar lo agradecido que estoy por tenerte a mi lado en este viaje.
A mis amigos, por todas las risas compartidas, por acompañarme en todo momento y por sacarme una sonrisa en los momentos difíciles o después de un examen agotador. Gracias por su apoyo incondicional, por las conversaciones sinceras y por ser un refugio de alegría y alivio en medio del estrés académico.

A la Dra. Mercedes Gregorio por ser mi asesora, por todas las lecciones y las correcciones durante la elaboración de este trabajo. Por último, gracias al Dr. Felipe Medina, M.C. David Ruelas y Dr. Leonardo Rojas por ser mis sinodales, por el tiempo y enseñanza durante este trabajo.

Muchas gracias a todos ustedes, esta tesis es más suya que mía.

Resumen

En los últimos años, se ha observado cómo la salud puede afectar la estabilidad financiera de las familias. El mercado asegurador ha creado un producto capaz de cubrir a las personas contra estos riesgos: el seguro de gastos médicos mayores.

Al mismo tiempo, se observa un encarecimiento de estos instrumentos financieros debido a la pandemia ocasionada por el *COVID-19*. Esta pandemia ha desencadenado una serie de impactos económicos como periodos hiperinflacionarios y desaceleración económica debido a la suspensión de actividades del programa "Jornada Nacional de sana distancia"teniendo un impacto en sectores como minería, contrucción, automotriz y aeroespacial (Esquivel, 2020). Como respuesta a este incremento en los precios, la propuesta de la investigación es mantener el precio del seguro constante por tres años.

La información con la que se cuenta para realizar el estudio son las bases de emisión de pólizas y pagos de siniestros de 2019 de una Institución de Salud Especializada. Esta compañía opera exclusivamente la operación de Accidentes y Enfermedades. Como información adicional, se cuenta con el Sistema Electrónico de siniestros y Avisos en Materia de Salud del año 2019 que proporciona la Comisión Nacional de Seguros y Fianzas.

Índice general

Introducción

El seguro de gastos médicos mayores es un instrumento financiero que ayuda a las personas a cubrirse contra pérdidas financieras potenciales derivadas de atención médica hospitalaria. En los últimos años, se ha observado un incremento en los precios de los seguros de gastos médicos derivado de la pandemia ocasionada por el virus *SARS-CoV-2* o *COVID-19* (WTW, 2022), además de los factores que anteriormente encarecían el producto como la edad, la inflación médica, el sexo y cualquier otra variable que utilicen las compañías de seguros para tarificar un seguro de gastos médicos.

El incremento en los precios hace que cada año que pasa sean más inaccesibles los seguros para los consumidores finales. Al mismo tiempo, se estima que en México, únicamente el 9.9 % (AMIS, 2023) de la población cuenta con un seguro de gastos médicos mayores.

Con el fin de incentivar la compra de seguros de gastos médicos, el tema central de la investigación será lograr una prima neta nivelada por tres años. Esto quiere decir que la prima del seguro no aumentará durante ese periodo.

Se cuenta con información de una compañía de seguros que opera exclusivamente la operación de Accidentes y Enfermedades por estar considerada como una Institución de Salud Especializada (ISES). La

aseguradora ha brindado información sobre su emisión y su siniestralidad con la finalidad de tarificar el producto. Los archivos anteriores contienen información histórica desde 2016 hasta 2019. La tabla de emisión contiene por asegurado datos como la edad, fecha de emisión, fecha de vencimiento, sexo y suma asegurada contratada. Por otro lado, la tabla de siniestros tienen la cantidad de reclamaciones, sexo, fecha de reclamación, reserva al momento de la reclamación y ajustes posteriores

Sin embargo, por ser una compañía de seguros relativamente joven comparada con el mercado, cuenta con poca experiencia para tarificar el producto. Es por eso que se requieren herramientas de inferencia distintas a las de la estadística frecuentista.

Se debe tener presente, que la estadística frecuentista basa sus resultados en la repetición de n eventos cuando n tiende a infinito. Debido a lo anterior, y a que no se cuenta con un volumen grande de información para suponer que la cantidad de observaciones es lo suficientemente grande para satisfacer las hipótesis de la estadística clásica, es que se requerirá el uso de Estadística Bayesiana.

Más adelante se profundizará en los fundamentos de la Estadística Bayesiana. Por ahora, basta con mencionar que la información adicional a la información provista por la compañía es el Sistema Estadístico del Sector Asegurador (SESA) de 2019.

Capítulo 1

Composición de cartera

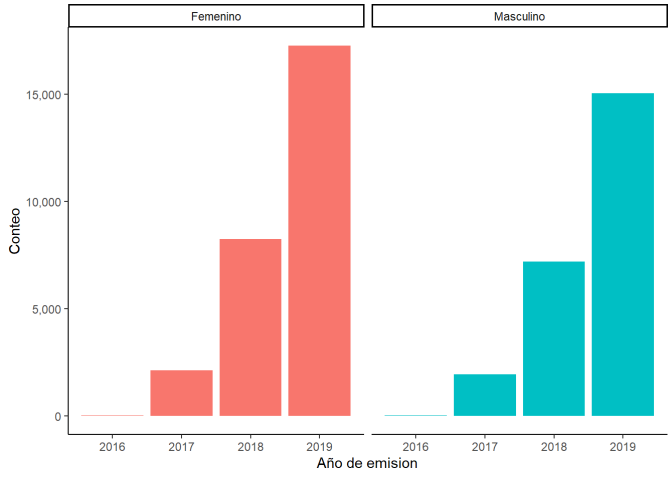
Antes de comenzar a hacer análisis de frecuencia y severidad, es necesario tener cierto conocimiento previo de la cartera. Para esto, se realizará un breve análisis exploratorio de las bases de vigores, emisión y siniestros históricos compartidas por la compañía.

De aquí en adelante, se llamará a la compañía como *Alfa seguros*. Se tendrá interés particular por conocer la distribución de edades en su emisión, la distribución de los siniestros, las sumas aseguradas emitidas, la distribución por sexo de la cartera, montos de reclamación por sexo, entre otros factores que conducirán a una exploración más exhaustiva y así poder modelar más puntualmente.

1.1. Base de emisión

En primera instancia, será de interés averiguar la cantidad de pólizas emitidas a lo largo de la historia de la compañía. Los datos se encuentran filtrados para no contemplar las cancelaciones y únicamente el ramo de Gastos Médicos mayores individuales.

Figura 1.1. Emisiones por año de emisión y sexo



Fuente: Elaboración propia con datos de Alfa seguros (2019)

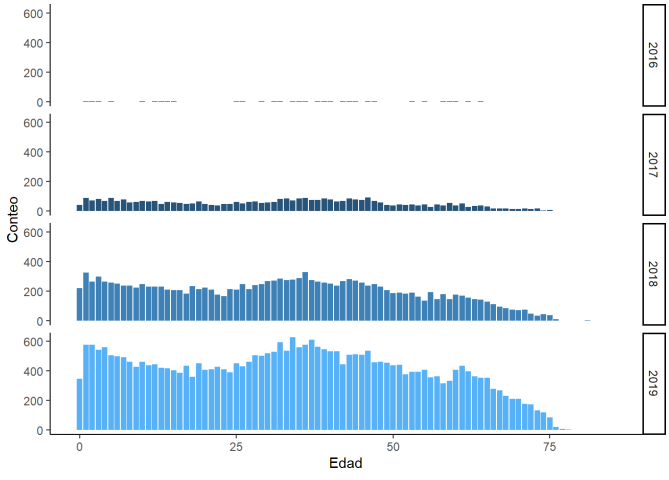
Cuadro 1.1. Resultados de emisión por año

Año de emisión	Femenino	Masculino
2016	23	27
2017	2,127	1,941
2018	8,237	7,194
2019	17,243	15,026

El figura 1.1. y el cuadro 1.1. ayudan a notar un crecimiento sostenido en la emisión de pólizas a lo largo del tiempo. Al ser una aseguradora emergente y que apenas está empezando a adquirir clientes, sería razonable pensar que seguirán creciendo. Al mismo tiempo, se puede observar que parece ser poca la diferencia en la cantidad pólizas emitidas para hombres y mujeres. Sin embargo, el caso será estudiado más adelante.

Será relevante revisar no sólo por sexo, sino por edad. El conteo por edades ayudará a reconocer si existen *huecos* en las gráficas. La falta de información sobre algunas edades podría complicar el análisis para esa edad particular.

Figura 1.2. Emisiones por año de emisión y edad

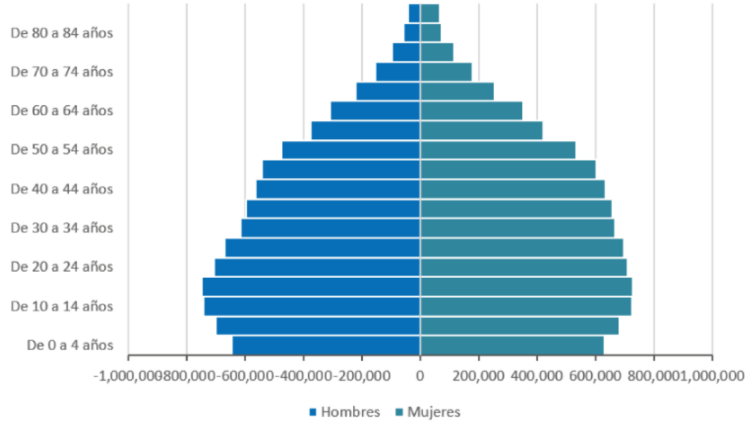


Fuente: Elaboración propia con datos de Alfa seguros (2019)

Como se puede observar, no existen edades sin asegurados en la base de emisión. Aparentemente, existen suficientes asegurados para realizar un análisis estadístico bajo un enfoque frecuentista. Sin embargo, no es posible asegurarlo todavía.

Se puede notar una leve caída en la emisión por edad en la medida en que avanzan las edades. Sin embargo, esta caída no es una particularidad de la cartera, sino de la población en general.

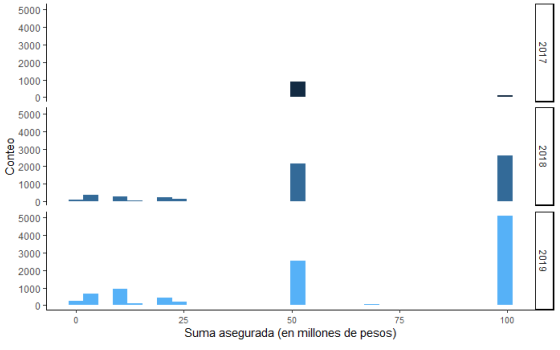
Figura 1.3. Piramide poblacional



Fuente: Consejo Estatal de Población del Estado de México (2020)

Por otro lado, es importante observar las sumas aseguradas emitidas, ya que éstas dan un indicador de exposición al riesgo.

Figura 1.4. Emisiones suma asegurada



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Existen dos sumas aseguradas que dominan la emisión de la cartera de

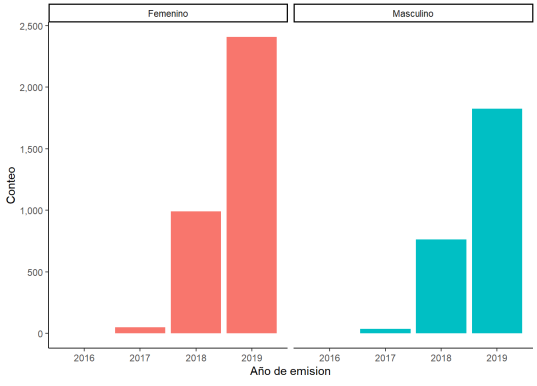
Alfa Seguros: 50 y 100 millones de pesos. Sin importar el año en que fue emitida la póliza, estas sumas aseguradas permanecen constantes. Más adelante, se elegirá una de estas dos sumas aseguradas.

Con esta gráfica, concluye un primer acercamiento a la base de emisión de Alfa seguros. Se puede concluir que la emisión es dependiente del sexo del asegurado, que las edades de los asegurados se deben en gran medida a la pirámide poblacional y que existen dos principales sumas aseguradas preferidas por los asegurados.

1.2. Base de siniestros

Para la base de siniestros, será de interés estudiar, en un primer momento, el comportamiento de las reclamaciones en conteo por sexo. Esto ayudará a reconocer si alguna de las dos categorías tiene un mayor número de reclamaciones.

Figura 1.5. Número total de reclamaciones por sexo

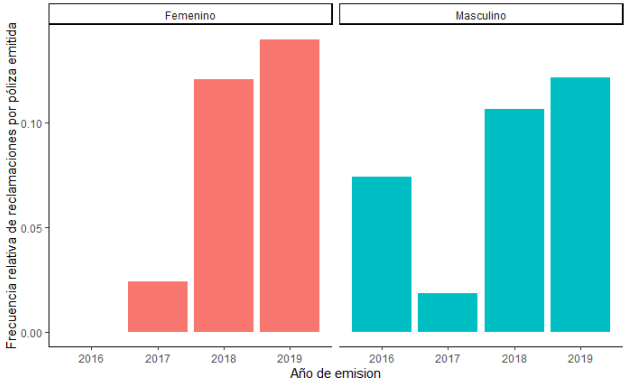


Fuente: Elaboración propia con datos de Alfa seguros (2019)

Debido a que no se ha alcanzado una estabilidad en la cantidad de

reclamaciones; es decir, que en todos los años existe una alta variabilidad respecto a la observación inmediata anterior, podría ser que la cartera no haya alcanzado la madurez suficiente para considerar realizar un análisis frecuentista Se observará más adelante, pero el número de reclamaciones por edad no es suficiente para cumplir las hipótesis de la estadística clásica. Note que, aunque se tienen poco más de 4 mil observaciones en reclamaciones, hay que dividir esas observaciones por edades.

Figura 1.6. Frecuencia relativa de reclamaciones condicional por año, póliza y por sexo



Fuente: Elaboración propia con datos de Alfa seguros (2019)

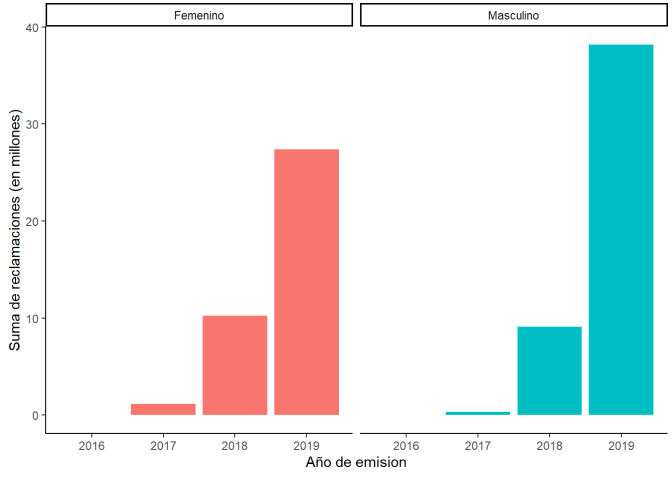
Se espera que, a lo largo del tiempo, el cociente del número de reclamaciones y el número de pólizas emitidas permanezca prácticamente constante. Aunque las variaciones parezcan grandes y porcentualmente lo sean, note que éstas no son más del 0.05 respecto al año inmediato anterior por lo que parece que la cartera ha alcanzado una estabilidad en estos últimos años.

Cuadro 1.2. Resultados de siniestros

Año de emisión	Sexo	Conteo	Número de reclamaciones promedio por póliza
2016	Femenino	0	0
2016	Masculino	2	0.074
2017	Femenino	51	0.024
2017	Masculino	36	0.019
2018	Femenino	992	0.120
2018	Masculino	765	0.106
2019	Femenino	2,407	0.140
2019	Masculino	1,825	0.121

A continuación, se observa cómo se comporta el monto de reclamaciones por sexo.

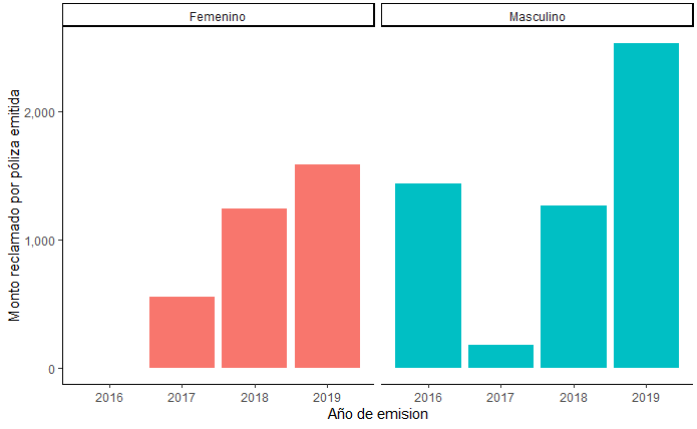
Figura 1.7. Monto total de reclamaciones por sexo



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Se puede notar que, aparentemente, existe un mayor monto reclamado para los hombres que para las mujeres. Sin embargo, hay movimiento a lo largo de los años. Es decir, en la cartera de Alfa Seguros, el monto reclamado de los hombres se encuentra por encima del de las mujeres en el último año. Por otro lado, en 2018 las reclamaciones hechas por mujeres exceden a las de los hombres.

Figura 1.8. Monto de reclamaciones por póliza y por sexo



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Así como en el conteo de reclamaciones, lo que se espera es que cuando la compañía entre en una etapa de madurez se pueda ver una razón constante. En este caso esa razón es entre el monto de los siniestros y la cantidad de pólizas emitidas.

Note que la caída en la cantidad de reclamaciones observadas en el año 2017 de la figura 1.6. afecta al monto de reclamaciones del mismo año. Sin embargo, lo que debería ser un poco alarmante es la alta siniestralidad reportada por el sexo masculino en el año 2019.

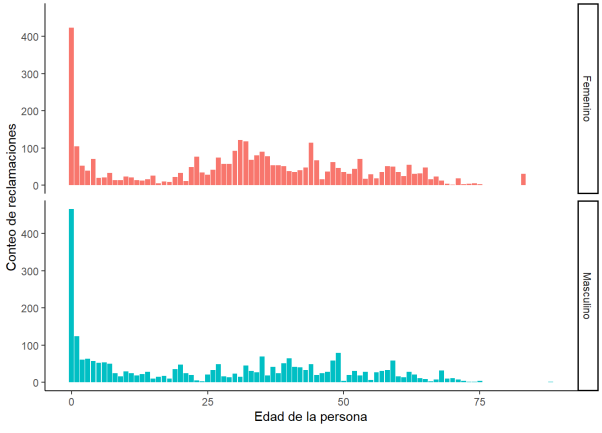
Se podría pensar que la siniestralidad de la cartera es aún muy

inestable y no conviene tomar la experiencia de la compañía para tarificar un seguro. Sin embargo, ahí es donde podría ser de ayuda usar Estadística Bayesiana.

Cuadro 1.3. Resultados de monto reclamado

Año de emisión	Sexo	Monto reclamado nominal	Monto reclamado real
2016	Femenino	0	0
2016	Masculino	38,836.56	44,523.09
2017	Femenino	1,181,239.81	1,290,134.16
2017	Masculino	348,738.24	380,887.19
2018	Femenino	10,246,214.44	10,661,356.41
2018	Masculino	9,107,109.60	9,476,098.89
2019	Femenino	27,378,779.02	27,378,779.02
2019	Masculino	38,149,264.46	38,149,264.46

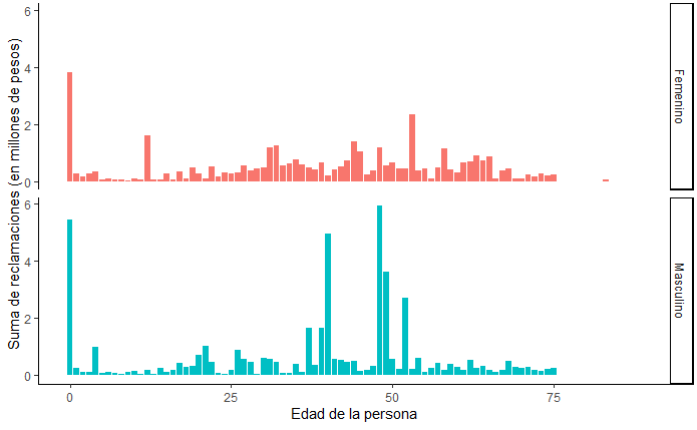
Figura 1.9. Número total de reclamaciones por edad



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Existe un número de reclamaciones grandes en edades tempranas para esta cartera. Esto se va a traducir en una tarificación más alta en edades tempranas que en edades intermedias. Esto se debe a las enfermedades congénitas con las que puede nacer una persona. Las enfermedades congénitas pueden definirse como anomalías estructurales o funcionales que ocurren durante la vida intrauterina de la persona. Estas pueden ser detectadas en el periodo prenatal, durante el parto o posterior a la primer infancia (OMS, 2023). Estos padecimientos pueden llegar a provocar la muerte en los primeros años de vida de las personas y son ocasionados por factores genéticos, socioeconómicos, demográficos y ambientales. En la nota técnica de *Alfa seguros* se mencionan algunos ejemplos de atención a enfermedades congénitas que van desde los tres mil hasta los 19 mil pesos. Sin embargo, este costo es solo el comienzo debido a que, en algunos casos, puede requerirse de apoyo de larga duración como fisioterapias, logoterapias, ergoterapias entre otras atenciones (OMS, 2023). Incluso la Comisión Nacional de Seguros y Fianzas, en su tabla de mortalidad, hace mención de esta fragilidad durante los primeros años de vida.

Figura 1.10. Monto total de reclamaciones por edad



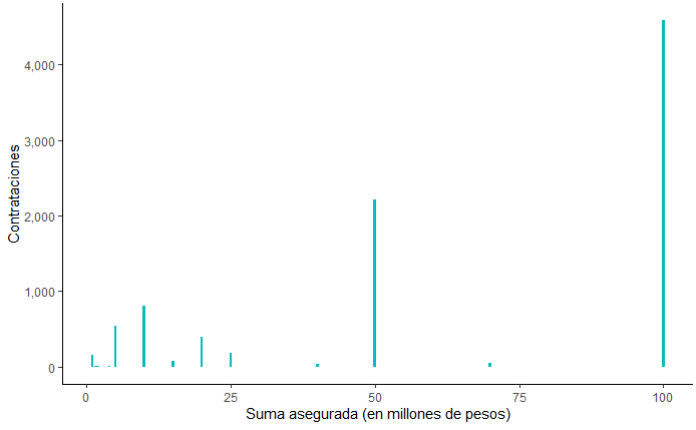
Fuente: Elaboración propia con datos de Alfa seguros (2019)

Es evidente que el monto de siniestralidad de las edades tempranas es por la frecuencia con la que las personas de esas edades hacen reclamaciones. Sin embargo, existen otros dos picos para los hombres. Estos dos picos son en edades medias entre 40 y 50 años. Esto encarecerá un poco los seguros de estas edades.

1.3. Vigor de diciembre 2019

La base de vigores de 2019 contiene información de pólizas vigentes con corte al 31 de diciembre de 2019. En este caso, únicamente se revisará la suma asegurada contratada por los asegurados. Con esta información se confirmará la hipótesis anterior sobre la preferencia en estos montos de suma asegurada.

Figura 1.11. Monto total de reclamaciones por edad



Fuente: Elaboración propia con datos de Alfa seguros (2019)

En la gráfica anterior se puede confirmar la observación de la figura 1.4. Ya sea que Alfa seguros emita principalmente pólizas de 50 y 100 millones o que sea por preferencia de los asegurados, deben tomarse estas dos como un indicador de exposición al riesgo y, más adelante, elegir una para la tarificación de nuestro seguro.

Capítulo 2

Análisis de frecuencia

2.1. Análisis exploratorio

Antes de iniciar con el análisis de frecuencia, vale la pena analizar cómo se comporta la base. La base de *Frecuencia* fue construida uniendo las bases de emisión con la de siniestros.

La base de emisión contiene la fecha de nacimiento, estado, sexo, suma asegurada, prima neta, inicio de vigencia, fin de vigencia, entre otros. Por otro lado, la base de siniestros contiene la antigüedad, el ramo, el monto del reclamo, deducible, coaseguro, edad, entre otros.

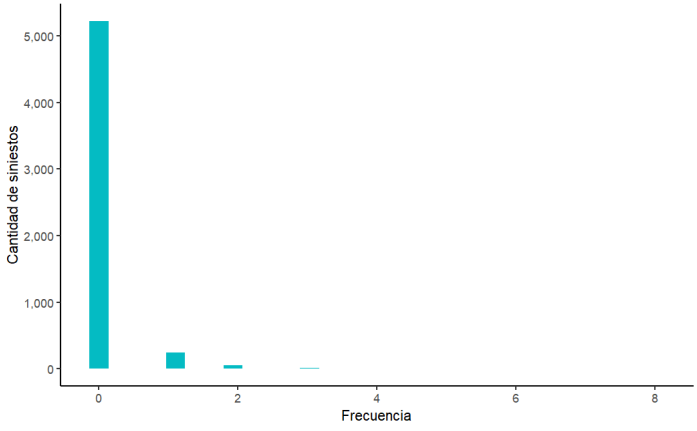
Al juntar las bases, se pueden filtrar los campos que se considerarán relevantes. En este caso particular, la base será segmentada por: edad, estado de la república, sexo. Al mismo tiempo, al realizar la intersección de las bases, puede contarse la cantidad de veces que se repite una póliza que se encuentra en la base de emisión, en la base de siniestros. Eso es útil para hacer un conteo de la cantidad de siniestros que tuvo una póliza a lo largo del año.

Existen pólizas que, al hacer la intersección, no se encuentran en la

base de siniestros, pero sí en la de emisión. Ese resultado indica que esas pólizas tuvieron cero siniestros a lo largo del año.

2.1.1. Análisis general

Figura 2.1. Distribución de reclamaciones

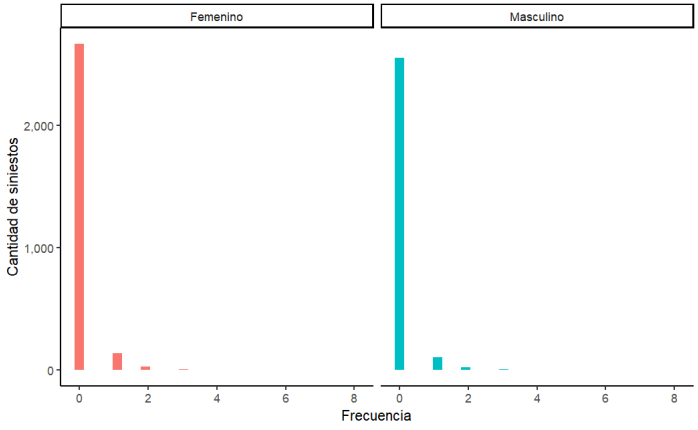


Fuente: Elaboración propia con datos de Alfa seguros (2019)

Los primeros resultados muestran lo usual en una cartera de seguros: la mayoría de los asegurados deben tener una muy baja o nula siniestralidad. La media de la frecuencia general es de 0.07 y su varianza 0.11.

2.1.2. Análisis por sexo

Figura 2.2. Distribución de reclamaciones por sexo



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Cuadro 2.1. Media y varianza del número de reclamaciones segmentando por sexo

Sexo	Media	Varianza
Masculino	0.0659	0.1159
Femenino	0.0765	0.1115

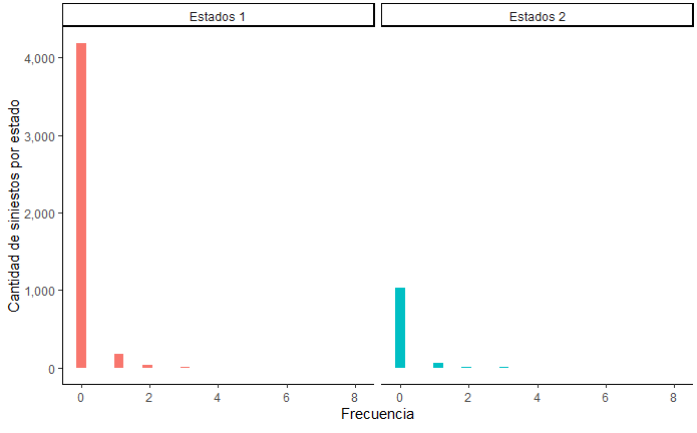
En un acercamiento más profundo a la base de frecuencia, tanto gráfica como numéricamente, los datos apuntan a que no existe una diferencia relevante entre la frecuencia con la que incurren en siniestros hombres y mujeres. Sin embargo, esta información se validará de manera posterior.

2.1.3. Análisis por estado

Para realizar el análisis de la frecuencia por estado de la República Mexicana no se considera la totalidad de los estados. Para propósitos de este estudio se tomarán dos claves de estados diferentes:

- Estados 1: los estados con esta clave son aquellos que cuentan con las ciudades más importantes del país. En esta ocasión fueron considerados la Ciudad de México, Nuevo León, Jalisco y Estado de México por ser considerado área metropolitana.
- Estados 2: el resto de los estados.

Figura 2.3. Distribución de reclamaciones por estado



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Cuadro 2.2. Media y varianza del número de reclamaciones segmentando por estado

Clave de estado	Media	Varianza
Estados 1	0.0646	0.1035
Estados 2	0.0980	0.1533

Debe tenerse cuidado con los resultados porque pueden ser contra intuitivos. A pesar de que en la figura 2.3 parece tener una masa más grande, note que ese desface es en los asegurados cuyo conteo de siniestros es 0.

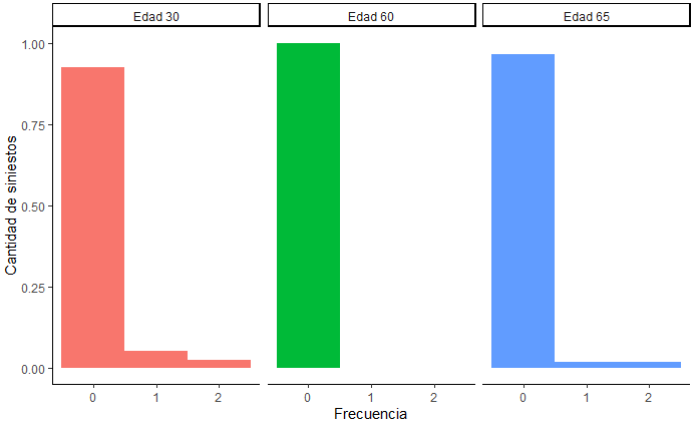
Es de notar que la media y la varianza de los *Estados 2* es muy similar.

2.1.4. Análisis por edad

Es complicado hacer un análisis exploratorio para cada una de las edades que aparecen en la base de datos de frecuencia, ya que prácticamente están todas las edades. Sin embargo, lo que se puede hacer es comparar algunas edades en un primer acercamiento para ver si éstas se parecen.

Para este ejemplo se van a usar las edades 30, 60 y 65. Las edades fueron elegidas arbitrariamente.

Figura 2.4. Distribución de reclamaciones por edad



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Cuadro 2.3. Media y varianza segmentando por edades

Edad	Media	Varianza
30	0.0977	0.1349
60	0	0
65	0.0545	0.0896

Tanto la figura 2.4 como el cuadro 2.3. son útiles para darse cuenta de dos cosas: las medias y las varianzas de la frecuencia por edades no son parecidas y existirán edades para las cuales no se registren reclamaciones. Por lo tanto, tendrán media cero y, más peligroso aún, varianza cero.

De los dos hallazgos, se pueden concluir dos cosas: la distribución de la frecuencia debe ser a través de datos agrupados para no estimar con media y varianza cero aquellas edades donde no haya evidencia de reclamaciones y que deben tener distribuciones diferentes según la edad

de las personas.

Antes de continuar, debe recordarse que éstos son resultados preliminares y que no necesariamente van a ser consistentes con un análisis más extenso de los datos.

2.2. Ajuste de modelos

2.2.1. Selección de modelos

Muchas veces, se pasa la mayor parte del tiempo intentando ver cuál es el modelo que mejor ajusta a los datos. Lo anterior puede hacerse tanto de manera paramétrica como no paramétrica. Al usar modelos paramétricos, existen varios métodos de calibración del modelo: momentos, cuantiles o verosimilitud.

La verosimilitud no es otra cosa que la probabilidad de observar un dato si el modelo es correcto. La función de verosimilitud se puede calcular sobre cualquier muestra. Sin embargo, en este trabajo se estará ocupando la función de verosimilitud de una muestra aleatoria. Es decir, sus observaciones son consideradas independientes e idénticamente distribuidas. Una vez obtenida la función de verosimilitud, se contruye una transformación monótona, creciente y derivable (el logaritmo). Esta transformación es conocida como log-verosimilitud. Finalmente, se maximiza la log-verosimilitud variando los parámetros que se desean estimar. Como ya se ha mencionado, lo que se está maximizando es la probabilidad de observar la muestra dado que los datos vienen del modelo propuesto.

Se dice que un modelo es más simple que otro si la cantidad de parámetros por estimar es menor. Es decir, un modelo con tres parámetros es más complejo que uno con dos parámetros. Dada una

muestra de tamaño n, a medida que el investigador aumenta la cantidad de parámetros del modelo, la verosimilitud aumenta también. Sin embargo, lo que realmente interesa es que la ganancia en verosimilitud (o log-verosimilitud) sea lo suficientemente grande como para justificar aumentar la complejidad del modelo.

Es decir, un buen modelo, además de explicar de una buena manera los datos de la muestra, es también simple. Se debe diferenciar entre un modelo simple y un modelo simplista. Sería ideal tener un modelo que aproxime bien los datos y que, por principio de parsimonia, sea lo más simple posible (la menor cantidad de parámetros).

Para hacer lo anterior, se compararán los modelos de este capítulo con los criterios de información Akaike y Bayesiano. Al mismo tiempo, se puede hacer una prueba de hipótesis llamada *cociente de verosimilitudes* la cual converge en distribución a una distribución χ^2.

Cuadro 2.4. Grupos de edades

Criterio o prueba	Estadístico
Akaike (AIC)	$\text{AIC} = 2K - 2\ell$
Bayesiano (BIC)	$\text{BIC} = 2ln(n) - 2\ell$
Cociente de verosimilitudes	$\chi^2 = 2(\ell_{\Theta_1} - \ell_{\Theta_0})$

Donde:

- K: es la cantidad de parámetros que estima el modelo
- n: es la cantidad datos con la que cuenta el modelo
- ℓ: es el resultado de la log-verosimilitd en el estimador de máxima verosimilitud
- ℓ_{Θ_1}: es el resultado de la maximización de la log-verosimilitd del modelo más complejo

- ℓ_{Θ_2}: es el resultado de la maximización de la log-verosimilitd del modelo más simple.

La prueba de hipótesis para el cociente de verosimilitudes se vería como sigue.

$$H_0 : \text{El modelo más simple es adecuado } vs.$$

$$H_1 : \text{ El modelo más complejo es adecuado}$$

Sea χ^2 , definido como se muestra en cuadro 2.4, el estadístico de prueba, se rechaza H_0 si $\chi^2 > \chi^2_{(1)(1-\alpha)}$ donde α es un nivel de significancia dado.

2.3. Modelos lineales generalizados

Para la frecuencia, se optará por usar métodos lineales generalizados para su estudio. Los modelos lineales generalizados, al igual que los modelos lineales, buscan expresar la relación entre una variable por explicar (o variable observada) Y y variables explicativas $\underline{X} = (X_1, X_2, ..., X_n)$.

Los modelos lineales generalizados son una extensión de los modelos de regresión tradicional donde se *generaliza* la variable observada (Y). Tradicionalmente, se dice que el error sigue una distribución $N(0, \sigma^2)$. Sin embargo, ahora la distribución de éste pertenecerá a la familia exponencial. Misma que incluye a la distribución normal.

Además, otra característica importante de los modelos lineales generalizados, *glm* por sus siglas en inglés, es una transformación en la media de la variable observada que la convierte en lineal. A esta función se le conoce como *liga*.

Antes de seguir profundizando en el tema de familias exponenciales y funciones liga, se mostrará en términos matriciales la diferencia entre los modelos lineales tradicionales y los generalizados:

$$\underline{Y} = \mu + \underline{\epsilon}, \quad donde \quad \mu = X\beta, \quad \epsilon_i \sim N(0, \sigma^2)$$

$$\underline{Y} = \mu + \underline{\epsilon}, \text{ donde } g(\mu) = X\beta, \quad \epsilon_i \text{ miembro de la familia exponencial}$$

Finalmente, la ventaja principal de usar modelos lineales generalizados para la estimación de la frecuencia es que permite restringir el rango de las variables observadas. En este caso, la restricción irá en el sentido de que las variables son discretas y no negativas.

2.3.1. Familia exponencial

Definición 1. *Se dice que una distribución pertenece a la familia exponencial si su función de densidad se puede escribir de la forma (Dobson & Barnett, 2008):*

$$f(y|\lambda) = e^{a(y)b(\theta)+c(\theta)+d(y)}$$

Recordando que existen tres distribuciones clásicas para la frecuencia (Poisson, Binomial y Binomial Negativa), se usarán esas tres distribuciones y será demostrado que efectivamente pertenecen a la familia exponencial.

Poisson

Si $y \sim Poisson(\theta), \theta > 0$

$$\Rightarrow f(y|\theta) = \frac{\theta^y e^\theta}{y!}, y = 0, 1, 2, ...$$

Sea:

$$a(y) = y, b(\theta) = ln(\theta), c(\theta) = -\theta, d(y) = -ln(y!)$$

$$\Rightarrow f(y|\theta) = \frac{\theta^y e^{-\theta}}{y!} = e^{yln(\theta)-\theta-ln(y!)}$$

$\therefore y|\theta \sim Poisson(\theta)$ pertenece a la familia exponencial.

Binomial

Si $y \sim Bin(n, \theta)$ con n conocida $0 < \theta < 1$

$$\Rightarrow f(y|n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y}, y = 0, 1, 2, ..., n$$

Sea:

$$a(y) = y, b(\theta) = ln\left(\frac{\theta}{1-\theta}\right), c(\theta) = nln(1-\theta), d(y) = ln\binom{n}{y}$$

$$\Rightarrow f(y|n,\theta) = \binom{n}{y}\theta^y(1-\theta)^{n-y} = e^{yln\left(\frac{\theta}{1-\theta}\right)+nln(1-\theta)+ln\binom{n}{y}}$$

$\therefore y|n, \theta \sim Bin(n, \theta)$ pertenece a la familia exponencial.

Binomial negativa

Si $y \sim BinNeg(r, \theta)$ con r conocida, $r > 0, 0 < \theta < 1$

$$\Rightarrow f(y|r,\theta) = \binom{y+r-1}{y}\theta^y(1-\theta)^r, y = 0, 1, 2, ...$$

Sea:

$$a(y) = y, b(\theta) = ln(\theta), c(\theta) = rln(1-\theta), d(y) = ln\binom{y+r-1}{y}$$

$$\Rightarrow f(y|r,\theta) = \binom{y+r-1}{y}\theta^y(1-\theta)^r = e^{yln(\theta)+rln(1-\theta)+ln\binom{y+r-1}{y}}$$

$\therefore y|r,\theta \sim BinNeg(r,\theta)$ pertenece a la familia exponencial.
Una vez demostrado que las distribuciones que clásicamente se ocupan para describir el comportamiento de la frecuencia pertenecen a la familia exponencial, se debe profundizar en la otra propiedad importante de los *glm*: las funciones liga.

2.3.2. Funciones liga

Definición 2. *La función liga es una función monótona diferenciable que relaciona el predictor lineal η con el valor esperado μ de un dato y. (Cullagh & Nelder, 1989)*

Note que, en los modelos de regresión clásicos, la media es lineal y, por lo tanto, la función liga sería la identidad. Sin embargo, para los casos que interesan estudiar (Poisson, Binomial y Binomial negativa), tomar la función identidad como liga podría resultar no tan efectivo debido a que, al ser distribuciones de conteo, su media será positiva, pero el predictor lineal podría no serlo.

Definición 3. *Se dice que una función es liga canónica para una distribución si ésta proviene del despeje de su parámetro canónico (Cullagh & Nelder, 1989).*

Se tendrá interés particular por encontrar las funciones liga para las distribuciones propuestas. Para encontrarlas, basta con fijarse en $b(\theta)$.

Cuadro 2.5. Funciones ligas canónicas

Distribución	Función liga
Poisson	$ln(\mu)$
Binomial	$ln\left(\frac{\mu}{1-\mu}\right)$
Binomial Negativa	$ln(\mu)$

Definición 4. *El parámetro canónico de una distribución no es otra cosa más que el despeje de la liga canónica respecto de otro al que se define como γ (Cullagh & Nelder, 1989).*

Se realiza un breve ejemplo con la distribución Binomial.

$$\gamma = ln\left(\frac{\theta}{1-\theta}\right)$$
$$\Rightarrow e^{\gamma} = \frac{\theta}{1-\theta}$$
$$\Rightarrow e^{\gamma}(1-\theta) = \theta$$
$$\Rightarrow e^{\gamma} = \theta(1+e^{\gamma})$$
$$\Rightarrow \theta = \frac{e^{\gamma}}{1+e^{\gamma}}$$

Intercambiando los parámetros para llegar a la función inversa:

$$\gamma = \frac{e^{\theta}}{1+e^{\theta}}$$

Es claro que para las otras dos distribuciones de interés, los parámetros canónicos resultan los siguientes:

Cuadro 2.6. Parámetros canónicos

Distribución	Parámetro canónico
Poisson	$\gamma = e^{\theta}$
Binomial	$\gamma = \frac{e^{\theta}}{1+e^{\theta}}$
Binomial Negativa	$\gamma = e^{\theta}$

2.3.3. *Offset*

Definición 5. *Un offset es un predictor estructural cuyo valor es relevante para el modelo, pero no necesita ser estimado. Es decir, su valor es conocido (Dobson & Barnett, 2008).*

En el caso particular de la estimación de frecuencia, el *offset* será el número de asegurados de la cartera.

2.3.4. Un primer acercamiento

Para dar una idea sobre el comportamiento de la base, se hará un primer ajuste con la distribución Poisson por ser la más simple (la de menor cantidad de parámetros). Para las edades, al no tener datos suficientes para darles un tratamiento individual, se agruparán haciendo saltos cada cinco años empezando con las personas de edad 1. El único grupo que tendrá más de cinco edades es el primero, ya que se le agregarán a las personas que tienen edad 0.

Es importante hacer notar que las variables sexo, grupo de edad y estado de residencia son politómicas. Por lo tanto, las variables explicativas tomarán únicamente los valores 0 o 1 dependiendo si cumple la condición. Es decir, se usará el siguiente modelo de regresión:

$$Y_i = e^{\beta_{0,i}+\beta_{1,i}X_{1,i}+\beta_{2,i}X_{2,i}+...+\beta_{n,i}X_{n,i}+\epsilon_i}$$

Donde, por ejemplo

$$X_{sexo,i} = \begin{cases} 1 & si \quad X_i \text{ es mujer} \\ 0 & si \quad X_i \text{ es hombre} \end{cases} \tag{2.1}$$

Más generalmente, se puede decir que

$$X_{j,i} = \begin{cases} 1 & si \quad X_i \text{ cumple la condición de pertenecer a } j \\ 0 & si \quad X_i \text{ no cumple la condición de pertenecer a } j \end{cases} \tag{2.2}$$

Note que estas variables caen en una de n categorías. De las n categorías anteriormente mencionadas, se le asignarán $n-1$ se

convertirán en indicadoras al construir una matriz de diseño. Aquella categoría a la que no se le asigne una indicadora tomará el nombre de *variable de referencia* y su efecto estará incluido en $\hat{\beta}_{0,i}$

Cuadro 2.7. Primer regresión Poisson

	$\hat{\beta}$	Error estándar	valor z	Pr(>|z|)
(Intercepto)	-2.0051	0.1486	-13.49	0.0000
Edad 2	-0.1999	0.2669	-0.75	0.4539
Edad 3	-0.8309	0.3789	-2.19	0.0283
Edad 4	-1.8977	0.5932	-3.20	0.0014
Edad 5	-1.0459	0.2921	-3.58	0.0003
Edad 6	-0.6316	0.1912	-3.30	0.0010
Edad 7	-0.6502	0.1863	-3.49	0.0005
Edad 8	-0.6450	0.1916	-3.37	0.0008
Edad 9	-0.7738	0.2135	-3.62	0.0003
Edad 10	-0.7279	0.2255	-3.23	0.0012
Edad 11	-0.8063	0.2722	-2.96	0.0031
Edad 12	-1.2002	0.3311	-3.63	0.0003
Edad 13	-0.8555	0.2669	-3.20	0.0014
Edad 14	-0.8066	0.5933	-1.36	0.1740
Edad 15	-0.3155	0.7204	-0.44	0.6614
Edad 16	-13.1517	1275.7539	-0.01	0.9918
Edad 17	1.3822	1.0098	1.37	0.1711
Estado 2	0.3978	0.1131	3.52	0.0004
Sexo M	-0.1458	0.1020	-1.43	0.1532

Como se puede observar, tanto el sexo como algunas edades de los asegurados parecen ser no significativos para el modelo. Sin embargo, existen parámetros que son muy significativos para el mismo. Lo anterior indica que es un buen acercamiento para la selección del modelo.

Al mismo tiempo, no se puede excluir a los grupos de edades que no son significativas para el modelo. Al ser la edad una variable que se considera relevante para la estimación de la frecuencia, lo que se hará será agregar más edades (no solo cinco años) a los grupos. Ahora los grupos no serán necesariamente simétricos. Los grupos abarcarán las edades como sigue:

Cuadro 2.8. Grupos de edades

Edades	Nombre del grupo
de 0 a 15	Edad 1
de 16 a 20	Edad 2
de 21 a 25	Edad 3
de 26 a 35	Edad 4
de 36 a 40	Edad 5
de 41 a 45	Edad 6
de 46 a 50	Edad 7
de 51 a 60	Edad 8
de 61 a 85	Edad 9

2.3.5. Modelo Poisson

Al hacer la agrupación por edades, como ya se mostró en el cuadro 2.8., el modelo resulta de una manera diferente. Antes de continuar, note que el grupo *Edad 1*, *Sexo F* y *Estado 1* está siendo considerado como grupo

de referencia.

Cuadro 2.9. Regresión Poisson con edades ajustadas

	$\hat{\beta}$	Error estándar	valor z	Pr(>|z|)
(Intercepto)	-2.1720	0.1253	-17.34	0.0000
Edad 2	-1.7321	0.5880	-2.95	0.0032
Edad 3	-0.8809	0.2814	-3.13	0.0017
Edad 4	-0.4764	0.1447	-3.29	0.0010
Edad 5	-0.4796	0.1747	-2.74	0.0061
Edad 6	-0.6082	0.1985	-3.06	0.0022
Edad 7	-0.5621	0.2113	-2.66	0.0078
Edad 8	-0.8087	0.2165	-3.74	0.0002
Edad 9	-0.6139	0.2289	-2.68	0.0073
Estado 2	0.4058	0.1128	3.60	0.0003
Sexo M	-0.1474	0.1020	-1.45	0.1483

Ahora se puede observar que casi todas las betas resultan significativas para el modelo que se va a escoger. Como ya se había interpretado en un primer acercamiento, la variable *Sexo* parece no estar aportando mucho al modelo. Sin embargo, para hacer un mejor análisis, se comparará paso a paso el modelo más complejo contra el más simple.

Es decir, por un lado, se tendrá al modelo propuesto en el cuadro 2.9. Por otro lado, el modelo más simple que se puede proponer es aquel que sólo tenga un intercepto. Se harán dos casos: comenzar por el modelo más simple e ir agregando variables que se encuentren como significativas para éste o comenzar por el modelo más complejo e ir eliminando variables.

En ocasiones, al hacer ambos métodos, no se llega a un mismo resultado. Sin embargo, se tiene el Criterio de Información de Akaike definido por comparar cuál es el mejor modelo. Note que siempre se preferirá el AIC más pequeño debido a que al maximizar una verosimilitud negativa y penalizando con una cantidad positiva que depende de la cantidad de parámetros.

En este caso, sin importar si se comienza a partir del modelo más complejo o el más simple, el método muestra que el mejor modelo es aquel que estima a través de la edad y el estado en el que vive la persona. Es decir, se ha eliminado un parámetro no significativo: el sexo.

Cuadro 2.10. Mejor regresión Poisson según AIC

	$\hat{\beta}$	Error estándar	Valor z	Pr(>|z|)
(Intercepto)	-2.2444	0.1156	-19.41	0.0000
Edad 2	-1.7380	0.5879	-2.96	0.0031
Edad 3	-0.8628	0.2811	-3.07	0.0021
Edad 4	-0.4616	0.1443	-3.20	0.0014
Edad 5	-0.4771	0.1747	-2.73	0.0063
Edad 6	-0.6097	0.1985	-3.07	0.0021
Edad 7	-0.5693	0.2113	-2.69	0.0070
Edad 8	-0.8137	0.2164	-3.76	0.0002
Edad 9	-0.6198	0.2289	-2.71	0.0068
Estado 2	0.4043	0.1128	3.58	0.0003

Como puede observarse, todas las variables son significativas con, al menos, un nivel de significancia $\alpha = 0.01$. Con esto, se puede concluir que este último es el mejor modelo que puede hacerse con los datos de la compañía.

2.3.6. Binomial negativa

Empíricamente, se ha observado de manera sistemática una varianza más grande que la media al hacer una estimación de la frecuencia. Por eso, además de la distribución Poisson, se hará un ajuste con la distribución Binomial Negativa y se compararán con base en los criterios de información definidos anteriormente.

Haciendo un símil con el modelo Poisson, se necesitan dos modelos: uno que incluya las tres variables explicativas (Edad, Sexo y Estado) y otro que sólo incluya el intercepto. Serán respetados los grupos de edades formados en el cuadro 2.8. Observe el modelo más complejo en el cuadro 2.11.

Cuadro 2.11. Regresión Binomial Negativa con edades ajustadas

	$\hat{\beta}$	Error estándar	Valor z	Pr(>|z|)
(Intercept0)	-2.1743	0.1481	-14.68	0.0000
Edad 2	-1.7221	0.5938	-2.90	0.0037
Edad 3	-0.8849	0.2994	-2.96	0.0031
Edad 4	-0.4695	0.1740	-2.70	0.0070
Edad 5	-0.4910	0.2008	-2.45	0.0145
Edad 6	-0.6021	0.2214	-2.72	0.0065
Edad 7	-0.5545	0.2331	-2.38	0.0174
Edad 8	-0.8160	0.2387	-3.42	0.0006
Edad 9	-0.6335	0.2516	-2.52	0.0118
Estado 2	0.3854	0.1258	3.06	0.0022
Sexo M	-0.1348	0.1162	-1.16	0.2461

De manera similar a la Poisson, se observa que aparentemente el Sexo no es una variable significativa para este modelo. Sin embargo, se repetirá el

procedimiento donde se parte del modelo más sencillo se irán añadiendo variables al modelo *(forward)*. Luego, se comenzará del modelo más complejo y se irán eliminando *(backward)*.

En este caso, al hacer los procedimientos mencionados, los métodos conducen a resultados diferentes. Por un lado, si se parte de un modelo que incluye únicamente al intercepto, el método arroja como variable significativa al Estado de residencia de la persona asegurada. Sin embargo, al partir del modelo más complejo e ir eliminando variables, el modelo elimina a la variable Sexo quedándose con Edad y Estado.

Cuando se procede a través de estos métodos y no son consistentes como en el caso de la Poisson, el criterio de selección de modelo es muy sencillo. De estos dos modelos propuestos, se elege el que tenga un AIC más pequeño.

Cuadro 2.12. AIC para los modelos finales

Método	AIC
Forward	204.3
Backward	201

Como se puede observar, el método *Backward* (aquel que comienza del modelo complejo y elimina variables una a una) tiene el menor AIC y, por lo tanto, es aquel que se preferirá.
Nuevamente, note que las variables del modelo son significativas para el mismo con al menos $\alpha = 0.05$. Este es un buen modelo Binomial Negativo con el que se podría estimar la frecuencia de la cartera de Alfa Seguros.

Cuadro 2.13. Mejor regresión Binomial Negativa según AIC

	$\hat{\beta}$	Error estándar	Valor z	Pr(>|z|)
(Intercept)	-2.2441	0.1452	-15.46	0.0000
Edad 2	-1.7187	0.5962	-2.88	0.0039
Edad 3	-0.8665	0.3061	-2.83	0.0046
Edad 4	-0.4479	0.1847	-2.43	0.0153
Edad 5	-0.4880	0.2108	-2.32	0.0206
Edad 6	-0.5998	0.2308	-2.60	0.0094
Edad 7	-0.5536	0.2421	-2.29	0.0222
Edad 8	-0.8187	0.2477	-3.31	0.0009
Edad 9	-0.6436	0.2609	-2.47	0.0136
Estado 2	0.3797	0.1305	2.91	0.0036

2.3.7. Poisson *vs.* Binomial Negativa

Como ya se mencionó, una vez seleccionados los mejores modelos tanto para la Distribución Poisson como la Binomial negativa, lo que queda es comparar ambos modelos y seleccionar el mejor modelo con los criterios anteriormente mencionados.

Es decir, lo que se busca es, a través de los criterios de información AIC, BIC y haciendo el cociente de verosimilitudes, un modelo.

Cuadro 2.14. AIC, BIC y ℓ para el modelo de frecuencia

Modelo	AIC	BIC	log-verosimilitud
Poisson	200.52	216.36	-90.26
Binomial Negativa	200.96	218.38	-89.48

Se puede observar que tanto el AIC como el BIC prefieren el modelo Poisson. Es decir, por un lado, el criterio de información que penaliza

el número de parámetros elige al modelo más simple. Por otro lado, el criterio que privilegia la cantidad de información también elige al modelo más simple.

La única prueba que hace falta realizar es el cociente de verosimilitudes. Esta prueba se puede usar debido a que, en el límite,la distribución Poisson es sólo un caso particular de la ditribución Binomial Negativa. Esta prueba solo es aplicables a modelos *anidados*.

$$H_0 : \text{El modelo Poisson es adecuado } \textit{vs.}$$

$$H_1 : \text{ El modelo Binomial negativo es adecuado}$$

Sea

$$\chi^2 = 2(\ell_{H_{Bin-Neg}} - \ell_{H_{Pois}}) = 2(-89.48 - (-90.26)) = 1.557$$

Por otro lado, sea $\alpha = 0.01 \Rightarrow \chi^2_{(1)(1-\alpha)} = \chi^2_{(1)(0.99)} = 6.634$
Comparando:

$$\chi^2 = 1.557 < \chi^2_{(1)(0.99)} = 6.634$$

Por lo tanto, se rechaza la hipótesis alternativa H_1 que suponía que la distribución Binomial negativa era el modelo más adecuado para describir los datos. Se conserva la hipótesis nula y se continua suponiendo que la frecuencia se distribuye Poisson.

2.4. Resultados

Una vez seleccionado el modelo, sería importante ver cómo se comporta el parámetro λ para cada una de las edades y los estados con los que se está trabajando.

Recuerde que si:

$$Y|\lambda \sim Pois(\lambda)$$

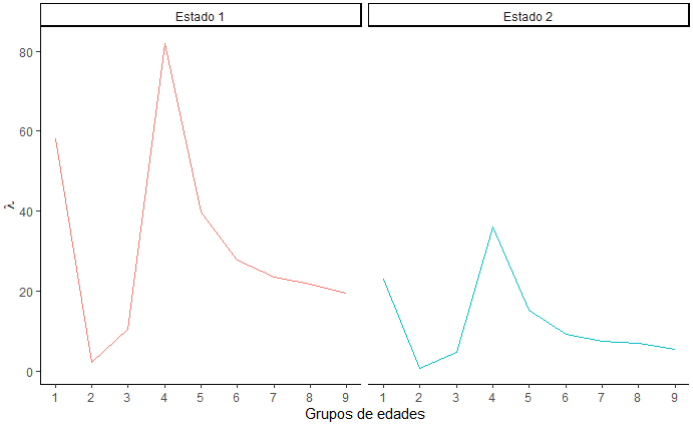
$$\Rightarrow E[Y|\lambda] = ln(X\beta) = \lambda$$

Sea $\hat{\beta}_0 = b_0, \hat{\beta}_1 = b_1, ...\hat{\beta_n} = b_n$

$$\hat{\lambda} = ln(X\hat{\beta}) = ln(b_0 + X_1 b_1 + ... + X_n b_n)$$

Este último será el estimador para la frecuencia. Además, recuerde que la función liga es la función $f(x) = ln(x)$. A continuación se podrá ver cómo fue el comportamiento del parámetro dada la muestra.

Figura 2.5. Parámetro λ para cada grupo de edad por estado



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Aparentemente, el parámetro λ de aquellos estados que contienen a las principales ciudades en México son más altas que en el resto del país. Sin embargo, lo que se debería hacer es dividir ese parámetro λ entre el número de asegurados que contiene el grupo para averiguar una medida más real.

Cuadro 2.15. Parámetro lambda de cada segmento de edad y estado

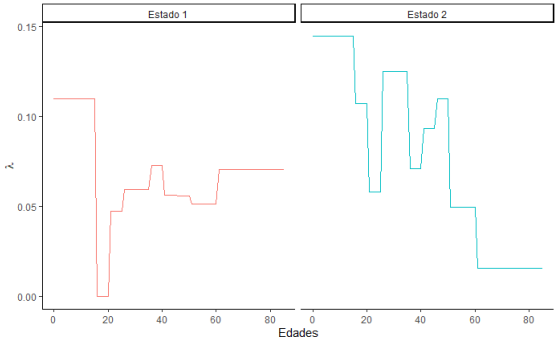
Edad	Estado	Reclamaciones	Asegurados	$\hat{\lambda}$	$\frac{\text{Reclamación}}{Asegurados}$
1	1	60	547	57.98	0.11
1	2	21	145	23.02	0.14
2	1	0	119	2.22	0.00
2	2	3	28	0.78	0.11
3	1	11	232	10.38	0.05
3	2	4	69	4.62	0.06
4	1	73	1227	81.97	0.06
4	2	45	360	36.03	0.12
5	1	44	604	39.73	0.07
5	2	11	155	15.27	0.07
6	1	27	482	27.77	0.06
6	2	10	107	9.23	0.09
7	1	22	394	23.63	0.06
7	2	9	82	7.37	0.11
8	1	24	466	21.89	0.05
8	2	5	101	7.11	0.05
9	1	24	341	19.45	0.07
9	2	1	65	5.55	0.02

Como se puede observar ahora, en general el cociente producido entre el número de reclamaciones y el número de asegurados es más grande para aquellos estados que no contienen a las ciudades más importantes en México. Lo anterior, deberá verse reflejado al momento de la tarificación. Esta característica parece ser una particular de la cartera

del Alfa seguros.

En la figura 2.6 se puede observar cómo se vería el parámetro λ para cada una de las edades. Note que en algunas partes de la gráfica permanece constante debido a que pertenecen al mismo grupo de edades.

Figura 2.6. Parámetro λ para cada edad por estado



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Capítulo 3

Análisis de severidad

3.1. Análisis exploratorio

Al igual que en el análisis de frecuencia, antes de empezar un análisis profundo de la severidad, se observará cómo se comportan los datos. Para la base de siniestros, se han hecho varios filtros para estudiar únicamente lo que interesa: siniestros reclamados por pólizas individuales de gastos médicos.

Además, para ser congruentes con el primer análisis exploratorio, se considerarán las reclamaciones hechas por las personas que tienen un determinado nivel de exposición al riesgo: 100 millones. La edad de las personas es calculada al 31 de diciembre de 2019 por ser esta última la fecha de corte con la que se están trabajando los datos.

En el mismo sentido de congruencia, las variables que se van a estudiar en el análisis exploratorio son la edad, estado de residencia de la persona y sexo. Los resultados de la frecuencia podrían ser independientes de los de la severidad. Por ejemplo: podría ser que el sexo de la persona resulte una variable que sea más determinante al

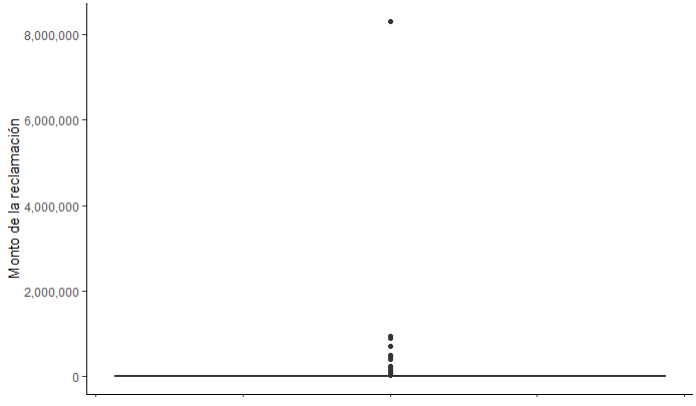
momento de estudiar la severidad.

Antes de continuar, vale la pena mencionar que, después de aplicar los filtros, queda una base de datos con un total de 690 reclamaciones.

3.1.1. Análisis general

Lo primero que resulta relevante de la base de datos ya filtrada es conocer, en términos muy generales, alrededor de dónde se encuentran los datos. Particularmente, su media y en qué intervalo (de monto) están la mayoría de las reclamaciones.

Figura 3.1. Caja de brazos de las reclamaciones

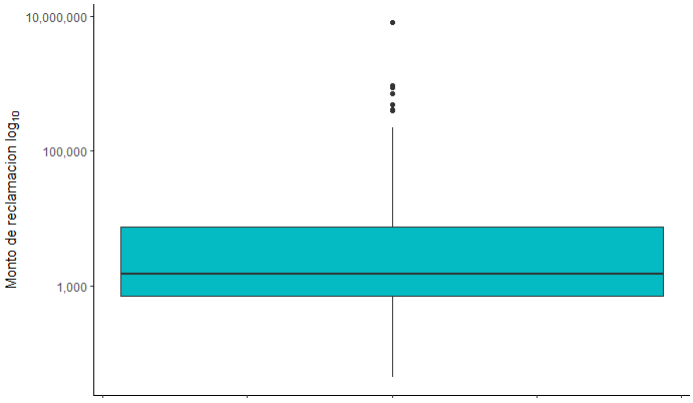


Fuente: Elaboración propia con datos de Alfa seguros (2019)

Aunque la figura 3.1. sea poco efectiva en términos visuales para lo que se buscaba, al menos indica que la mayoría de las observaciones son menores a 2 millones y que existe una enorme desviación provocada por una pérdida de 8 millones.

Para facilitar la visualización de la gráfica, lo que se hará será escalar el eje vertical con la función logaritmo.

Figura 3.2. Caja de brazos de las reclamaciones

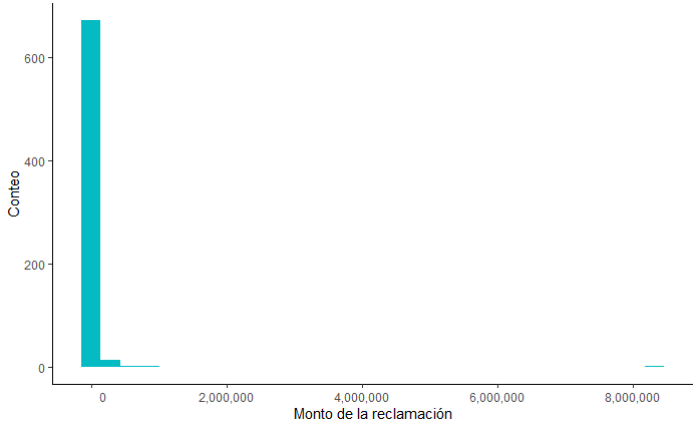


Fuente: Elaboración propia con datos de Alfa seguros (2019)

Debe tenerse mucho cuidado con la figura 3.2. porque aunque la mediana parezca estar muy cerca de mil, recuerde que está en escala logarítmica. Por lo tanto, esta mediana podría ser mucho mayor de lo que aparenta. Por otro lado, en esta gráfica es más complicado ver los *outliers*.

Además de los gráficos de caja y brazos, otra herramienta muy útil para obtener información sobre la distribución de un conjunto de datos es el gráfico de barras.

Figura 3.3. Gráfica de barras de las reclamaciones



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Es de esperarse en una cartera de Gastos médicos que tenga reclamaciones de colas muy pesadas. Como lo es este caso. Sin embargo, también es de esperarse que la mayoría estén alrededor del cero. Observe, numéricamente, cómo se comporta a grandes rasgos la cartera.

Cuadro 3.1. Principales características de las reclamaciones

Estadístico	Valor	Estadístico	Valor
Mínimo	44	Máximo	8,318,794
Primer cuantil	700	Media	30,758
Mediana	1,516	Rango intercuartílico	6,711
Tercer cuantil	7,411	Desviación estándar	323,434

Se ha confirmado que la mayoría de las reclamaciones son relativamente pequeñas. Sin embargo, note que la desviación estándar es gigante. Además, la mediana no es cercana a mil como lo parecía en

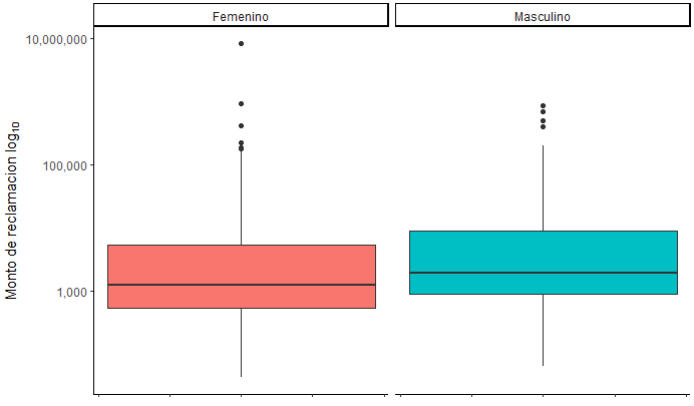
la figura 3.2.

A continuación, un análisis un poco más profundo.

3.1.2. Análisis por sexo

Para el análisis exploratorio de las reclamaciones por sexo y posteriores es importante hacer notar que se estarán escalando los ejes con la función logaritmo para que su visualización sea más sencilla. Para los siguientes casos no solo se estará escalando el gráfico de caja y brazos, sino también el gráfico de barras.

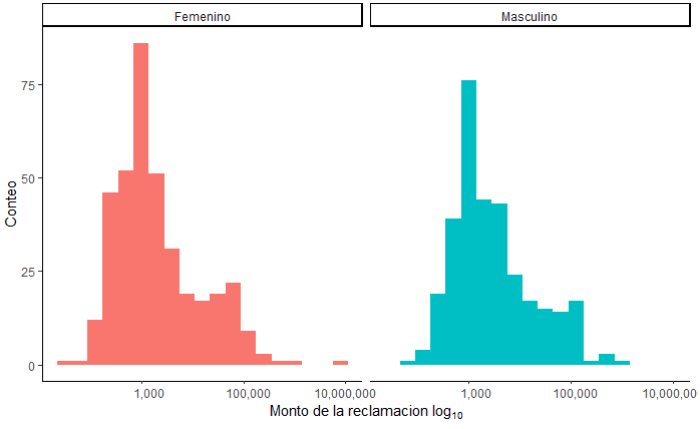
Figura 3.4. Gráfica de caja separado por sexo



Fuente: Elaboración propia con datos de Alfa seguros (2019)

En este caso, se puede observar que, en términos generales, no hay diferencias que parezcan significativas ni la mediana ni en el gran volumen de reclamaciones. Sin embargo, el caso deberá seguir siendo estudiado.

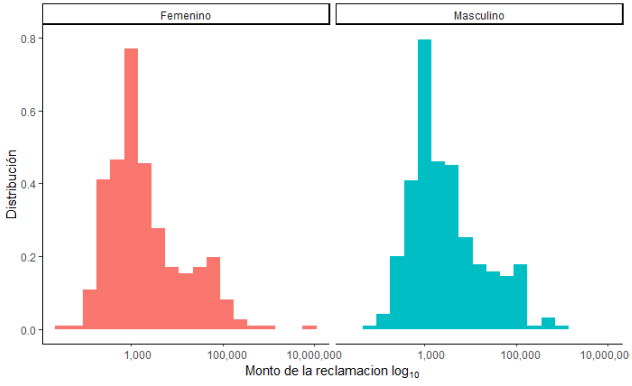
Figura 3.5. Diagrama de barras para monto de reclamaciones por sexo



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Al segregar por sexo, en la figura 3.5., que tiene en escala logarítmica el eje horizontal, parece no haber diferencias significativas. Sin embargo, antes de sacar conclusiones, no sólo hay que considerar el conteo de las reclamaciones, sino cómo se distribuyen.

Figura 3.6. Histograma del monto reclamaciones por sexo



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Aun escalando por distribución al eje vertical de la figura 3.6., no es perceptible alguna diferencia significativa. Por último, observe numéricamente cómo se comportan los datos.

Cuadro 3.2. Principales características de las reclamaciones

Sexo	Media	Desviación estándar
Masculino	15,960	60,573
Femenino	38,280	434,711
Femenino *	21,958	77,848

Para el cuadro 3.2., además de incluir la media y la desviación estándar de las reclamaciones segregadas por sexo, se ha decidido hacer un tercer renglón definido como *Femenino* *. Puede verse en la figura 3.1. que existe una reclamación de 8 millones que está muy desviada del resto de las reclamaciones. Debido a lo anterior, se ha decidido excluirla del cálculo y la desviación estándar del último renglón para ver cuánto está

afectando esa desviación.

Como puede verse, al quitar este siniestro, la media y la desviación estándar de las reclamaciones son más cercanas. Además, se cuenta con evidencia de que para la frecuencia, la variable sexo es una variable que no es significativa para el modelo. Por lo tanto, en el caso de la severidad no debería ser extraño que pase algo parecido.

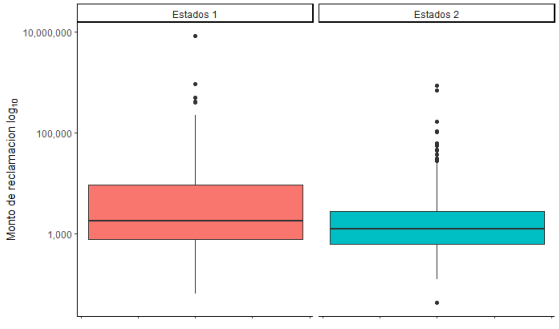
Hasta el momento, esta variable parece no ser significativa, pero el caso seguirá siendo estudiado.

3.1.3. Análisis por estado

La siguiente variable que se quiere estudiar es el estado de residencia de los asegurados. Para ser congruentes con el análisis de frecuencia, se seguirá haciendo la segregación de estos de la misma manera.

Igualmente, se continuará trabajando bajo una escala logarítmica y se presentarán las mismas gráficas.

Figura 3.7. Caja y brazos de reclamaciones por estado de residencia



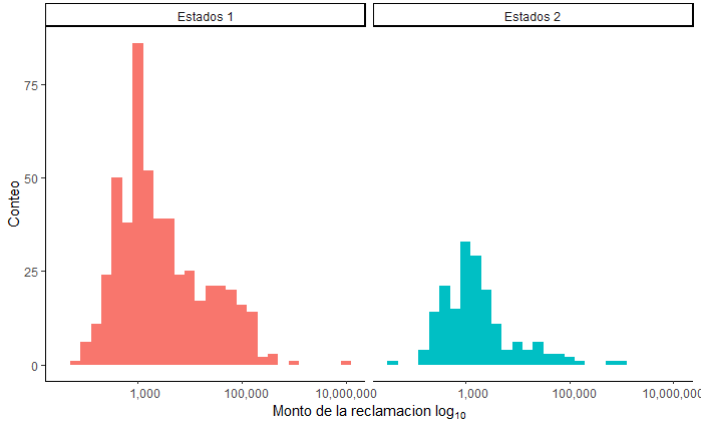
Fuente: Elaboración propia con datos de Alfa seguros (2019)

La primera impresión podría ser que existe entre las reclamaciones que

tienen las personas del grupo *Estados 1* y *Estados 2*. Se puede apreciar como el segundo gráfico está más comprimido que el primero en la figura 3.7.

Lo anterior, sugiere que la cola del primer grupo es más pesada que la segunda. Es posible decir que ese siniestro ha estado causando desviaciones grandes es del sexo femenino y reside en alguno de los estados que contiene una ciudad importante. Esa desviación puede ser causada por el costo de la atención médica.

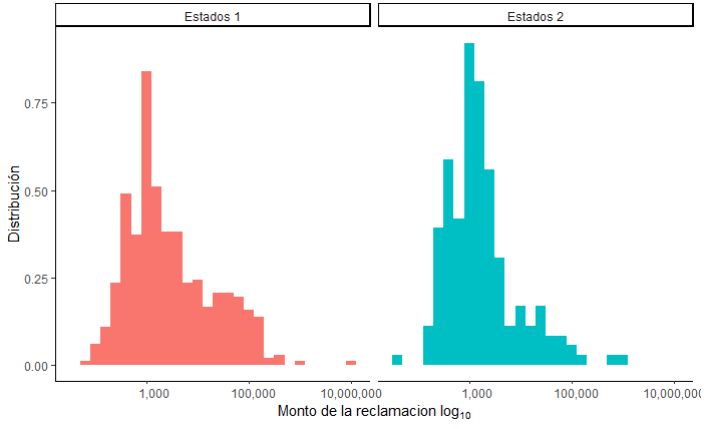
Figura 3.8. Diagrama de barras para los montos de reclamaciones por estado de residencia



Fuente: Elaboración propia con datos de Alfa seguros (2019)

En este caso, es notorio un volumen mayor de reclamaciones hechas por las personas del primer grupo que del segundo. Además, como ya se había previsto, la caída es mucho más lenta y se observan muchos más siniestros en la cola de la derecha del primer grupo, mientras que en el segundo caen ligeramente más rápido.

Figura 3.9. Histograma de monto de reclamaciones por estado de residencia



Fuente: Elaboración propia con datos de Alfa seguros (2019)

La figura 3.9. muestra la importancia de no sólo mostrar la cantidad de reclamaciones, sino su distribución. La distribución permite ver la abrupta caída antes de los 100 mil pesos de reclamaciones en el segundo grupo. Al mismo tiempo, se observa cómo las reclamaciones se mantienen *altas* para el primer grupo de asegurados.

Ahora se tiene evidencia para confirmar la hipótesis de colas más pesadas en el primer grupo. Es decir, aun cuando los resultados numéricos salgan desviados uno del otro, no lo es atribuible a un solo siniestro, como lo fue el caso del sexo.

Cuadro 3.3. Principales características de las reclamaciones

Estados	Media	Desviación estándar
Estado 1	36,078	372,363
Estado 2	15,570	85,679

Como puede observarse la media y la varianza de la categoría Estado 1 no están cerca de las mismas medidas de la otra categoría. Por lo que esta variable sí resultará significativa para el resto del análisis de severidad.

3.1.4. Análisis por edad

Para el análisis de severidad, así como para el de la frecuencia, si se quisiera segregar por edades individuales, no sólo se tendría poca información para algunas edades, sino que no habría información para muchas otras. Por lo anterior, se ha optado por segregar las edades en tres grandes grupos: edades tempranas, edades intermedias y edades avanzadas. Estos rangos fueron determinados por cuota, con la esperanza que cada uno contuviera una densidad adecuada de observaciones para poder calcular estadísticas de manera confiable y realizar inferencias robustas. Numéricamente, las edades quedaron catalogadas como sigue:

Cuadro 3.4. Agrupación de edades para la severidad

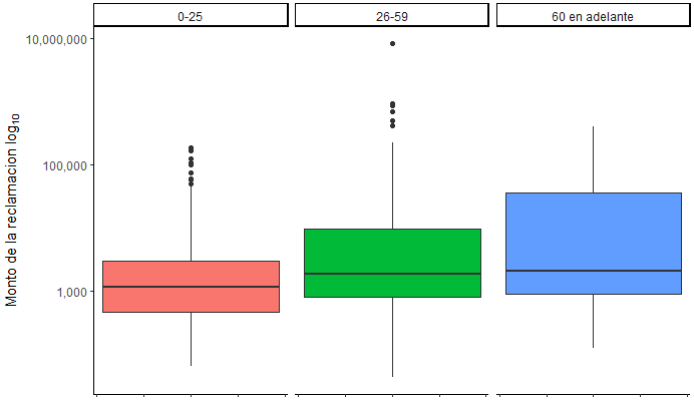
Edades agrupadas	Asegurados	Media	Desviación estándar
de 0 a 25	204	8,588	25,531
de 26 a 59	451	40,414	398,881
de 60 en adelante	35	35,547	77,266

Note que aun segmentando en grupos bastante amplios de edades, la información con la que se cuenta es relativamente reducida. Es decir, considere que será necesario que $n \to \infty$ para que los resultados cumplan con dos teoremas importantes: La ley de los grandes números y el Teorema del límite central.

Para el análisis por edad, no fueron mostrados los resultados

numéricos antes de las gráficas para mostrar que los tres grupos no son homogéneos y puede confirmarse gráficamente.

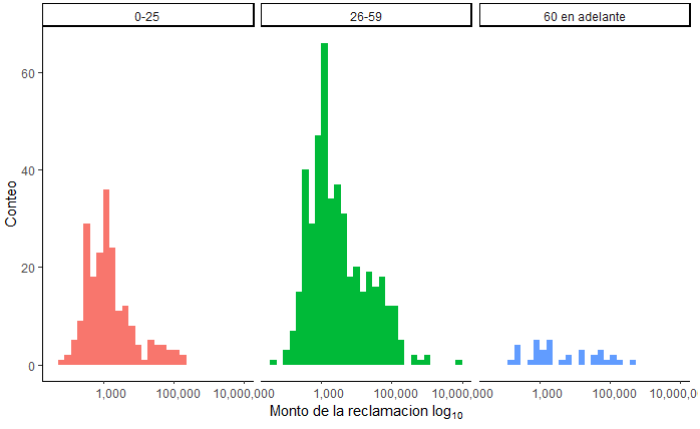
Figura 3.10. Caja y brazos de reclamaciones por grupo de edad



Fuente: Elaboración propia con datos de Alfa seguros (2019)

La primera gráfica de la figura 3.10. ayuda a confirmar algunos aspectos que podrían imaginarse desde antes: la mediana del segundo y tercer grupo son muy parecidas. Sin embargo, al tener un mayor número de asegurados, este segundo grupo tiene un rango intercuartílico menor y tiene una caja más *aplastada* que la del tercer grupo.

Figura 3.11. Diagrama de barras para monto de reclamaciones por grupo de edad

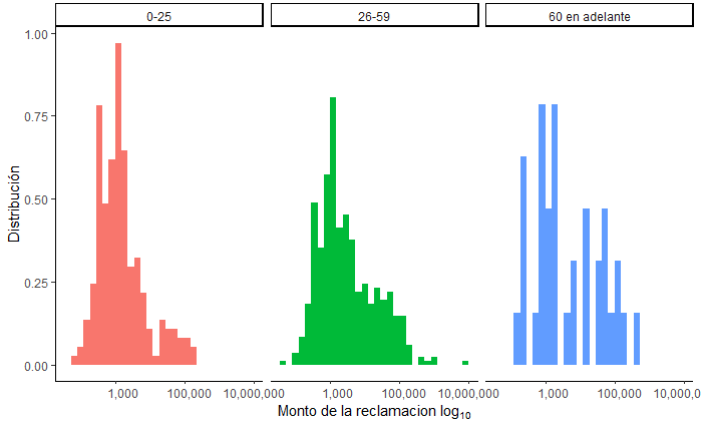


Fuente: Elaboración propia con datos de Alfa seguros (2019)

Los conteos de la figura 3.11. muestran dónde se encuentra la mayor parte de los asegurados: en el segundo grupo. Nuevamente, este resultado no debería ser sorpresivo debido a que ya se había visto que esto es fácilmente atribuible a la pirámide poblacional.

Sin embargo, lo que sí puede observarse es que la cola de ese segundo grupo es mucho más pesada que el resto. Esto provocará una prima un poco más grande para este grupo que para el resto.

Figura 3.12. Histograma de monto reclamaciones por grupo de edad



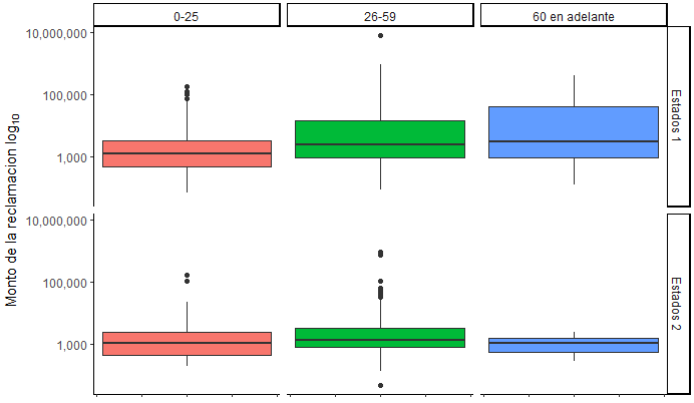
Fuente: Elaboración propia con datos de Alfa seguros (2019)

La figura 3.12. es una de las mejores aproximaciones que pudiera tenerse sobre el comportamiento de los datos. Por lo anterior, se puede decir que efectivamente la cola del segundo grupo es la más pesada de la segregación. Por otro lado, se observa una gran concentración de siniestros alrededor de los 100 mil en el primer grupo. Estas observaciones sobresalen incluso por encima de las observaciones del mismo monto del segundo grupo.

3.1.5. Análisis por edad y estado

La figura 3.13. muestra cómo se comportan las distribuciones al segregar la base en seis grandes grupos que resulten las posibles combinaciones entre edad y estado de residencia.

Figura 3.13. Caja y brazos de reclamaciones por grupo de edad y estado



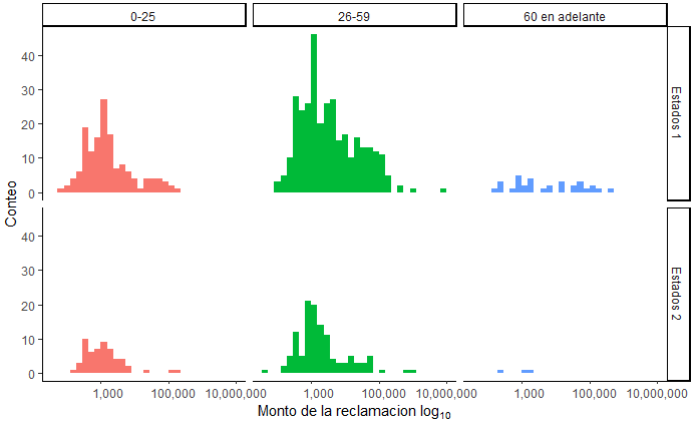
Fuente: Elaboración propia con datos de Alfa seguros (2019)

Como puede observarse, la siniestralidad de los asegurados del Estado 1 tienen un rango intercuartílico es un poco más amplio que los del segundo grupo. Al mismo tiempo, se pudo ver que no es una constante la amplitud de intervalo dependiendo la edad.

Se sigue observando un *outlier* que está muy desviado del resto en el grupo Estado 1 y entre las edades 26 y 59. Al mismo tiempo, el segundo dato más desviado se encuentra en el mismo rango de edades, pero el segundo grupo de residencias.

Al mismo tiempo, observe que la gráfica de los Estados 2 y de edades de 60 en adelante podría parecer un grupo con edades muy homogéneas. Sin embargo, esto puede explicarse porla escasez de datos.

Figura 3.14. Diagrama de barras de monto de reclamaciones por grupo de edad y estado

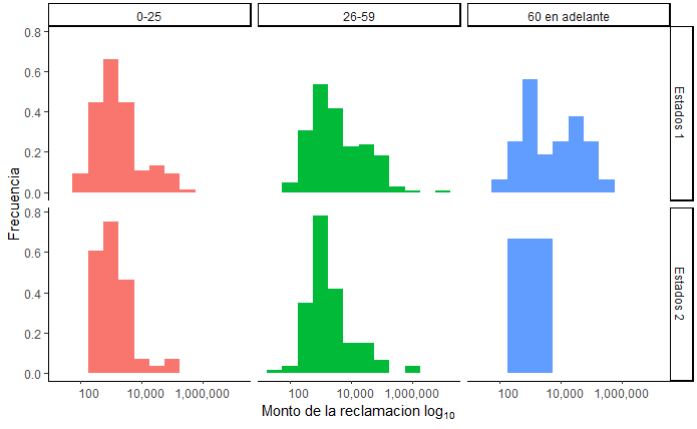


Fuente: Elaboración propia con datos de Alfa seguros (2019)

Note que al igual que en la frecuencia, se puede ver un mayor número de observaciones en el primer grupo de Estados. Al mismo tiempo, apenas se alcanzan a apreciar observaciones del tercer grupo de edades y el segundo de estado.

Sin embargo, esto no parece ser un rasgo particular del segundo grupo de estados. El primer grupo de estados tiene el mismo problema, pero no es tan notable como el primer grupo mencionado.

Figura 3.15. Histograma de monto de reclamaciones por grupo de edad



Fuente: Elaboración propia con datos de Alfa seguros (2019)

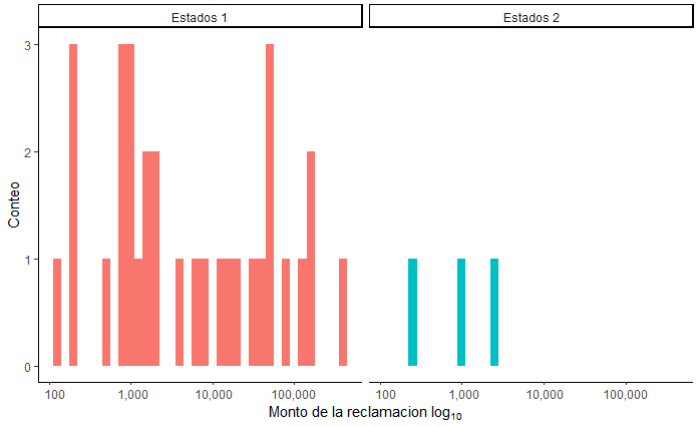
Se puede observar que escalar el eje vertical con frecuencia relativa ayuda a no ser alarmante la cantidad de datos con los que se cuenta. Lo anterior debido a que supone que, de tener más datos, las distribuciones tendrían, por lo menos, la misma forma.

Sin embargo, se sigue teniendo un problema: las observaciones del Estado 2 y edades mayores a 60 años son prácticamente nulas. Para intentar corregirlo, se agruparán las observaciones de ambos grupos de estados para esas Edades.

3.1.6. Agrupación de edades de 60 en adelante

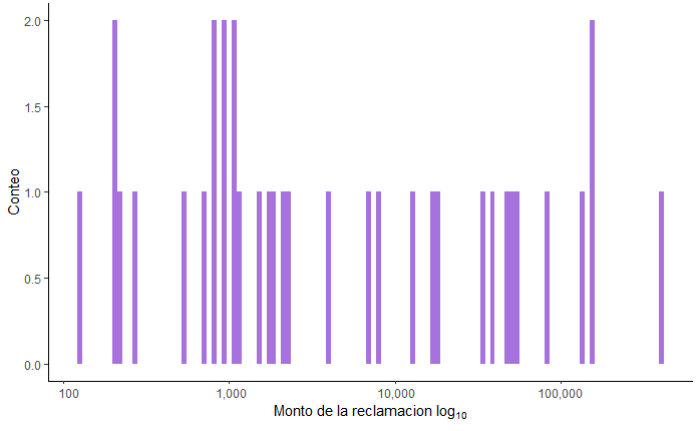
Para evaluar si es significativo agrupar los datos obtenidos del *Estados 1* y *Estados 2*, se observa su comportamiento conjunto y por separado.

Figura 3.16. Reclamaciones sin agrupar los estados



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Figura 3.17. Reclamaciones agrupando los estados



Fuente: Elaboración propia con datos de Alfa seguros (2019)

Al observar la figura 3.16. y 3.17., se pude observar que en ninguno de los casos es posible asociar la forma de las reclamaciones con otra distribución que no sea la uniforme. Por lo tanto, se podría considera que las observaciones de ambos estados provienen de una misma muestra hasta que haya evidencia de lo contrario.

3.2. Estadística aplicada a los seguros

La teoría de los seguros funciona en gran parte por considerar que la cartera cumple con dos principios: Teorema del límite central y Ley de los grandes números.

Teorema 1 (Ley de los grandes números)**.** *Sea $X_1, X_2, ..., X_n$ variables aleatorias independientes e idénticamente distribuidas con media μ desconocida. Se dice que la media aritmética converge en probabilidad a la media teórica (Ross, 2013).*

$$\overline{X}_n = \frac{X_1 + X_2 + ... + X_n}{n}$$

En otras palabras, sea $\epsilon > 0$

$$\lim_{n\to\infty} P(|\overline{X}_n - \mu| < \epsilon) = 1$$

Teorema 2 (Teorema de límite central)**.** *Sea $X_1, X_2, ..., X_n$ variables aleatorias independientes e idénticamente distribuidas con media μ y varianza σ^2 desconocidas. Se dice que la media aritmética sigue una distribución normal (Ross, 2013).*

$$\lim_{n\to\infty} \frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0, 1)$$

Note que tanto μ como σ^2 son parámetros desconocidos.

$$\text{Sea } Z = \frac{\mu - \overline{X}_n}{\frac{\sigma}{\sqrt{n}}} \text{ y } Q = \frac{(n-1)S_x^2}{\sigma^2}$$

$$\text{donde } S_x^2 = \sum_{i=1}^{n} \frac{(X_i - \overline{X})^2}{n-1} = \hat{\sigma}^2$$

En realidad se tiene que

$$T = \frac{Z}{\sqrt{\frac{Q}{n-1}}} = \frac{\frac{\overline{X}_n - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{1}{n-1}\left(\frac{(n-1)S_x^2}{\sigma^2}\right)}} = \frac{\overline{X}_n - \mu}{\sqrt{n}S_x} \sim t_{(n-1)}$$

De lo anterior, se puede obtener que

$$P\left[\overline{X}_n - t_{(n-1)\left(\frac{1-\alpha}{2}\right)}\frac{S_x}{\sqrt{n}} \leq \mu \leq \overline{X}_n + t_{(n-1)\left(\frac{1-\alpha}{2}\right)}\frac{S_x}{\sqrt{n}}\right] = 1 - \alpha$$

Una vez obtenidas la media y la varianza muestral (fijas), se define $\alpha = 0.01$ y se obtiene el intervalo de confianza fijo para μ al nivel $(1-\alpha) \times 100\,\%$ como sigue

$$\left[\overline{X}_n - t_{(n-1)\left(\frac{1-\alpha}{2}\right)}\frac{S_x}{\sqrt{n}} \ , \ \overline{X}_n + t_{(n-1)\left(\frac{1-\alpha}{2}\right)}\frac{S_x}{\sqrt{n}}\right]$$

Cuadro 3.5. Intervalos de confianza al 99 % para μ

Grupo	n	Límite inferior	Límite superior
Edad 1 y Estado 1	148	3,865	14,664
Edad 1 y Estado 2	56	-2,525	16,125
Edad 2 y Estado 1	331	-17,805	113,420
Edad 2 y Estado 2	120	-4,474	44,520
Edad 3	35	31,158	37,682

Hasta ahora, se han usado el teorema del límite central para construir los intervalos de confianza de los diferentes grupos junto con un supuesto de normalidad en las muestras. Al mismo tiempo, note del cuadro 3.5. que hay estimaciones negativas en el intervalo de la media. La estimación negativa no es derivada de observaciones negativas, sino de una inmensa variabilidad en los datos.

Al observar el último intervalo, podría parecer el más razonable por estar en un espacio más acotado. Sin embargo, por el momento, una hipótesis más razonable es que, al haber una cantidad de datos casi nula (35), no existen desviaciones significativas con una muestra de un tamaño tan pequeño.

Por lo tanto, se puede decir que no existe evidencia de que los datos cumplan con la ley de los grandes números para cada grupo por separado por haber pocos datos y la tarificación de este seguro es imposible mediante estadística clásica.

Sin embargo, existe otro enfoque estadístico que será útil para tarificar un seguro de gastos médicos mayores: la Estadística bajo enfoque Bayesiano.

3.3. Análisis de distribuciones

A lo largo de esta sección se tendrá interés particular por evaluar la razonabilidad de la elección de un modelo que describa apropiadamente a las reclamaciones de la cartera de Alfa Seguros.

3.3.1. Verosimilitud

Con base en la sección anterior, se observa un comportamiento de campana en escala logarítmica. Lo anterior, al revertir la escala a la original, produce distribuciones de cola pesada. Al estar estudiando

reclamaciones, debe tenerse presente que éstas son estrictamente positivas. Por lo tanto, la elección estará restringida a aquellas distribuciones cuyo soporte sea estrictamente positivo y de colas pesadas.

Para este caso, se usarán las distribuciones: Gamma Inversa, Log-normal, T de Student, Cauchy, Pareto y F de Snedecor. Siendo todas distribuciones de colas pesadas y aunque las distribuciones T de Student y Cauchy no son estrictamente positivas, es posible parametrizar las distribuciones para que la probabilidad de valores negativos sea cercana a 0. Haciendo ajustes paramétricos por máxima verosimilitud, la log verosimilitud resulta como sigue.

Cuadro 3.6. Log verosimilitud de familias paramétricas

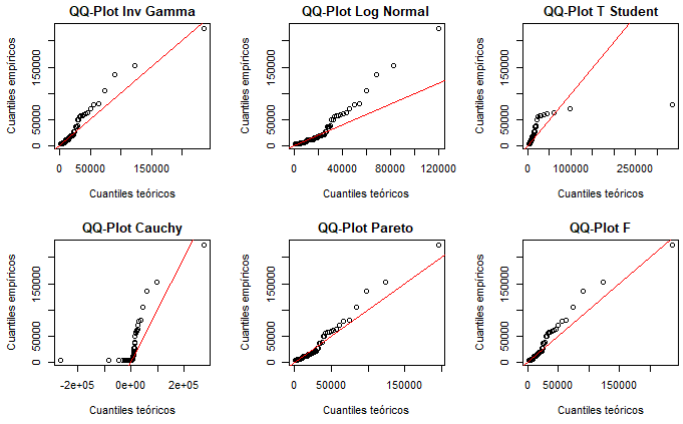
Familia	Edad 1 Estado 1	Edad 1 Estado 2	Edad 2 Estado 1	Edad 2 Estado 2	Edad 3
Inv Gamma	-1534	-578	-3745	-1294	-409
Log Norm	-1552	-582	-3765	-1308	-413
T no central	-1497	-576	-3713	-1278	-400
Cauchy	-1581	-580	-3837	-1328	-417
Pareto	-1573	-594	-3792	-1324	-420
F no central	-1534	-578	-3745	-1294	-409

Resultados redondeados sin decimales

Observe que, a excepción de la F, la cual tiene tres parámetros, el resto de distribuciones tienen únicamente dos parámetros. Lo anterior se debe a que según lo que se ha mencionado en el cuadro 2.4. el AIC penaliza dependiendo de la cantidad de parámetos y el BIC por la cantidad de datos utilizados. Al usar la misma cantidad de datos sería irrelevante usar criterios de selección de modelos como el AIC o BIC. Sin embargo, la log verosimilitud sí explica cuál es un mejor ajuste.

Antes de hacer la elección de un modelo, se graficarán los QQ-Plots de las familias paramétricas estudiadas. Las gráficas serán hechas para el grupo *Edad 1 - Estado 1* y el resto se dejarán en anexos por la similitud presentada.

Figura 3.18. QQ-Plot de familias paramétricas de la distribución de severidad



Ahora, además del resultado numérico de la log verosimilitud, se puede ver gráficamente el ajuste de las familias paramétricas. Lo anterior ayuda a descartar a la distribución Cauchy. Se observa que la T de Student también hace una mala inferencia sobre el comportamiento de la distribución. Las familias que parecen ser aceptables son la Gamma Inversa, Pareto, y Lognormal. La distribución F fue descartada por la cantidad de parámetros y no tener una mejora significativa en verosimilitud.

Más adelante en la investigación será elegida alguna de estas tres distribuciones. Por ahora, observe que la única de las 3 que sobrestima la observación más grande es la Inversa Gamma. Lo anterior significa que es

la de colas más pesadas. La segunda y tercer distribución calificándolas por las cosas son Pareto y Lognormal. Siendo esta última la de colas más ligeras.

3.3.2. Sistema Estadístico del Sector Asegurador (SESA)

No se ha hablado mucho acerca de esta base de datos en particular. Lo anterior es debido a que la SESA no forma parte de la información de la compañía. Esta es más bien una recompilación de siniestros de todas las compañías de seguros.

Se tiene interés por conocer el comportamiento de los parámetros de las distribuciones de la *Subsección 3.3.1.* El procedimiento para cada uno de los grupos será exactamente el mismo. Debido a la similitud en el procedimiento, se observará a detalle el del primer grupo y se dejará en los anexos el resto de los grupos.

Primero, de la SESA se filtrarán los datos como correspondan según el grupo estudiado, se ajustarán por máxima verosimilitud parámetros que correspondan para las tres familias paramétricas del apartado anterior: Inversa Gamma, Pareto y Lognormal.

Se tomarán muestras de tamaño 1,000 de la SESA. A esa muestra se le harán el ajuste por máxima verosimilitud para hallar los parámetros. Este proceso de muestreo se repetirá 1,000 veces.

Un escenario ideal sería que los estimadores de máxima verosimilitud de los parámetros estuvieran lo menos correlacionados posibles para que la distribución conjunta sea únicamente el producto de sus marginales al suponer independencia. Para este estudio se graficará cada pareja de estimadores en un plano cartesiano.

Figura 3.19. Parámetros estimados $\hat{\alpha}$ y $\hat{\beta}$ de una distribución Inversa Gamma

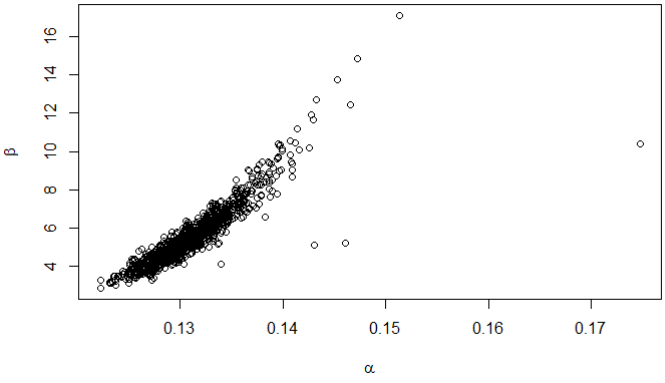


Figura 3.20. Parámetros estimados $\hat{\alpha}$ y $\hat{\theta}$ de una distribución Pareto

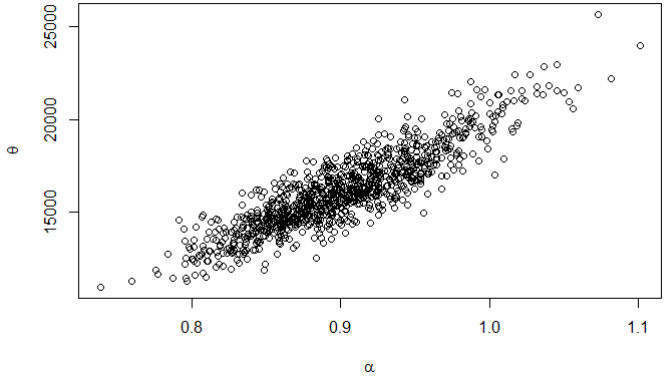
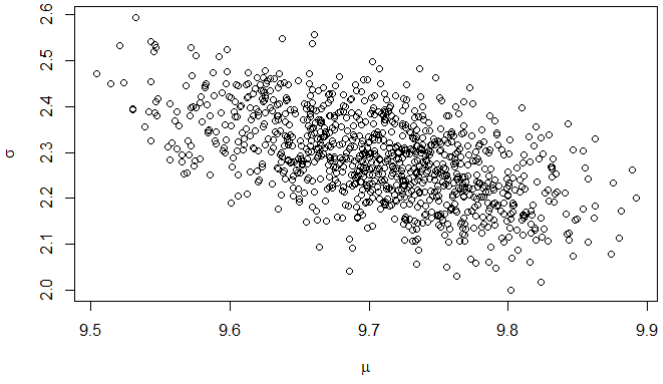


Figura 3.21. Parámetros estimados $\hat{\mu}$ y $\hat{\sigma}$ de una distribución Log Normal



Clásicamente, se pueden usar tres medidas de dependencia para ver su relación: Correlación de Pearson, Correlación de Spearman y τ de Kendall.

Cuadro 3.7. Medidas de dependendia para los parámetros de las distintas familias paramétricas

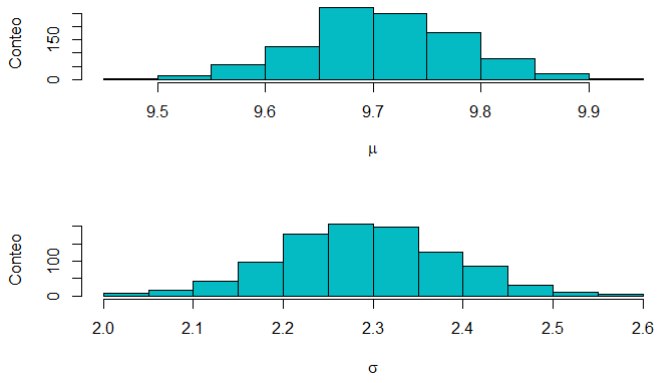
Medida de dependencia	Inversa Gamma	Pareto	Lognormal
Pearson	0.904	0.893	-0.549
Spearman	0.940	0.697	-0.534
Kendall	0.795	0.879	-0.374

Como se puede observar, los coeficientes de correlación son más altos para la Inversa Gamma que para la Pareto y la Lognormal. Siendo esta última la de menor grado en cuanto a medidas de dependencia. Aun cuando la correlación no es un criterio de optimalidad, el hecho de tener una baja correlación ayudará en un futuro. Por lo tanto, los parámetros

estimados de la Lognormal son los mejores para este propósito.

Antes de continuar, observe el comportamiento marginal de los parámetros estimados de la Lognormal.

Figura 3.22. Gráfica de barras de los parámetros estimados $\hat{\mu}$ y $\hat{\sigma}$ de una Lognormal



La figura 3.21. y el cuadro 3.7. evidencian que a pesar de que la dependencia es débil, no es nula. Por lo tanto en este punto de la investigación parece razonable hacer uso de cópulas. En términos simples, una cópula (denotada por C) es una función que une (*couple*) funciones de densidad multivariadas a partir de sus marginales univariadas. Similarmente, las cópulas pueden ser definidas como funciones de distribución multivariadas cuyas marginales univariadas son uniformes en el espacio (0,1) (Nelsen, 2007).

Formalmente,

Definición 6. *C es una función tal que $C : [0,1]^p \to [0,1]$, donde $p > 1$ y que cumple con las siguientes propiedades (Nelsen, 2007).*

1. $C(1,1,..,U_j,1,...,1) = U_j, j = 1,2,...,p$
2. Si $U_i < V_i \ \ \forall i \in \{1,2,...p\} \Rightarrow C(U) \leq C(V)$
3. Si $0 \leq a_i \leq U_i \leq b_i \leq 1 \ \ \forall i \in \{1,2,...p\} \Rightarrow C([\mathbf{a},\mathbf{b}]) \geq 0$

Como se puede observar hasta el momento, no se han mencionado las variables aleatorias en la definición formal. Sin embargo, las cópulas son especialmente útiles cuando se quiere unir dos funciones marginales en una función de densidad conjunta. Existe un teorema que explica la relación que tienen las cópulas con la teoría Probabílistica: Teorema de Sklar

Teorema 3 (Teorema de Sklar.)**.** *Sean F y G dos funciones de distribución que han sido escaladas al intervalo [0,1] y H la función de distribución bivariada. Si, además, F y G son funciones continuas, existe una única función C tal que:*

$$H(x,y) = C(F(x),F(y))$$

Más generalmente, sea $X \in \mathbb{R}^p \ \exists \ C$ *tal que:*

$$F(x_1,x_2,...,x_p) = C(F(x_1),F(x_2),...,F(x_p))$$

Del teorema anterior se deduce que

$$C(x_1,x_2,...,x_p) = F(F^{-1}(x_1),F^{-1}(x_2)...,F^{-1}(x_p))$$

Se ha encontrado que dentro de las cópulas existen familias. Estas familias son parecidas a las familias paramétricas conocidas en Probabilidad. Sin embargo, antes de seguir avanzando, es necesario proponer distribuciones marginales para μ y σ.

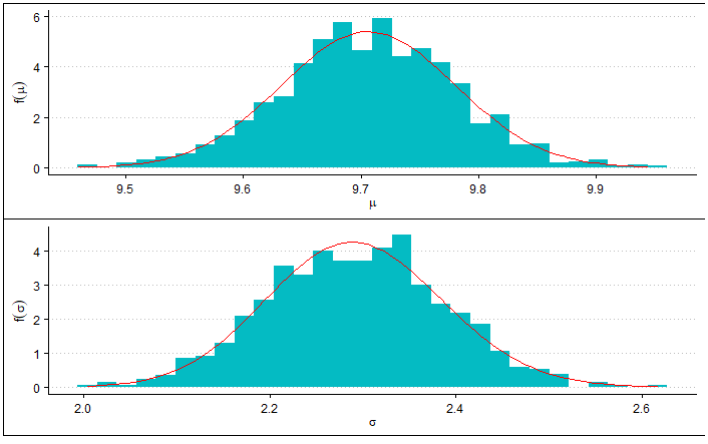
Para ello, de los vectores μ y σ, se propone una distribución Gamma y nuevamente se le aplica una prueba Kolmogorov-Smirnov. En ninguno

de los casos, se rechaza la hipótesis de que los parámetros μ y σ se distribuyen Gamma.

Observe que a μ se le asigna una distribución con soporte positivo a pesar de que, en general, $\mu \in \mathbb{R}$. Lo anterior se debe a que, después de realizar el análisis exploratorio, parece poco razonable que la media de la distribución sea pequeña. Como una media pequeña en una Lognormal se obtiene al hacer μ cercano a cero o negativo, parece que un parámetro μ estrictamente positivo, hace más sentido para el problema que se busca resolver.

Gráficamente, se observa cómo sigue.

Figura 3.23. Ajuste de distribución Gamma para μ y σ



Con ayuda de la paquetería *VineCopula*[1] de R se determinará cuál es la familia de Cópulas que se debe usar para relacionar las marginales Gammas. Para lo anterior, recuerde que es importante escalar las variables aleatorias al rectángulo $[0, 1] \times [0, 1]$.

[1] https://cran.r-project.org/web/packages/VineCopula/VineCopula.pdf

La función *BiCopSelect* indica que la Cópula Gaussiana es la mejor alternativa para modelar los distribuciones marginales de manera conjunta. La función hace el ajuste mediante máxima verisimilitud y compara mediante el AIC, aunque podría hacerlo mediante el BIC o cualquier otra medida especificada.

Definición 7 (Copula Gaussiana)**.**

$$C(u,v) = \int_{-\infty}^{\Phi^{-1}(u)} \int_{-\infty}^{\Phi^{-1}(v)} \frac{1}{\sqrt{2\pi(1-\rho^2)}} e^{-\frac{s^2-2\rho st+t^2}{2(1-\rho^2)}} dsdt$$

Donde Φ es la función de distribución de una normal.

Asimismo, recuerde que u y v deberán ser las funciones de distribución de una Gamma.

Es decir, sabiendo que $\mu \sim Gamma(\eta_\mu, \zeta_\mu)$ y $\sigma \sim Gamma(\eta_\sigma, \zeta_\sigma)$, se deduce que:

$$F(\mu) = \frac{\gamma\left(\eta_\mu, \zeta_\mu \mu\right)}{\Gamma(\eta_\mu)}$$

Donde:

$$\gamma\left(\eta_\mu, \zeta_\mu \mu\right) = \int_0^{\zeta_\mu \mu} t^{\eta_\mu - 1} e^{-t} dt$$

$$\Gamma\left(\eta_\mu\right) = \int_0^{\infty} t^{\eta_\mu - 1} e^{-t} dt$$

La definición de $F(\sigma)$ es análoga.

$$\therefore F(\mu,\sigma) = C(F(\mu), F(\sigma)) =$$

$$\int_{-\infty}^{\Phi^{-1}(F(\mu))} \int_{-\infty}^{\Phi^{-1}(F(\sigma))} \frac{1}{\sqrt{2\pi(1-\rho^2)}} e^{-\frac{s^2-2\rho st+t^2}{2(1-\rho^2)}} dsdt$$

Finalmente,

$$f(\mu,\sigma) = \frac{\partial^2 F(\mu,\sigma)}{\partial\mu\partial\sigma} = \frac{\partial^2 F(\mu,\sigma)}{\partial\sigma\partial\mu}$$

Si bien no se conoce ni se va a desarrollar la forma analítica de la función de densidad ni distribución conjunta, numéricamente es posible calcular su densidad. La cual será útil más adelante.

Los parámetros de las distribuciones marginales para μ y σ fueron estimados por máxima verosimilitud y se ven como sigue.

Cuadro 3.8. Parámetros de distribuciones marginales de μ y σ

Grupo	η_μ	ζ_μ	η_σ	η_σ
Edad 1 Estado 1	19,301	1,988	548	239
Edad 1 Estado 2	25,053	2,597	570	285
Edad 2 Estado 1	23,195	2,278	607	277
Edad 2 Estado 2	28,167	2,832	598	303
Edad 3	23,671	2,291	976	458

Resultados redondeados sin decimales

En el cuadro 3.9. se podrá observar la familia seleccionada y los parámetros estimados.

Cuadro 3.9. Familias paramétricas de Cópulas ajustadas

Grupo	Cópula	Parámetro
Edad 1 Estado 1	Cópula Gaussiana	$\rho = -0.532$
Edad 1 Estado 2	Cópula Gaussiana	$\rho = -0.441$
Edad 2 Estado 1	Cópula Gaussiana	$\rho = -0.482$
Edad 2 Estado 2	Cópula Gaussiana	$\rho = -0.473$
Edad 3	Cópula Gaussiana	$\rho = -0.532$

Resultados redondeados a tres decimales

Si bien la paquetería de R sugiere que es mejor para el grupo **Edad**

2 Estado 2 una cópula BB1 [2] rotada 270°, analizando la propuesta de R es posible cambiarla a una Cópula Gaussiana por simplicidad del ejercicio y porque el impacto en verosimilitud es mínimo.

3.4. Estadística Bayesiana

La Estadística Bayesiana es una teoría axiomática de Inferencia Estadística. El darle una estructura axiomática a la Estadística, no sólo resulta atractiva desde un punto de vista matemático, sino que asegura que los resultados y procedimientos se desprenden de dichos axiomas que, además, no se contradicen ni son incompatibles.

En este momento, para la investigación, es importante destacar que la principal ventaja de la Estadística Bayesiana sobre la Estadística Frecuentista es que complementa la información de los datos incorporando información previa (o adicional) al modelo.

La información adicional se incorpora a partir de distribuciones de probabilidad llamadas *a priori* y se obtienen a partir de bases de datos externas, opiniones expertas o creencias de aquella persona que estudia un fenómeno incierto.

Al usar información que no sólo proviene de la muestra aleatoria, la Estadística Bayesiana producirá inferencias más estables y ayudará a quitar la alta variabilidad provocada por la poca cantidad de datos.

Contrario a lo que se pudiera pensar hasta el momento, la información adicional con la que se cuenta no es sobre los datos, sino

[2] Una Cópula BB1 es una Cópula bivariada que generaliza las Cópulas univariadas Gumbel y Clayton. En particular, si en una Cópula BB1 el parámtero a tiende a cero, entonces se estaría en presencia de una Cópula Gumbel. Por otro lado, si el parámetro c es igual a 1, entonces se obtendría una Cópula Cayton. El hecho de rotarla 270 grados es para empatar mejor con las colas de la distribución bivariada.

sobre el parámetro (o parámetros) que se quieren estudiar. En ese sentido, los datos serán parte de la función de verosimilitud cuyo parámetro $\boldsymbol{\theta}$ es desconocido, pero su distribución es conocida a priori.

La forma de corregir la información a priori con la función de verosimilitud es a través del Teorema de Bayes. Este último producirá la distribución posterior.

Es decir: sea $f(\boldsymbol{\theta})$ una distribución a priori y $f(X|\boldsymbol{\theta})$ una función de verosimilitud cuyos datos provienen de una muestra aleatoria,

$$f(\boldsymbol{\theta}|X) = \frac{f(\boldsymbol{\theta})f(X|\boldsymbol{\theta})}{C}$$

donde C es conocida como una constante de proporcionalidad y se define como sigue:

$$C = \int_{\boldsymbol{\Theta}} f(\boldsymbol{X}|\boldsymbol{\theta})f(\boldsymbol{\theta})d\boldsymbol{\theta}$$

Se puede definir análogamente para $\boldsymbol{\theta}$ discreta.
Más aún, una vez obtenida la distribución posterior del parámetro, se pueden hacer inferencias sobre el comportamiento de las siguientes observaciones condicionadas en la observación de la muestra. A esto se le conoce como distribución predictiva posterior. Se puede demostrar que la distribución predictiva posterior de la siguiente observación se obtiene de la ecuación:

$$f(X_{n+1}|\boldsymbol{X}_n) = \int_{\boldsymbol{\Theta}} f(X_{n+1}|\boldsymbol{\theta})f(\boldsymbol{\theta}|\boldsymbol{X}_n)d\boldsymbol{\theta}$$

Por último, la teoría de credibilidad es una herramienta que ocupan los actuarios para tarificar seguros. Esta metodología, al ocupar Estadística Bayesiana, permite al actuario reconocer la experiencia del mercado o

cualquier otro agente externo y mezclarla con las observaciones de su propia cartera.

Para este estudio particular esto último resultará muy útil debido a que la cartera no cuenta con un volumen de datos suficiente.

3.5. Estimación Bayesiana

Antes de empezar con el cálculo de la severidad, debe tenerse presente que el objetivo es encontrar la esperanza de la distribución predictiva posterior:

$$E[X_{n+1}|\boldsymbol{X}_n] = \int_{\boldsymbol{\Theta}} E(X_{n+1}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{X}_n) d\boldsymbol{\theta}$$

Para lo anterior, debe primero definirse una distribución a priori, una función de verosimilitud y encontrar la distribución posterior.

Antes de continuar, vale la pena proponer una distribución para el monto de las reclamaciones. Aunque la cantidad de datos que se tienen resulta poca en términos de convergencia de estimadores, es posible asignar una distribución a los datos.

A lo largo de la **Sección 3.3** fueron exploradas algunas de las distribuciones que pudieran ser adecuadas para modelar una distribución cuyo soporte es estrictamente positivo y de colas pesadas.

Entre las distribuciones que resultaron adecuadas fueron: Inversa Gamma, Pareto y Lognormal. En esta investigación, se usará aquella distribución que sea computacionalmente más efectiva al momento de producir resultados con menor correlación y que cumpla con ciertos criterios que serán definidos en las siguientes secciones.

En este caso, la elección de una distribución Log Normal resulta ser una distribución adecuada para modelar apropiadamente las reclamaciones de la cartera de Alfa Seguros.

3.5.1. Distribución a priori

Una de las mayores ventajas de la Estadística Bayesiana puede ser también el punto más criticable de la investigación. La Estadística bajo enfoque Bayesiano es comúnmente criticada por la definición de una creencia inicial sobre el parámetro desconocido.

Las críticas provienen de que el investigador podría establecer una distribución a priori que arroje los resultados que le convengan. Por lo anterior, la definición de esta distribución estará basada en estadísticas del mercado: SESA en materia de Salud.

El comportamiento de los parámetros fue estudiado en la **Sección 3.3.2** cuyas distribuciones marginales para los parámetros μ y σ son distribuciones Gammas.

Asimismo, la distribución conjunta probablemente no tenga una forma analítica conocida. Sin embargo, se puede estimar la densidad para los propósitos que sean definidos más adelante.

3.5.2. Función de verosimilitud

Los datos que forman parte de la función de verosimilitud, fueron estudiados previamente a lo largo de este capítulo: desde la forma de segmentarlo, hasta proponer una distribución que nos parece razonable que sigan los datos. Sin embargo, aún se debe observar que los datos de la base son datos de reclamaciones; no de siniestros. Por lo tanto, en esta sección será definida correctamente la función de densidad de los siniestros.

Sean Y las reclamaciones, X los siniestros, d el deducible asociado al siniestro, cc el coaseguro y u la suma asegurada de la póliza:

$$f_y(y) = \begin{cases} \frac{f_x(x)}{1-F_x(d)} & si \quad d \leq x \leq \frac{u}{1-cc} + d \\ \\ \frac{1-F_x(u)}{1-F_x(d)} & si \qquad x > \frac{u}{1-cc} + d \end{cases}$$

donde

$$Y = \begin{cases} 0 & si & x < d \\ \\ (x-d)(1-cc) & si & d \leq x \leq \frac{u}{1-cc} + d \\ \\ u & si & x > \frac{u}{1-cc} + d \end{cases}$$

Como se puede observar, un siniestro está definido como la reclamación dividida por 1 menos el porcentaje de coaseguro más el deducible del contrato y además se divide entre la función de supervivencia del deducible porque, para observar una reclamación, el siniestro debe rebasar el deducible.

Por otro lado, es posible, aunque en el caso particular de Alfa Seguros es complicado por el volumen tan alto de sumas aseguradas, observar que el siniestro esté topado en lo que se obliga a pagar la aseguradora: la suma asegurada. Para contemplar que el siniestro es más grande que la suma asegurada, se debe poner una función de supervivencia a partir de dicho monto y, de nuevo, una función de supervivencia en el deducible para poder observar esa reclamación.

Al no conocer el parámetro $\boldsymbol{\theta}$, no es posible decir mucho hasta el momento. Sin embargo, recuerde que la distribución es totalmente conocida.

3.5.3. Distribución posterior

Hasta el momento, se conocen las distribuciones a priori y de los siniestros.

La distribución a priori no tiene forma analítica conocida, pero es el resultado de incluir marginales gammas en una Cópula Gaussiana. Por otro lado,

$$\boldsymbol{X}|\boldsymbol{\theta} \sim Lognormal(\mu, \sigma)$$

Ahora, la inferencia bajo enfoque Bayesiano se complica al momento de calcular la distribución posterior, ya que, aunque su forma es, en principio, fácil de obtener mediante el Teorema de Bayes, la densidad a priori no se ha podido expresar de forma analítica. Si bien, multiplicar formas proporcionales la distribución a priori por verosimilitud es fácil, encontrar una forma analítica para la posterior no sería tan fácil.

Es decir, la distribución posterior podría no tener una forma analítica conocida. Es por eso que se han desarrollado métodos computacionales que ayudan a estimar la distribución posterior.

Por lo tanto, la idea será aproximar numéricamente lo que analíticamente es complicado de expresar. Es decir, la distribución posterior. Uno de los métodos más populares es como *Markov Chain Monte Carlo* (MCMC) y lo que propone es aproximar la distribución posterior por medio de simulación.

En la literatura existen tres formas clásicas de hacer muestreo aleatorio mediante MCMC: Gibbs, Metropolis y Metropolis-Hastings, siendo este último el caso más general y los otros dos casos particulares del mismo (Kruschke, 2015).

En esta tesis se usará un muestreador de Metropolis-Hastings. Este último está fundamentado en varios artículos científicos: Metropolis *et*

al. (1953) y Hastings (1970). Muchos autores también consideran a Barker (1965) y Peskun (1973) como autores que contribuyen al método al proponer una nueva regla de aceptación rechazo y realizando análisis sobre la eficiencia de la regla. El artículo original de Metropolis trata de solucionar cálculos de propiedades químicas de una sustancia y fue publicado en *Journal of Chemical Physics*.

Metropolis-Hastings, en términos muy generales, funciona a través de Cadenas de Markov cuyos estados iniciales son arbitrarios. En cada uno de los pasos de la cadena, el siguiente paso dependerá únicamente de su estado actual.

El objetivo de usar cadenas de Markov es converger a una distribución límite de la cadena. Para lo anterior, la cadena debe cumplir con ser irreducible (no tener estados absorbentes), aperiódica (que el máximo común divisor del número de pasos necesario para volver a un estado sea 1) y que sea recurrente positiva (que el número de pasos promedio que necesite la cadena para volver a un estado sea menor a infinito).

Para demostrar que una cadena de Markov converge a una distribución límite es relativamente sencillo cuando la matriz de transición es finita. Sin embargo, en este caso, se tendrá una matriz de dimensiones infinitas y se mostrará tanto gráfica como numéricamente que la cadena alcanza una distribución límite.

Hace algunos párrafos se mencionaba que es irrelevante el estado inicial de la cadena. Esto se debe a que la cadena pasará por un proceso de adaptación conocido como *Burn-in*. Las observaciones producidas en el periodo de adaptación (o calentamiento) de la cadena serán ignoradas por no pertenecer al proceso estable.

El proceso de *burn-in* se hace con la intención de ignorar aquellas observaciones que no pertenezcan a la distribución límite a la que se

pretende llegar. Una vez pasado el proceso, se hará inferencia con las observaciones que ya pertenecen a la distribución límite.

Para este método hay que hacer algunas definiciones. Recuerde:

$$\pi(\boldsymbol{\theta}) = f(\boldsymbol{\theta}|\boldsymbol{X}_n) = \frac{f(\boldsymbol{X}_n|\boldsymbol{\theta})f(\boldsymbol{\theta})}{C}$$

Ahora bien, el algoritmo de Metropolis-Hastings recurre a una función q con la intención de explorar la distribución posterior $\pi(\boldsymbol{\theta})$

Además, Metropolis-Hastings es un algoritmo que ocupa la idea de aceptación-rechazo. Es decir, una $\boldsymbol{\theta}$ propuesta será aceptada si es mejor que la anterior. Si se añade un índice j las observaciones, $\boldsymbol{\theta}^{j-1}$ es la observación anterior y $\boldsymbol{\theta}^*$ es la $\boldsymbol{\theta}$ propuesta que se comparará con la anterior para ver si se acepta o se rechaza.

La forma de comparar dichas observaciones con la finalidad de aceptar o rechazar a través de un cociente. De nuevo, note que la constante de proporcionalidad es irrelevante para el cálculo de la distribución posterior, pues:

$$\frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{j-1})}{\pi(\boldsymbol{\theta}^{j-1})q(\boldsymbol{\theta}^{j-1}|\boldsymbol{\theta}^*)} = \frac{f(\boldsymbol{\theta}^*|\boldsymbol{X}_n)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{j-1})}{f(\boldsymbol{\theta}^{j-1}|\boldsymbol{X}_n)q(\boldsymbol{\theta}^{j-1}|\boldsymbol{\theta}^*)}$$

$$= \frac{\frac{f(\boldsymbol{X}_n|\boldsymbol{\theta}^*)f(\boldsymbol{\theta}^*)}{C}q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{j-1})}{\frac{f(\boldsymbol{X}_n|\boldsymbol{\theta}^{j-1})f(\boldsymbol{\theta}^{j-1})}{C}q(\boldsymbol{\theta}^{j-1}|\boldsymbol{\theta}^*)} = \frac{f(\boldsymbol{X}_n|\boldsymbol{\theta}^*)f(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{j-1})}{f(\boldsymbol{X}_n|\boldsymbol{\theta}^{j-1})f(\boldsymbol{\theta}^{j-1})q(\boldsymbol{\theta}^{j-1}|\boldsymbol{\theta}^*)}$$

Más aún, defina la regla de aceptación/rechazo (Gelman *et al.*, 2013) $\alpha(\boldsymbol{\theta}^{j-1}, \boldsymbol{\theta}^*)$

$$\alpha(\boldsymbol{\theta}^{j-1}, \boldsymbol{\theta}^*) = min\left\{1, \frac{\pi(\boldsymbol{\theta}^*)q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{j-1})}{\pi(\boldsymbol{\theta}^{j-1})q(\boldsymbol{\theta}^{j-1}|\boldsymbol{\theta}^*)}\right\}$$

Además, se define $\boldsymbol{\theta}^*$ como sigue como una caminata aleatoria. Es decir, $\boldsymbol{\theta}^{(*)} = \boldsymbol{\theta}^{(j-1)} + \boldsymbol{w}_j$ donde $\boldsymbol{w}_j$ es una variable aleatoria independiente de la cadena.

Para este caso, se define a esa variable aleatoria como una Normal Bivariada centrada en la observación aceptada inmediata anterior y una matriz de varianzas y covarianzas cuyas varianzas sean relativamente bajas y covarianzas 0.

$$\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{j-1} \sim N_2(\boldsymbol{\mu} = \boldsymbol{\theta}^{(j-1)}, \boldsymbol{\Sigma})$$

A continuación, se muestran un resumen del método de muestreo Metrópolis-Hastings:

1. Defina arbitrariamente la observación inicial $\boldsymbol{\theta}^{(1)}$. En este caso, se usará la distribución a priori para obtener esta observación, pero podría ser cualquiera.
2. Simule $\boldsymbol{\theta}^{*}|\boldsymbol{\theta}^{j-1}$
3. Calcule $\alpha(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{*})$
   - Si $\alpha(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{*}) = 1$ se acepta la observación.
   - Si $\alpha(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{*}) < 1$ simule $u \sim Unif(0, 1)$
     - Si $u > \alpha(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{*})$ se rechaza la observación.
     - Si $u < \alpha(\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{*})$ se acepta la observación.
4. Repita el proceso un número suficiente grande de veces. En este caso, se realizaron 130,000 observaciones.

Antes de continuar, es necesario hacer notar una particularidad del cociente de aceptación rechazo. En particular, observe que si

$$\boldsymbol{\theta}^{*}|\boldsymbol{\theta}^{j-1} \sim N_2(\boldsymbol{\theta}^{j-1}, \boldsymbol{\Sigma}) \text{ y } \boldsymbol{\theta}^{j-1}|\boldsymbol{\theta}^{*} \sim N_2(\boldsymbol{\theta}^{*}, \boldsymbol{\Sigma})$$

donde

$$q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{j-1}) \propto e^{-\frac{(\boldsymbol{\theta}^*-\boldsymbol{\theta}^{j-1})^T\boldsymbol{\Sigma}^{-1}((\boldsymbol{\theta}^*-\boldsymbol{\theta}^{j-1})}{2}}$$

$$q(\boldsymbol{\theta}^{j-1}|\boldsymbol{\theta}^*) \propto e^{-\frac{(\boldsymbol{\theta}^{j-1}-\boldsymbol{\theta}^*)^T\boldsymbol{\Sigma}^{-1}((\boldsymbol{\theta}^{j-1}-\boldsymbol{\theta}^*)}{2}}$$

$$\Rightarrow \frac{q(\boldsymbol{\theta}^*|\boldsymbol{\theta}^{j-1})}{q(\boldsymbol{\theta}^{j-1}|\boldsymbol{\theta}^*)} \propto 1$$

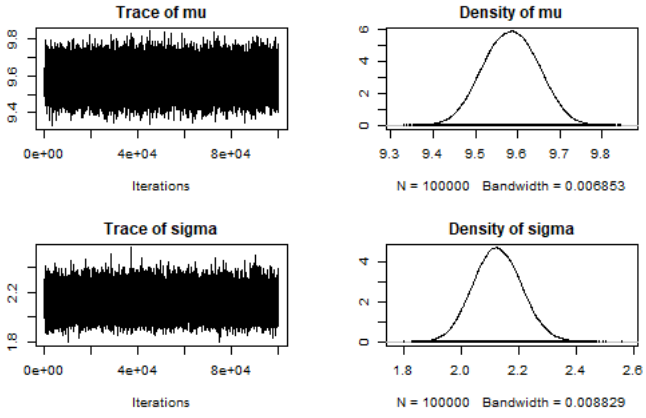
De hecho, en este caso, como ambas distribuciones tienen la misma matriz de varianzas y covarianzas, resulta que el cociente es exactamente igual a 1. Es decir, al tener el mismo kernel, se puede concluir que la distribución es la misma.

Por lo tanto, no es necesario calcular las densidades de esta función para el cociente. La caminata aleatoria será útil al momento de proponer el siguiente paso, pero no para ver si se acepta o se rechaza dicha observación.

En la figura 3.24. se muestran gráficas de densidad y *traceplots*[3] para las distribuciones posteriores de los parámetros μ y σ. Para este ejercicio, se consideró como periodo de calentamiento a las primeras 30,000 observaciones.

[3]Un *traceplot* es un gráfico que que muestra los estados visitados por la cadena (eje vertical) a lo largo de las iteraciones (eje horizontal).

Figura 3.24. Densidad y *Traceplot* de μ y σ



3.5.4. Convergencia de la distribución posterior

Como ya se había mencionado, el objetivo es obtener una muestra de la distribución posterior. Podría decirse que el muestreo es correcto si la muestra cumple que con que sea **estable, *representativa* y eficiente** (Kruschke, 2016). Estable en las medidas de tendencia central (valor esperado) y eficiente en el tamaño.

Recuerde que no es posible asegurar que una muestra es representativa en el sentido de que caracteriza a la perfección a la distribución posterior. Lo anterior debido a que si se pudiera caracterizar a la perfección la distribución posterior, no serían necesarios los métodos numéricos MCMC. Por lo tanto, la palabra *representativa* se refiere a que la cadena converge a su distribución límite independientemente del punto inicial de la misma. Este concepto es abordado por otros autores (Ross, 2013) y es definido como *ergodicidad*.

A continuación, se mostrarán métodos gráficos y numéricos que

demuestran que se cumplen las características propuestas.

1. **Estable**

La forma más sencilla en la que se puede mostrar que las media es estable para el proceso es una gráfica. Para lo anterior se calculará la media acumulada del proceso y se graficará a lo largo del millón de observaciones.

Figura 3.25. Media acumulada para μ

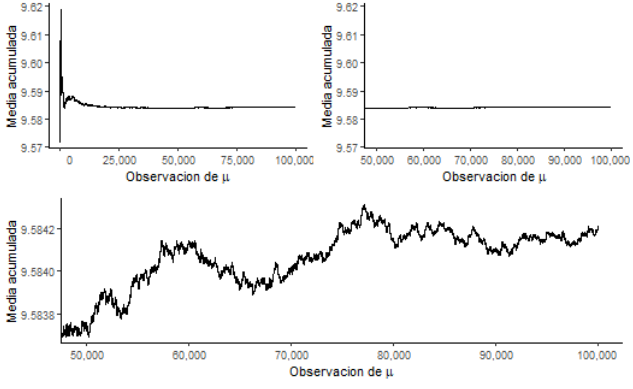
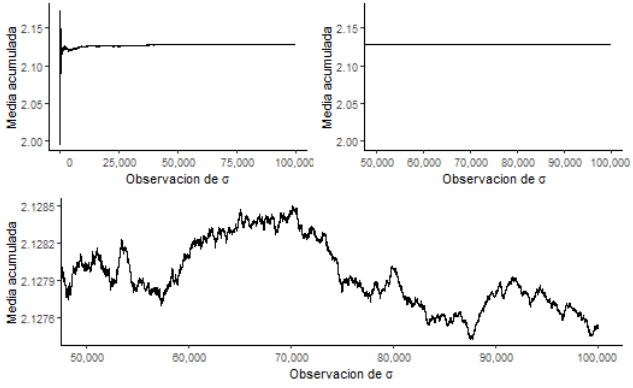


Figura 3.26. Media acumulada para σ



Cada una de las gráficas anteriores se dividen en tres partes. En la esquina superior izquierda se puede observar toda la gráfica de la media acumulada para ambos parámetros. Por otro lado, la gráfica de la esquina superior derecha es la misma media acumulada haciendo énfasis en las últimas 50 mil observaciones. Por último, la gráfica de la parte inferior hace un zoom sobre el eje vertical para observar mejor las variaciones de la segunda gráfica.

Es decir, una vez observado el comportamiento completo de la media acumulada, se hace un corte en la observación 50 mil debido a que a partir de entonces se observa un proceso estable. Para confirmar la hipótesis anterior, se hace un zoom a dicho proceso estable y se observa alrededor de dónde están esas observaciones.

Como se puede observar en ambas gráficas, las variaciones son mínimas en los últimos pasos de la cadena, por lo que gráficamente se puede decir que este es un proceso estable.

2 ***Representativa (ergódica)***

Como ya se había mencionado, lo que se procurará demostrar en este

punto es que la cadena converge prácticamente al mismo punto sin importar el estado inicial de la cadena.

Para lo anterior, se simularán cinco cadenas con estados iniciales aleatorios simulados a partir de la distribución a priori. Asimismo, no se considerarán los periodos de calentamiento de la cadena, ya que se quieren observar los comienzos de la cadena y las variaciones. Se simularán cien mil observaciones para el siguiente ejercicio.

Figura 3.27. Estados iniciales de μ

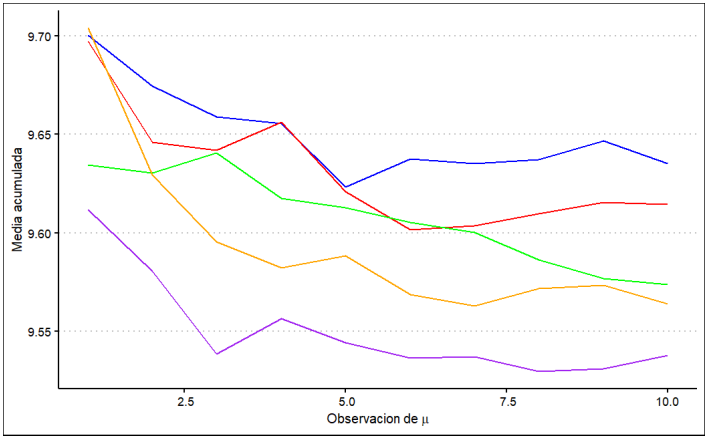
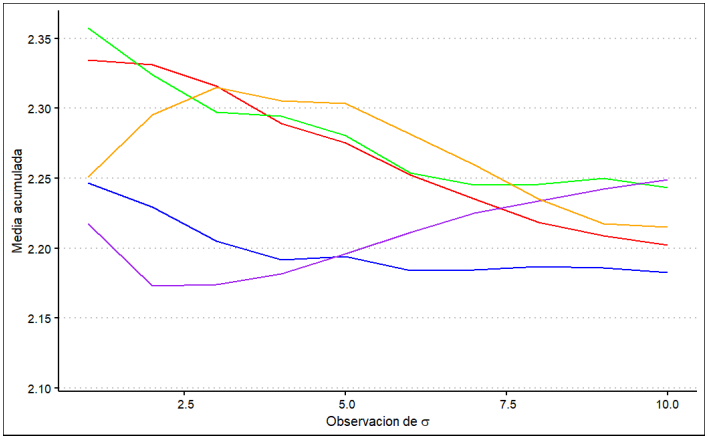


Figura 3.28. Estados iniciales de σ



En las figuras 3.27. y 3.28. se muestran los estados iniciales de μ y σ. Aun en estas gráficas se puede observar la correlación negativa que guardan los parámetros. Por ejemplo, observe como, a medida que crece la línea morada de σ, μ decrece.

Figura 3.29. Medias acumuladas para varias cadenas de μ

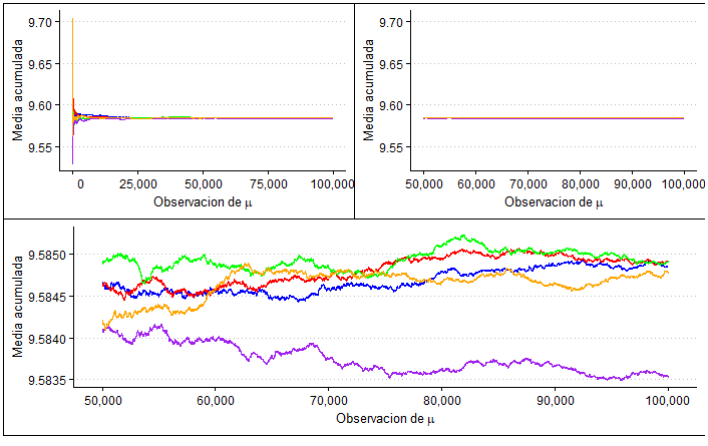
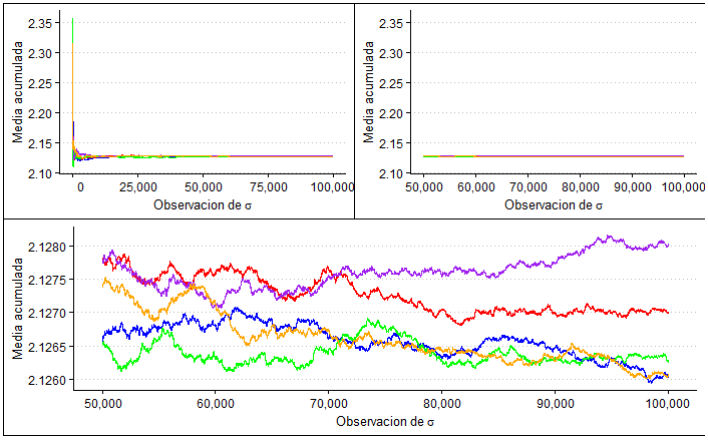


Figura 3.30. Medias acumuladas para varias cadenas de σ



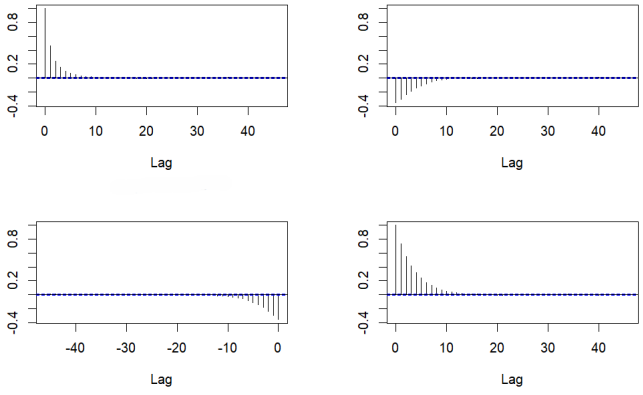
Note que el acomodo de las gráficas es igual que cuando se explora una sola cadena. Más interesante aún, note cómo sin importar el estado inicial, las cadenas oscilan su promedio en cantidades similares e incluso intercambian de lugar. Por lo tanto, se puede decir que la muestra es efectivamente *representativa*.

3 Eficiente

Se requiere una cadena que sea efectiva en términos de tamaño. En las figuras 3.29. y 3.30. se observa cierta estabilidad a partir de la observación 50 mil, pero entonces ¿por qué se hizo una muestra de cien mil observaciones?

La respuesta es que la cadena, al tener una mediana correlación en los parámetros, genera autocorrelación con observaciones anteriores y son pocos los pasos independientes que da la cadena. Lo anterior se puede observar con la correlación y que, a pesar de la función muestra hasta 10 rezagos (*lags*), esta la autocorrelación no disminuye.

Figura 3.31. Función de autocorrelación de μ y σ



Asimismo, existe la medida conocida como tamaño de muestra efectivo (*effective sample size* o ESS) que intenta explicar cuántas observaciones de las cadenas son realmente independientes y se define como sigue.

$$ESS = \frac{n}{1 + 2\sum_{i=1}^{\infty} \rho_i}$$

Donde n es el tamaño de muestra y ρ es la autocorrelación.

$$\rho_i = \frac{Cov(X_t, X_{t+i})}{Var(X_t)}$$

Aún con cien mil observaciones el ESS es de 30,477 para μ y 14,364 para σ lo que produce un tamaño de muestra no tan eficiente.

Antes de alarmarse, recuerde que esta medida cuenta solo las observaciones independientes en la cadena. Esto sería útil si se tuviera interés particular por los límites de los intervalos de máxima credibilidad. Sin embargo, anteriormente se ha mencionado que son más importantes las medidas de tendencia central que la apertura de los límites. Por lo anterior, y dado que no existe tanta variabilidad en

las medidas de tendencia central, la cadena es poco efectiva, pero no nula.

En la literatura (Kruschke, 2016), se dice que el tamaño del *effective sample size* no necesita ser tan grande si el interés son las medidas de tendencia central. Sin embargo, si bien de interés particular encontrar la media, la cola derecha de la distribución sí es relevante para el propósito de la investigación. Por lo tanto, para tener un resultado razonable de los intervalos de máxima credibilidad, el *ESS* es de al rededor de 10 mil. Tamaño con el cual cumplen las simulaciones.

Cuadro 3.10. Algunas de métricas para todos los grupos

Mediada	Edad 1 Estado 1	Edad 1 Estado 2	Edad 2 Estado 1	Edad 2 Estado 2	Edad 3
$Media_{\mu}$	9.594	9.601	9.936	9.859	10.300
$Media_{\sigma}$	2.128	1.921	1.841	1.856	2.085
SD_{μ}	0.064	0.061	0.058	0.057	0.069
SD_{σ}	0.083	0.079	0.066	0.072	0.067
L.I. HDI_{μ}	9.46	9.48	9.83	9.74	10.16
L.S HDI_{μ}	9.71	9.71	10.05	9.97	10.43
L.I. HDI_{σ}	1.96	1.76	1.71	1.71	1.93
L.S HDI_{σ}	2.29	2.07	1.97	1.99	2.21
ESS_{μ}	30,447	33,098	40,167	36,000	21,687
ESS_{σ}	14,364	16,097	21,473	18,709	20,755
Cociente	0.448	0.612	0.403	0.401	0.449

En el cuadro 3.10. se pueden observar la media y desviaciones para cada uno de los grupos observados. Sin embargo, además de algunas de las medidas de tendencia central, también se muestran los límites superiores e inferiores de los invertalos de máxima credibilidad con probabilidad

de 0.95 (L.S. HDI y L.I. HDI respectivamente).

Además, se deja como referencia el *ESS*. Como se mencionó, esta medida indica la cantidad de pasos en la cadena que son estrictamente independientes.

Por otro lado, otra forma con la que se mide la eficiencia de la simulación es el cociente de aceptación del algoritmo de muestreo Metrópolis-Hastings. Este cociente no es otra cosa que el tamaño de la cadena entre el número de observaciones necesarias para llegar a ese tamaño de la cadena.

Autores como Kruschke menciona que para problemas de dimensión baja como éste, el cociente de aceptación debe ser de aproximadamente 0.5 y es más flexible en la medida en la que la dimensión del problema aumenta. Como se puede observar en el cuadro 3.10., en este caso todas se encuentran alrededor del valor clásicamente aceptado.

Por las razones anteriores, podemos concluir que las cadenas que se construyeron son eficientes en tamaño.

3.5.5. Distribución predictiva posterior

Una distribución predictiva posterior se refiere a la función de densidad de X_{n+1} dado que ya se ha observado la muestra $X_1, X_2, ..., X_n$. Si además se cuenta con información adicional, la distribución predictiva posterior se puede escribir como sigue:

$$f(X_{n+1}|\boldsymbol{X}_n) = \int_{\boldsymbol{\Theta}} f(X_{n+1}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|\boldsymbol{X}_n) d\boldsymbol{\theta}$$

Sin embargo, observe que ha sido imposible caracterizar la distribución posterior debido a que la forma analítica de la distribución posterior no es conocida. Por lo tanto, la forma en la que se puede encontrar esta distribución predictiva posterior es usando muestras aleatorias de

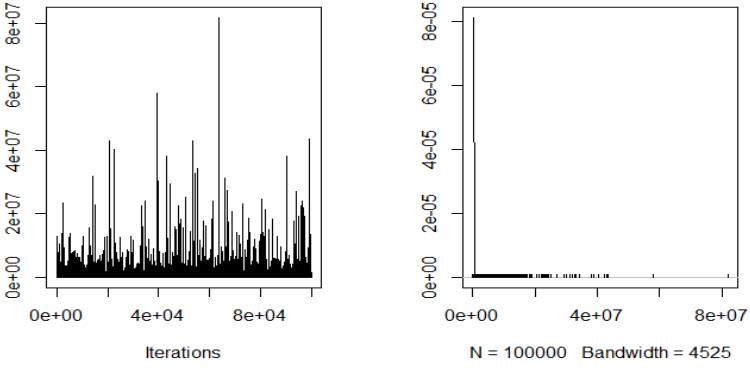
los parámetros de la distribución posterior y simulando la distribución propuesta para la verosimilitud: la distribución Lognormal.

Es decir, numéricamente lo que se hará es tomar una pareja de μ y σ. Esta pareja son observaciones de la distribución posterior encontrada anteriormente. Una vez elegida esta pareja de manera aleatoria, se simulará una observación de una Lognormal cuyos parámetros serán los anteriormente elegidos de manera aleatoria.

Lo anterior se puede entender como otra cadena de Markov. Por lo tanto, se usarán las mismas pruebas de convergencia que para la distribución posterior. Para esta simulación, se realizó utilizando un periodo de calentamiento de 30 mil observaciones y una cadena con un total de 100 mil observaciones.

Observe, para empezar, el *traceplot* y la función de densidad.

Figura 3.32. *Traceplot* y densidad de distribución predictiva posterior

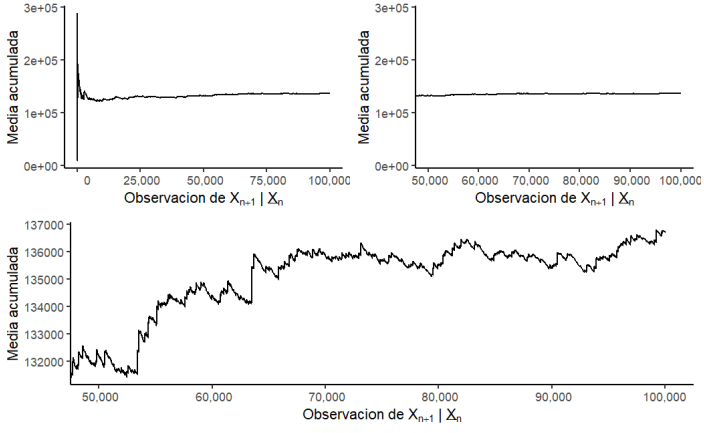


En la figura 3.32. se observa un *traceplot* diferente a los que se observan en la figura 3.24. Lo anterior se debe a que los siniestros no pueden ser menores a cero. Al mismo tiempo, se observa una función de densidad

muy densa en siniestros de montos bajos, pero de colas muy pesadas.

En la figura 3.33. se observarán las gráficas para mostrar la estabilidad de la cadena. Recuerde que la medida que se ocupa para evidenciar la estabilidad es la media acumulada.

Figura 3.33. Media acumulada distribución predictiva posterior



Una vez demostrada la estabilidad, debe probarse que, sin importar el estado inicial de la cadena, las medidas deben converger al mismo punto.

Figura 3.34. Estados iniciales distribución predictiva posterior

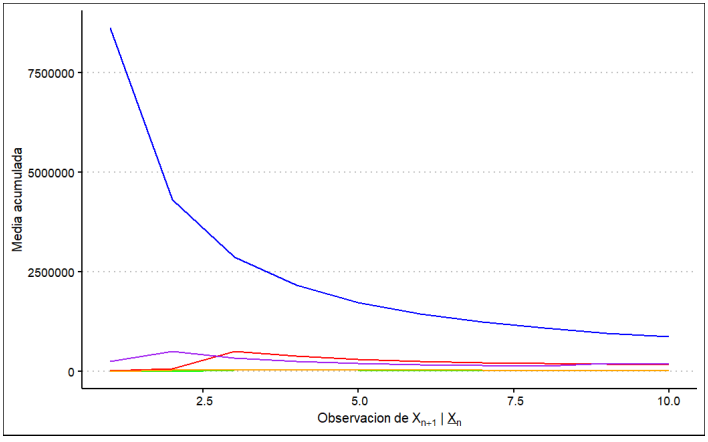
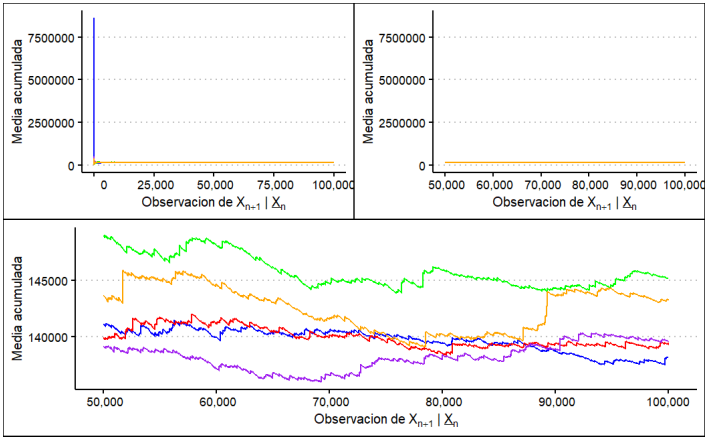


Figura 3.35. Media acumulada para varias cadenas de distribución predictiva posterior

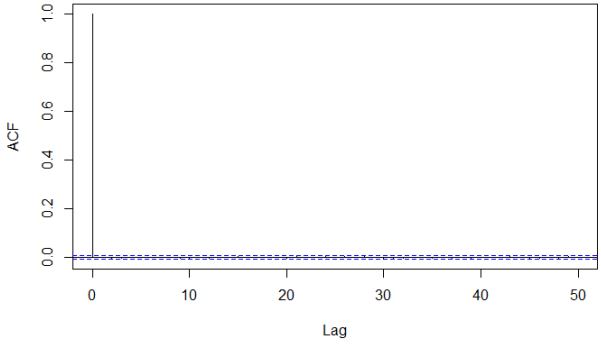


En las gráficas de la figura 3.35. puede notarse que sin importar el estado inicial de la cadena, ésta, en promedio, converja a cierta estabilidad después de un cierto número de observaciones. Observe

cómo la simulación azul tiene un punto inicial muy alto y, a pesar de esto, en promedio, esta termina por debajo de algunas cadenas simuladas cuyos puntos iniciales eran menores.

Por último, debe demostrarse la efectividad del proceso. Esto puede probarse a partir de su *effective sample size* y la gráfica de autocorrelaciones. En este caso, al no utilizar el algoritmo Metrópolis-Hastings, el cociente de aceptación no hace sentido.

Figura 3.36. Función de autocorrelación de la distribución predictiva posterior



Cuadro 3.11. Algunos resultados de la distribución predictiva posterior

Mediada	Edad 1 Estado 1	Edad 1 Estado 2	Edad 2 Estado 1	Edad 2 Estado 2	Edad 3
Media	136,738	96,650	112,292	99,616	258,755
Dev. est.	852,298	608,142	500,711	504,967	1,601,541
$\pi_{0.25}$	3,449	3,986	6,050	5,492	7,325
Mediana	14,525	14,781	20,580	18,747	30,031
$\pi_{0.75}$	60,652	54,231	72,282	63,752	121,953
L.I. *HDI*	1.44	4.07	16.61	13.20	5.43
L.S. *HDI*	491,176	351,277	430,424	366,708	911,537
ESS	100,000	100,000	96,402	100,000	100,000

En este caso, se han obtenido mejores resultados de eficiencia en la distribución predictiva posterior que en la distribución posterior. Esto se puede ver en que la función de autocorrelación cae rápidamente a cero, a diferencia de las gráficas mostradas en la figura 3.31. de la distribución posterior. Por otro lado, el *effective sample size* es mucho más cercano a la cantidad de observaciones simuladas.

Sin embargo, no se debe perder de vista que la distribución predictiva posterior está íntimamente ligada con la distribución posterior. Misma que a pesar de haber resultado efectiva en tamaño, contaba con algunos pasos autocorrelacionados.

Capítulo 4

Modelo colectivo de riesgos

4.1. Modelo coletivo de riesgos

Las compañías de seguros agrupan riesgos parecidos con el fin de que, en conjunto, los riesgos individuales se combinan para reducir el riesgo total de la cartera. Por lo anterior, se ha definido la agrupación de estos riesgos como sigue

Definición 8 (Modelo colectivo de riesgos)**.** *El modelo colectivo de riesgos es la suma de N variables aleatorias de pérdidas individuales, S, donde N es también una variable aleatoria y se puede escribir como sigue (Kuglman, 2013).*

$$S = X_1 + X_2 + ... + X_N, N = 0, 1, 2, ...$$

Este modelo cumple con los siguientes supuestos

1. *Si se condiciona $N = n$, las variables aleatorias de pérdidas individuales son independientes e idénticamente distribuidas.*

2. *Si se condiciona $N = n$ las variables aleatorias de pérdidas individuales son independientes de N.*

3. *La variable aleatoria N no depende de las variables aleatorias de pérdidas individuales.*

En resumen, se define N como la distribución del conteo de reclamaciones; mejor conocida como distribución de frecuencia. Por otro lado, X es la distribución de monto reclamado; o bien distribución de severidad.

A continuación serán definidas algunas propiedades de S.

$$E[S] = E[E(S|N)] = E[NE(X)] = E(N) * E(X)$$

$$Var[S] = E[N]Var[X] + Var[N]E[X]^2$$

Las distribuciones de frecuencia y severidad para la cartera de Alfa Seguros fueron definidas en los capítulos 2 y 3 respectivamente. Antes de continuar, recuerde que

$$N \sim Pois(\lambda)$$

Además, recuerde que X tiene la distribución cuya forma analítica es desconocida, pero se sabe que proviene de la distribución predictiva posterior del análisis Bayesiano realizado en el capítulo anterior.

Clásicamente, la prima de un seguro de *no vida* es definida como la esperanza del modelo colectivo de riesgos.

$$Prima = E(S) = E(N)E(X)$$

Cuadro 4.1. Primas individuales de riesgo de modelo colectivo de riesgos

Edades	Prima Estado 1	Prima Estado 2
Edad 1	14,492.57	15,347.15
Edad 2	2,548.71	2,699.00
Edad 3	6,115.74	6,476.67
Edad 4	7,347.84	10,908.85
Edad 5	7,234.31	10,740.28
Edad 6	6,335.99	9,406.61
Edad 7	6,597.06	9,794.21
Edad 8	5,167.07	7,671.20
Edad 8*	12,205.64	18,286.48
Edad 9	14,816.90	22,198.70

Edad 8* Son personas de 60 años las cuales caen en edad 8 para la frecuencia, pero que fueron tomadas como adultos mayores para la severidad.

Una de las formas en las que es posible validar que el resultado teórico es verdadero para el modelo considerado, es a través de simulaciones. El procedimiento para simular sería el siguiente:

1. Simule una observación $n \sim$Poisson(λ)
2. Tome una muestra de tamaño n del vector de siniestros
3. Una vez obtenidos $X_1, X_2, ..., X_n$ súmelos. Esa suma es conocida como S.
4. Obtenga un número suficientemente grande de realizaciones de S. Para este estudio, fue considerado que 50,000 era un número grande de observaciones. Es decir, $S_1, S_2, ..., S_{50,000}$

5. Promedie el vector de S_i

Cuadro 4.2. Primas de riesgo de modelo colectivo de riesgos

Edades	Prima Estado 1 Teórica / Simulada	Prima Estado 2 Teórica / Simulada
Edad 1	7,927,433 / 7,930,320	2,225,337 / 2,251,960
Edad 2	303,296 / 305,415	75,572 / 76,061
Edad 3	1,418,852 / 1,420,200	446,869 433,902
Edad 4	9,015,811 / 8,999,628	3,927,185 / 3,944,494
Edad 5	4,369,523 / 4,351,137	1,664,744 / 1,653,669
Edad 6	3,053,946 / 3,056,992	1,006,507 / 1,021,355
Edad 7	2,599,243 / 2,588,753	803,125 / 802,867
Edad 8	2,407,855 / 2,411,569	774,790 / 773,706
Edad 8*	5,687,821 / 5,749,943	1,846,935 / 1,867,686
Edad 9	5,052,564 / 5,033,852	1,442,916 / 1,427,188

Observe que el valor simulado, no es tan lejano del valor teórico.

4.2. Suficiencia de la prima

La investigación ha producido a una *prima justa*, pero ¿es esta prima suficiente para hacer frente a las obligaciones de la aseguradora? En teoría, la respuesta inmediata debería ser: en promedio, sí. Sin embargo, en Actuaría existe todo un campo de investigación dedicado a estudiar si la prima cobrada es suficiente para que las aseguradoras no caigan en su principal preocupación: la insolvencia.

Para lo anterior, se define una función que depende del capital inicial u, un monto de primas p y la convolución de siniestros S. Además, la

ruina, al ser un evento que ocurre a través del tiempo, se define como proceso estocástico. Por lo tanto, es necesario involucrar una variable t.

$$R_t = u + pt - \sum_{i=1}^{N_t} X_i$$

Donde:

- u es el capital inicial
- p son las primas recibidas
- t es el tiempo al momento de medición
- N_t es el número de siniestros al tiempo t
- X_i es una siniestro.

De esta forma, el proceso de ruina no es otra cosa que conocer el estado del capital a lo largo de tiempo. Ahora, la preocupación es que este proceso se encuentre por debajo de 0 para alguna $t > 0$. Si la ruina es definida como la probabilidad de caer en insolvencia en cualquier momento ψ dado un capital inicial, entonces puede escribirse como sigue (Søren & Hanshörg, 2010):

$$\psi(u) = P\left(\inf_{t\geq 0} R_t < 0\right) = P\left(\inf_{t\geq 0} R_t < 0 \,|R_0 = u\right)$$

De manera similar, puede definirse un proceso de ruina a tiempo finito. De esta forma, la probabilidad de ruina a tiempo finito se escribe como sigue (Søren & Hanshörg, 2010):

$$\psi(u,T) = P\left(\inf_{0\leq R_t\leq T} R_t < 0\right) = P\left(\inf_{0\leq R_t\leq T} R_t < 0 \,|R_0 = u\right)$$

Al ser el proceso de ruina a tiempo infinito el más complicado de lograr en términos de solvencia, lo que se podría hacer es simular para un periodo de tiempo prolongado cómo sería el comportamiento de las reclamaciones y la recepción de primas a lo largo del tiempo dado un capital inicial. Para estos ejercicios, se ocuparán tanto primas como siniestros en términos reales.

Lo anterior, resulta ser una hipótesis razonable debido a que los seguros de gastos médicos son seguros de corto plazo. Es decir, pueden retarificarse todos los años. Por lo tanto, las tarifas podrían ser ajustadas para que, en términos reales, sean idénticas a las del año anterior.

Se observará la probabilidad de ruina con una prima justa y sin considerar un capital inicial. Como ya se mencionó, es imposible simular infinitamente, pero para efectos prácticos, se considerará tiempo infinito como 1,000 años.

Si bien un horizonte de 1,000 años parece exagerado, lo anterior fue definido de esa forma debido a la propiedad de escalamiento del proceso Poisson; el cual es generador del proceso S.

Ahora, sea $N(t)$ un proceso Poisson, este cumple con 3 condiciones (Ross, 2013):

i) $N(0) = 0$

ii) $N(t)$ es independiente de $N(t+s) - N(t)$, $t < s$ (incrementos independientes)

iii) $\forall t, s \geq 0$ y $h > 0$, $N(t+h) - N(t) \sim N(s+h) - N(s)$ (incrementos estacionarios; el incremento solo depende de h)

Recuerde que $N(t) \sim Pois(\lambda t)$. Por otro lado, sea k el tamaño de la cartera de Alfa Seguros donde $k = 1$ es el tamaño de la cartera actual,

entonces:

$$N(t) = N\left(\frac{t}{k}k\right) \sim Pois\left(\frac{t}{k}k\lambda\right)$$

De esta forma, es posible hacer alguna combinación lineal que tenga un sentido más práctico. Por ejemplo, sea $k = 10$, entonces el proceso puede interpretarse como que el proceso simulado por mil años es equivalente a simular por cien años una cartera 10 veces más grande que la que tiene Alfa Seguros. Recuerde que $t = 1,000$; defina a $t' = \frac{t}{k}$.

$$N(t) = N(1,000) = N\left(\frac{1,000}{10}(10)\right) = N(100(10)) = N(t'k)$$

$$N(t'k) \sim Pois(t'k) = Pois((100)(10)) = Pois(1,000)$$

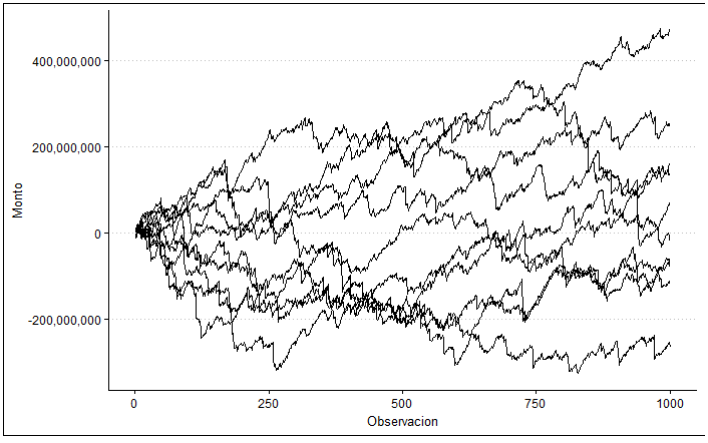
Las probabilidades de ruina mostradas en el cuadro 4.3. fueron obtenidas con un capital inicial de 0.

Cuadro 4.3. Probabilidad de ruina para una prima justa sin capital inicial

Edades	Probabilidad de ruina Estado 1	Probabilidad de ruina Estado 2
Edad 1	0.466	0.452
Edad 2	0.423	0.356
Edad 3	0.452	0.433
Edad 4	0.498	0.464
Edad 5	0.480	0.464
Edad 6	0.478	0.456
Edad 7	0.459	0.445
Edad 8	0.465	0.443
Edad 8*	0.471	0.438
Edad 9	0.464	0.434

Como se puede observar, las probabilidades están cerca de 0.5. Lo anterior, puede esperarse debido a que la prima calculada es *justa*. Como se mencionó, una prima justa es aquella que logra cubrir en promedio las reclamaciones de un año.

Figura 4.1. Gráfica del proceso de ruina sin capital inicial



En la figura 4.1., se pueden observar 10 trayectorias del proceso de ruina simulado. Gráficamente también se puede ver que la mitad de las trayectorias pasan por debajo del 0. Debido a la similitud de las gráficas, se mostrará el de Edad 1 y Estado 1 y el resto se dejarán en los Anexos.

Existen dos formas de reducir la probabilidad de ruina a partir de sus parámetros: el capital inicial y las primas recibidas. Con la finalidad de priorizar el no incrementar la prima del cliente, lo primero que se intentará es ajustar el capital inicial igualándolo al Capital Mínimo Pagado el cual es de de 1,704,243 UDIs para la Instituciones de seguros que operen el ramo de salud, incluyendo accidentes personales y/o gastos médicos mayores según el Anexo 6.1.2. de la Circular Única de Seguros y Fianzas. El valor de la UDI fue tomado de

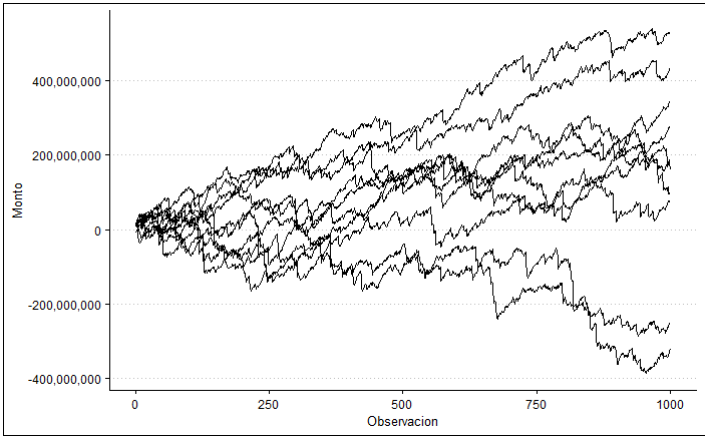
la fecha del 30 de junio de 2023 (7.766768) publicada por Banco de México.

Cuadro 4.4. Probabilidad de ruina para una prima justa con el Capital Mínimo Pagado

Edades	Probabilidad de ruina Estado 1	Probabilidad de ruina Estado 2
Edad 1	0.448	0.378
Edad 2	0.267	0.096
Edad 3	0.354	0.236
Edad 4	0.422	0.402
Edad 5	0.376	0.332
Edad 6	0.369	0.302
Edad 7	0.326	0.294
Edad 8	0.355	0.292
Edad 8*	0.428	0.293
Edad 9	0.420	0.372

Como se puede observar, al comparar el cuadro 4.4. contra cuadro 4.3. se observa que si bien la probabilidad de ruina disminuyó, esta disminución no hace que la probabilidad de ruina llegue a niveles en los cuales un inversionista se sienta cómodo. A continuación se muestran 10 trayectorias de este proceso.

Figura 4.2. Gráfica del proceso de ruina con Capital Mínimo Pagado



Gráficamente, no se nota mucha diferencia debido a que en ambas las utilidades y pérdidas de 1,000 años está entre 400 millones y -400 millones de pesos. Por lo tanto, un capital de 9 millones de pesos no impacta significativamente a la mayoría de las categorías.

Con base en los resultados anteriores, una compañía de seguros requeriría alrededor de 400 millones de pesos para sostener al grupo que se muestra en la figura 4.2.

Si aumentar el capital inicial no funciona, la otra alternativa es aumentar la prima cobrada. La forma en la que se puede aumentar la prima es conocida como recargos de seguridad. Para este estudio, el recargo de seguridad se hará como sigue:

$$\text{Prima recargada} = Prima(1+\theta)$$

Donde θ será definida en función de la probabilidad de ruina. Es decir, se busca una θ tal que $\psi(u) < 0.005$. Lo anterior para buscar la

consistencia con el cuantil del Requerimiento de Capital de Solvencia.

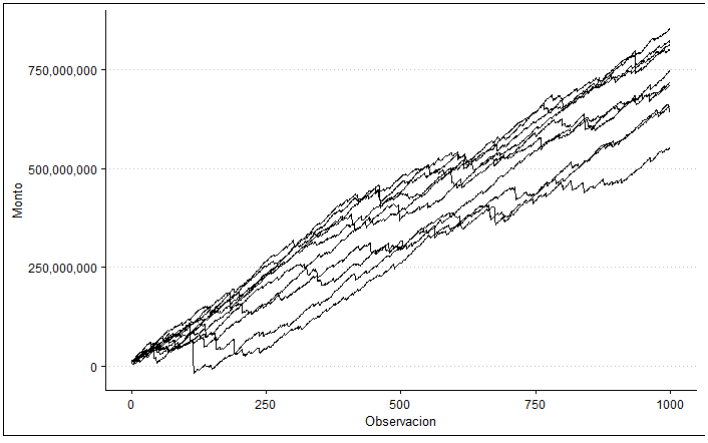
A continuación se el cuadro 4.5. donde se podrán observar las probabilidades de ruina y los recargos de seguridad requeridos para lograr las probabilidades de ruina menores al umbral fijado.

Cuadro 4.5. Probabilidad de ruina para una prima recargada con el Capital Mínimo Pagado

Edades	Probabilidad de ruina Estado 1	θ	Probabilidad de ruina ruina Estado 2	θ
Edad 1	0.002	0.30	0.001	0.50
Edad 2	0.002	1.10	0.003	1.40
Edad 3	0.002	0.50	0.003	0.70
Edad 4	0.004	0.10	0.002	0.25
Edad 5	0.001	0.20	0.004	0.30
Edad 6	0.002	0.1	0.004	0.35
Edad 7	0.002	0.1	0.003	0.45
Edad 8	0.002	0.1	0.002	0.50
Edad 8*	0.001	0.60	0.003	0.60
Edad 9	0.001	0.60	0.001	1.30

Como se puede observar, la probabilidad de ruina ha disminuido a niveles con los que un inversionista se sentiría cómodo. Por lo tanto, se puede decir que estas primas, a pesar de no ser justas, garantizan la solvencia de la compañía de seguros.

Figura 4.3. Gráfica del proceso de ruina con Capital Mínimo Pagado y prima recargada



Asimismo, la imagen muestra cómo una prima con recargo de seguridad cambia la tendencia del proceso. Tendría que ocurrir un año catastrófico para lograr que esta tendencia a la alza caiga en ruina. En algunos casos, el recargo de seguridad es más de 1. Sin embargo, al observar las primas de estos grupos, no fueron exageradamente altas.

Cuadro 4.6. Primas individuales recargadas de riesgo de modelo colectivo de riesgos que garantizan la solvencia

Edades	Prima Estado 1	Prima Estado 2
Edad 1	18,840.33	23,020.72
Edad 2	5,352.296	6,477.61
Edad 3	9,173.609	11,009.82
Edad 4	8,082.63	13,636.06
Edad 5	8,681.17	13,962.37
Edad 6	7,603.19	12,698.92
Edad 7	7,916.47	14,201.60
Edad 8	6,200.48	11,506.79
Edad 8*	19,529.00	29,258.37
Edad 9	23,707.04	51,057.01

Como se puede observar en la cuadro 4.6., las primas de riesgo obtenidas al querer garantizar la suficiencia no son tan elevadas. Sin embargo, no es posible compararlas con las primas del mercado debido a que al cotizar cualquier seguro de gastos médicos, se estaría observando una prima recargada con gastos de administración, costos de adquisición y un porcentaje de utilidad. Debido a lo anterior, a pesar de recargar las primas para garantizar la solvencia, aún harían falta estudios de gastos para poder recargar la prima con los porcentajes. El estudio anterior, queda fuera del alcance de la investigación.

Capítulo 5

Cálculo de prima nivelada

En seguros, el mercado mexicano puede dividese en seguros de largo y corto plazo (CUSF). Al mismo tiempo, aquellos seguros de largo plazo tienen una periodicidad mayor a un años y son, en la mayoría de los casos, de vida. Sin embargo, la propuesta de esta investigación es lograr un seguro de largo plazo para el ramo de Gastos Médicos.

La diferencia técnica va más allá de la periodicidad del seguro. Técnicamente, lo complicado de este seguro será lograr nivelar la prima de riesgo. Una prima nivelada, no es otra cosa que tomar las primas *naturales* y a través de una anualidad tomar una prima equivalente que sea la misma por un periodo determinado.

Es decir, en vez de que la prima aumente anualmente, la prima sería la misma a lo largo de la vigencia del seguro. Para esta investigación se ha optado por tomar una vigencia de tres años debido a la alta volatilidad de la tasa de inflación médica. Esto complica hacer un pronóstico adecuado para un plazo más amplio.

Para los clientes, entre más tiempo se nivele la prima, mejor. Sin embargo, después de platicar con algunas personas de la institución y

de un periodo hiperinflacionario como la pandemia, no se encuentran tan abiertos a ampliar demasiado la vigencia de un seguro.

Si bien, un seguro de prima nivelada de gastos médicos resulta muy atractivo para los clientes, para los dueños de las compañías de seguros representa un riesgo más grande. Lo anterior debido a que les impide retarificar. Por lo tanto, si hay altas desviaciones a la alza en las tasas de inflación médica, habría pérdidas técnicas debido a la tarificación del producto. Este tema será abordado más adelante.

La forma en la que se va a nivelar la prima de este seguro será la siguiente.

$$\Pi_x = \frac{\pi_x + \pi_{x+1}(1+i_{med_1})\nu_1 + \pi_{x+2}(1+i_{med_1})(1+i_{med_2})\nu_1\nu_2}{\ddot{a}_{x:\overline{3}|}}$$

$$\ddot{a}_{x:\overline{3}|} = 1 + p_x\nu_1 +_2 p_x\nu_1\nu_2$$

Donde:

- Π_x es la prima nivelada para una persona de edad x
- π_x es la prima natural para una persona de edad x
- i_{med_j} es la tasa de interés médica del año j
- ν_j es el factor de descuento con tasa libre de riesgo. Tasa a la cual serán invertidos los activos.
- $\ddot{a}_{x:\overline{3}|}$ es la anualidad de una persona de edad x temporal por 3 años a tasa libre de riesgo.
- ${}_kp_x$ es la probabilidad de que una persona de edad x sobreviva k años.

Como se puede observar, no se cuentan con las tasas de inflación médica ni las tasas para los factores de descuento. Para la inflación, existen herramientas con las que se pueden estimar estas tasas. Por otro lado, inversiones consideradas libres de riesgo para invertir las primas del seguro serán los CETES.

5.1. Tasa libre de riesgo

BBVA define a la tasa de referencia como el porcentaje al que presta dinero Banco de México en función de su política monetaria. Además, se menciona que esta tasa sirve como guía para saber a qué tasa van a prestar los bancos y otras instituciones financieras.

Por lo anterior, las tasas de instrumentos *libres de riesgo* están fuertemente ligadas a la tasa de referencia establecida por Banco de México. En particular, la tasa de referencia de junio de 2023 fue fijada en 11.25 %.

A continuación se muestran las tasas de interés que ofrecen los CETES a distintos plazos.

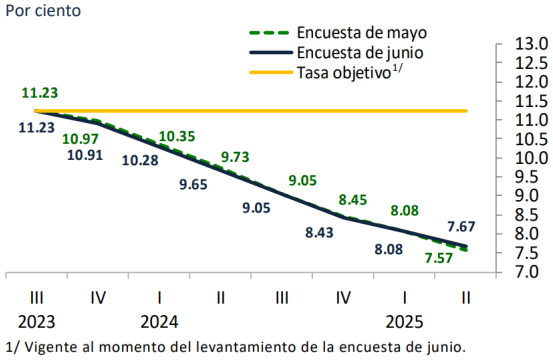
Cuadro 5.1. Tasas ofrecidas por CETES el 29 de junio de 2023.

Plazo	Tasa ofrecida	Plazo	Tasa ofrecida
CETES 28	11.02 %	CETES 182	11.31 %
CETES 91	11.20 %	CETES 364	11.12 %

Por otro lado, en junio, fue publicada la Encuesta sobre las Expectativas de los Especialistas en Economía del Sector Privado: Junio 2023. Dicha encuesta fue realizada por Banxico a 36 grupos de análisis y consultoría económica del sector privado. Dentro de estos destacan Banorte, BBVA,

Citibanamex, HSBC, JP Morgan, Monex, entre otros. En esta se revela que la tendencia de las tasas de referencia y, por lo tanto, las tasas de CETES empiecen a descender a partir de ese momento.

Figura 5.1. Expectativas promedio de la tasa de referencia al cierre de cada trimestre



Fuente: BANXICO 2023

Como se ha analizado, y según las creencias de los expertos, sería razonable tarificar el seguro con una tasa libre de riesgo del 11 % para el primer año y 9.5 % para el segundo año. Lo anterior, fue definido tomando en cuenta la encuesta de junio mostrada en la figura 5.1. y restando aproximadamente 0.25 % debido a que reducir las tasas aumenta las obligaciones de la aseguradora. Esta es una medida conservadora.

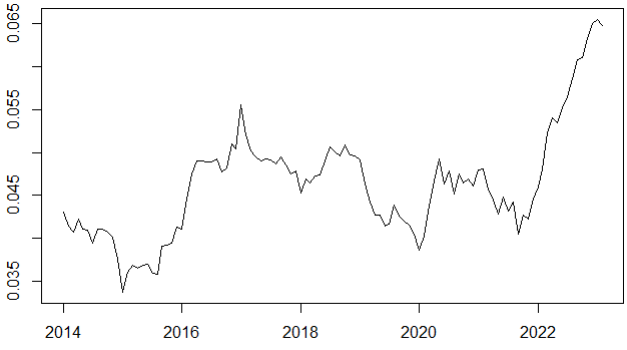
5.2. Inflación médica

La tasa de inflación médica es uno de los mayores retos de esta investigación debido a su alta volatilidad. Además, debido a la

pandemia ocasionada por el COVID 19, la inflación médica ha aumentado a niveles no observados anteriormente.

Para hacer un pronóstico de la inflación médica, se ocupará análisis de series de tiempo. Este análisis tendrá como insumos datos desde enero de 2014 y hasta febrero de 2023. Estos datos de inflación médica son proporcionados por la Asociación Mexicana de Instituciones de Seguros a las aseguradoras que no cuenten con un departamento especializado al análisis de la inflación.

Figura 5.2. Serie de tiempo de inflación médica



Fuente: AMIS 2023

Después de graficar la serie de tiempo de inflación médica, se puede observar una tendencia a lo largo de esta. Sin embargo, a pesar de ser evidente, se utilizará la prueba de hipótesis Dickey Fuller (1979 y 1980) con los 3 modelos: Raíz unitaria, Raíz unitaria con deriva y Raíz unitaria con deriva y tendencia determinista.

Para la prueba Dickey Fuller,

H_0 : Raíz unitaria $\Rightarrow$ Serie no estacionaria v.s. H_1 : Serie estacionaria

Cuadro 5.2. *P - values* para las distintas variantes de la prueba Dickey Fuller.

Variante de prueba	P - value
Raíz unitaria	0.437
Raíz unitaria con deriva	0.952
Raíz unitaria con deriva y tendencia	0.460

En ninguno de los modelos de la prueba de hipótesis fue rechazada la hipótesis nula. Por lo tanto, se puede decir que la serie de tiempo tiene al menos una raíz unitaria. Para eliminar esta raíz unitaria debe aplicarse una diferencia a la serie y repetir los modelos necesarios de la prueba de hipótesis Dickey Fuller hasta garantizar, con un nivel de significancia bajo, que la serie es estacionaria una vez realizada la diferencia.

Cuadro 5.3. *P - values* de la prueba Dickey Fuller para la serie de tiempo diferenciada.

Variante de prueba	P - value
Raíz unitaria	0

Con los resultados anteriores es posible rechazar la hipótesis nula y concluir que, una vez aplicada una diferencia a la serie de tiempo, esta serie ya es estacionaria. Ahora, es necesario evaluar las funciones de autocorrelación acumulada y función autocorrelación parcial.

Figura 5.3. Función de autocorrelación de la serie de tiempo diferenciada

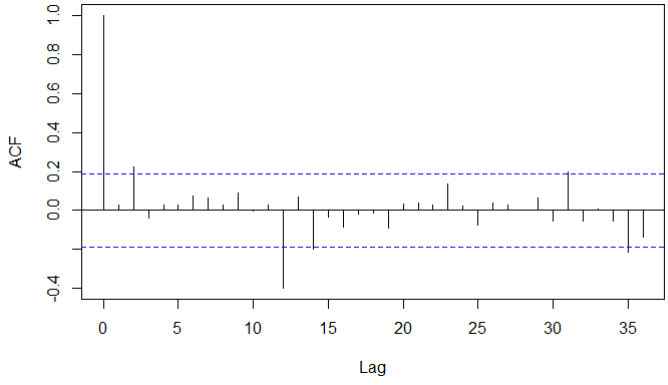
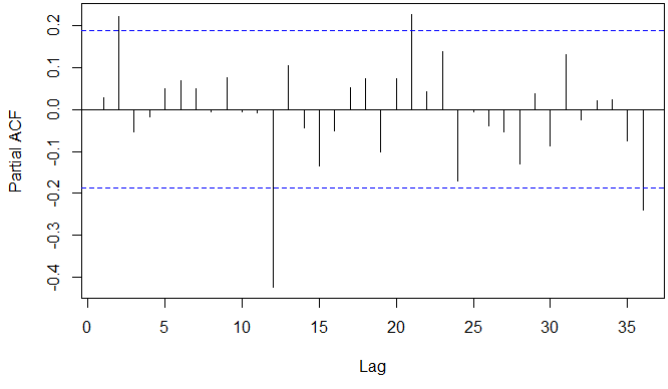


Figura 5.4. Función de autocorrelación parcial de la serie de tiempo diferenciada

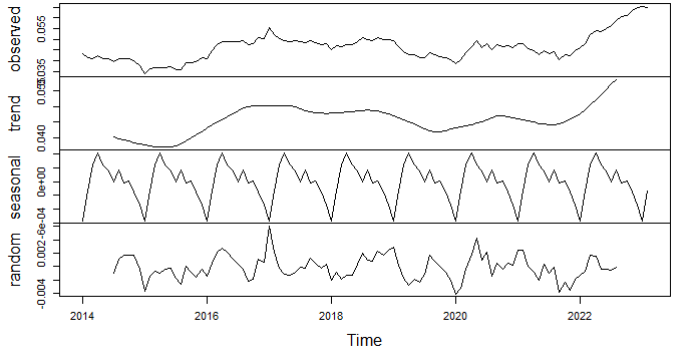


Como se puede observar de la gráfica de autocorrelación parcial apunta a que hay a lo más dos parámetros autoregresivos. Por otro lado, al

observar la gráfica de autocorrelación, se puede observar que hay a lo más dos parámetros de medias móviles. En ambas gráficas hay una desviación grande en el tiempo 12. Lo anterior no implica que hay 12 parámetros autoregresivos o de medias móviles, sino que existe un factor estacional con una temporalidad de 12 periodos. Al estar analizando una serie de tiempo cuyos datos se registran mensualmente, se está hablando de que le serie tiene comportamientos similares de manera anual.

Para validar la hipótesis anterior, es posible descomponer la serie de tiempo en tres componentes: tendencia, estacionalidad y error aleatorio. Existen métodos para hacerlo paso a paso. Sin embargo, la paquetería *stats* [1] de R ayuda a descomponer la serie de tiempo en los componentes que se habían mencionado anteriormente.

Figura 5.5. Descomposición de la serie de tiempo de la inflación médica



Después de observar la gráfica, es claro que la estacionalidad de la serie de tiempo es de forma anual. Por lo tanto, queda validada la hipótesis que se realizó al observar las funciones de autocorrelación y

[1] https://rdocumentation.org/packages/stats/versions/3.6.2

autocorrelación parcial sobre el parámetro estacional y la frecuencia de este. Sin embargo, además de dar soporte a la idea, esta gráfica ayuda a suponer que la relación con el parámetro estacional es autoregresivo más que de medias móviles.

Una vez analizado lo anterior, lo que sigue es determinar un modelo de series de tiempo. En la mayoría de los casos se ocupan modelos $ARIMA(p, d, q)$ para describir a la serie de tiempo. Sin embargo, en este caso particular, será necesario un modelo $SARIMA(p, d, q)(p_s, d_s, q_s)[12]$. Recuerde que el factor autoregresivo es $AR(p)$, el factor de integrando es el $I(d)$ y el de medias móviles es el $MA(q)$. Similarmente para los parámetros estacionales.

Es claro que el parámetro $d = 1$ debido a lo que se analizó en la prueba de hipótesis Dickey Fuller. Sin embargo, antes de proponer un modelo derivado del análisis anterior, se puede observar el resultado de la función *auto.arima* de la librería *forecast* [2].

Esta función propone el modelo $SARIMA(0, 1, 0)(2, 0, 0)[12]$. Sin embargo, proponer dos parámetros autoregresivos estacionales parece excesivo cuando hay una clara tendencia anual. Por otro lado, si bien establecer 0 parámetros autoregresivos y 0 de medias móviles es menor a lo que se había analizado a partir de las gráficas de autocorrelación y autocorrelación parcial, debemos analizar modelos similares a este a fin de poder compararlos.

A continuación se deja una tabla con el modelo propuesto por la función *auto.arima* y una serie de modelos que, con base en el análisis, se cree que podrían ser mejores. Además, la manera de elegir uno será mediante los criterior de información Akaike y Bayesiano, cuyos valores se dejan en la misma tabla.

[2] https://cran.r-project.org/web/packages/forecast/forecast.pdf

Cuadro 5.4. Modelos SARIMA probados con AIC y BIC

Modelo	AIC	BIC
SARIMA(0,1,0)(2,0,0)[12]	-1,094.33	-1,083.57
SARIMA(2,1,0)(1,0,0)[12]	-1,092.60	-1,081.84
SARIMA(2,1,1)(1,0,0)[12]	-1,090.93	-1,077.48
SARIMA(1,1,0)(1,0,0)[12]	-1,089.52	-1,081.44
SARIMA(1,1,1)(1,0,0)[12]	-1,090.57	-1,079.80

Después de analizar los distintos modelos y tomando en cuenta ambos criterios de información, se considera el modelo $SARIMA(2,1,1)(1,0,0)[12]$ como el mejor entre los probados en esta investigación.

Finalmente, antes de realizar un pronóstico se debe observar que los errores residuales de este modelo ya no cuentan con autocorrelación. Para lo anterior, se aplica una prueba de hipótesis llamada Ljung-Box (1978) la cual propone que $n(n+2)\frac{\hat{\rho}^2}{n-k}$ converge en distribución a una $\chi^2_{(1)}$. Como se quiere demostrar que no hay correlación para ningún tiempo, se deben sumar para, después, comparar con algún cuantil el estadístico y rechazar (o no) la hipótesis de que la correlación entre los errores residuales es 0.

Es decir, en este caso:

$$H_0 : \rho_1 = \rho_2 = ... = \rho_k = 0 \text{ v.s. } H_1 : \rho_j \neq 0 \text{ p.a. } j \in \{1, 2, ..., k\}$$

Equivalentemente se podría decir que para los errores del SARIMA ajustado:

H_0 : Proceso de ruido blanco v.s. H_1 : Autocorrelación significativa en el proceso

Definición 9 (Proceso de ruido blanco)**.** *Se dice que un proceso es de ruido blanco si (Brooks, 2019) :*

1. $E[Y_t] = \mu$
2. $Var(Y_t) = \sigma^2$
3. $\gamma_{t-r} = 0 \forall t \neq r$

donde γ es la función de autocorrelación.

El estadístico de prueba Q se escribe como sigue:

$$Q = (n)(n+2)\sum_{j=1}^{h} \frac{\hat{\rho}^2}{n-k}$$

mismo que será comparado contra una $\chi^2_{(h)}$ donde h, para tamaños reducidos de muestra como el de esta investigación, investigadores mencionan que con $h = 3$ se mejora la potencia de la prueba (Hassini & Reza, 2020). Al mismo tiempo, la literatura menciona que haciendo $h \approx ln(n) = 4.7$ la potencia también es alta (Tsay, 2010).

Cuadro 5.5. Prueba Ljung Box para el modelo seleccionado.

h Seleccionada	Estadístico Q	*P - value*
$h = 3$	0.3496	0.95
$h = 5$	1.8188	0.87

Por lo tanto, el modelo seleccionado cumple con los supuestos de series de tiempo y es posible hacer pronósticos a partir de éste. El modelo seleccionado y los parámetros son:

$$(1 - \phi_1 L_1 - \phi_2 L^2)(1 - \Phi_1 L^{12})(1 - L)y_t = (1 + \theta)L\epsilon_t$$

Donde:

- y_t es la serie de tiempo de la inflación médica
- ϕ_1 y ϕ_2 son los coeficientes autoregresivos no estacionales
- Φ es el coeficiente autoregresivo estacional
- θ_1 es el coeficiente de media móvil no estacional.
- L es el operador rezago (lag)
- ϵ_t es un error aleatorio (ruido blanco)

Recuerde que $L^k y_t = y_{t-k}$

Cuadro 5.6. Parámetros estimados del modelos seleccionado.

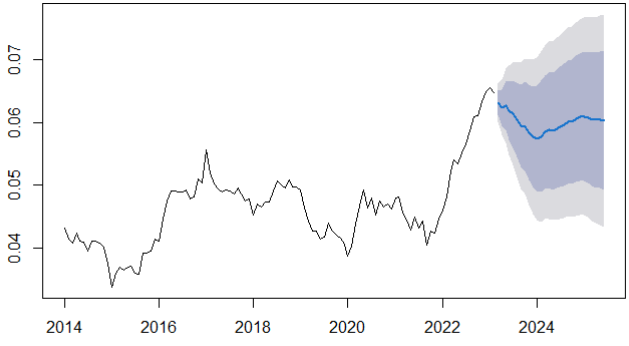
Parámetro	Estimador	Error estandar
ϕ_1	-0.113	0.331
ϕ_2	0.233	0.094
θ_1	0.200	0.336
Φ_1	-0.437	0.900

Al hacer pronósticos de series de tiempo, lo recomendable es hacerlo por un periodo corto de tiempo debido a que, en la mayoría de los casos, el proceso se estabiliza velozmente alrededor de su media y se empieza a observar un pronóstico constante después de algunos periodos.

Sin embargo, en este caso muy particular, por estar cargando información del año pasado, la cantidad de pronósticos que este modelo permite hacer es mayor a un modelo de series de tiempo que no contiene un factor estacional. Lo anterior, no quiere decir que la

varianza no vaya a aumentar a manera que aumente el número de periodos de los cuales se quiere hacer inferencia.

Figura 5.6. Pronóstico de la serie de tiempo con sus bandas de confianza al 80 % y 95 %



Cuadro 5.7. Pronósticos realizados para periodos de interés.

Periodo	Estimador	Estimado al 80 %	Estimado al 95 %
Junio 2024	5.89 %	6.85 %	7.35 %
Junio 2025	6.03 %	7.14 %	7.73 %

Al observar la gráfica, se nota una media bastante razonable. Tal vez, en algún otro contexto, podría ser tomada alguna de las bandas de confianza superiores con el argumento de ser conservadores. Sin embargo, note que las últimas observaciones son el máximo alcanzado por el proceso en toda su historia. Dicho máximo ocasionado por la pandemia.

Asimismo, note un ligero descenso en el último periodo de febrero, lo que no indica que la inflación empiece a descender. Además, en caso de

tomar un pronóstico al 80 % o 95 % se estaría sugiriendo que la inflación médica será mayor que en épocas de pandemia. Lo anterior parece, por ahora, poco razonable.

5.3. Prima nivelada

Como ya se comentó al principio del capítulo, la intención de nivelar la prima es darle seguridad a los asegurados sin comprometer la solvencia de la compañía. Ya se ha hablado de la complejidad del caso particular de gastos médicos por el momento tan complicado que se vive en el mercado.

Para observar cómo se hará para todos los casos, se realizará un ejemplo de la cuenta necesaria. Para realizar este ejemplo, será calculada la prima de riesgo (con los recargos necesarios para garantizar la solvencia mencionados en el **capítulo 4**) para una persona de edad 6 y del grupo de estados de las ciudades por 3 años.

Al calcular las primas naturales por 3 años a partir de la edad 7, tenemos lo siguiente

$$\pi_7 = 18,840.33$$

$$\pi_8 = 18,840.33(1 + 0.0589) = 19,950.03$$

$$\pi_9 = 18,840.33(1 + 0.0589)(1 + 0.0603) = 21,153.02$$

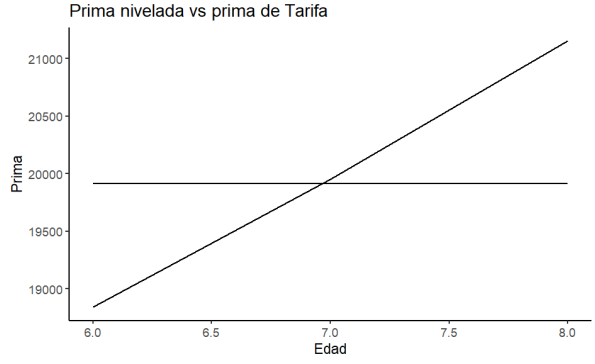
Una vez obtenidos esos valores, solo hace falta nivelar la prima y traer a valor presente lo que sea necesario.

$$\Pi_7 = \frac{18,840.33 + 19,950.03\left(\frac{1}{1.1100}\right) + 21,153.02\left(\frac{1}{1.1100}\right)\left(\frac{1}{1.0950}\right)}{1 + 0.99956\left(\frac{1}{1+0.1100}\right) + 0.99912\left(\frac{1}{1+0.1100}\right)\left(\frac{1}{1+0.0950}\right)}$$

$$\therefore \Pi_7 = 19,915.19$$

Gráficamente, se observa como sigue

Figura 5.7. Prima natural vs prima nivelada para una persona de edad 6



Aunque éste es solo un ejemplo, se puede replicar para el resto de las edades variando la prima natural y la mortalidad según la tabla publicada por la Comisión Nacional de Seguros y Fianzas en el Anexo 5.3.3-a de la Circular Única de Seguros y Fianzas. La tabla con el resto de las primas, debido a su longitud se quedará en los Anexos de la investigación.

Capítulo 6

Conclusiones

6.1. Recapitulación

Esta investigación tiene como propósito principal incentivar la compra de seguros de gastos médicos a través de una prima nivelada. Para lograrlo, se propuso una prima nivelada para un periodo de tres años.

A lo largo de esta investigación, se exploraron problemas de clasificación, selección de modelos, enfoques estadísticos, tarificación y análisis econométrico.

No se encontró que alguno de los dos sexos se accidente con mayor frecuencia ni incurra en mayores costos promedio. Sin embargo, factores como la edad y el estado de residencia resultaron ser determinantes.

Aunque *Alfa Seguros* está en una etapa temprana de su desarrollo, la compañía cuenta con información suficiente para hacer inferencias sobre la distribución de frecuencia de los siniestros. No obstante, debido a la falta de datos en todas las categorías estudiadas (sexo, edad y estado de residencia), fue necesario agrupar las variables de edad y estado para obtener resultados más estables. Estas agrupaciones se basaron en la

significancia de las variables en el modelo estudiado.

Después de realizar una serie de pruebas estadísticas, se determinó que la distribución Poisson describe correctamente la distribución de la frecuencia de los siniestros. Esta elección fue crucial para la tarificación del modelo colectivo de riesgos, dadas las propiedades de la media y la varianza de la distribución seleccionada.

Si bien la agrupación fue suficiente para describir efectivamente la frecuencia de los siniestros, al explorar la severidad y tratar de hacer agrupaciones similares, se descubrió que la variabilidad de los datos proporcionaba resultados demasiado inestables para hacer inferencias sobre su media, asumiéndola desconocida y fija bajo el enfoque frecuentista.

Para resolver este problema, se cambió el enfoque estadístico asumiendo el parámetro desconocido como aleatorio e incorporando información adicional de la Sistema Estadístico del Sector Asegurador de salud, adoptando un enfoque bayesiano. A lo largo del análisis bayesiano, se justificó la selección de la distribución a priori y la distribución de verosimilitud. Una vez obtenidas ambas distribuciones, se buscó la distribución posterior de los parámetros desconocidos. La complicación fue que no existe una forma analítica conocida para dicha distribución.

Al no existir una forma analítica conocida de la distribución, fue necesaria la aplicación de métodos numéricos que, a través de cadenas de Markov, buscan lograr un proceso que simule una muestra aleatoria de la distribución posterior. Para ello, se seleccionó el método de muestreo Metrópolis-Hastings.

Después de simular la muestra, se demostró que la cadena simulada es estable, representativa y eficiente. Si bien demostrar la convergencia de una cadena es complicado, que ésta tenga esas tres características

proporciona evidencia de que es un buen acercamiento a la muestra aleatoria de la distribución límite.

El paso de encontrar la distribución posterior es importante, pero no es el objetivo final del estudio de severidad. En este caso, fue de mayor interés la distribución de la siguiente observación condicionada a los datos ya observados. Esta distribución fue obtenida a través de un método de muestreo simulado. Por lo tanto, se realizaron las mismas pruebas de convergencia de la cadena para mostrar un comportamiento favorable de ésta.

Una vez propuesta la distribución de los siniestros y obtenida una muestra aleatoria de la distribución predictiva posterior, se procedió a tarificar el seguro. Dicha tarificación se hizo con el modelo colectivo de riesgos, agrupando los riesgos individuales para reducir el riesgo del total de la cartera.

Estimar la media de este modelo, fue sencillo debido a las propiedades de la distribución de frecuencia y a la capacidad de poder de cómputo para muestrear de la distribución predictiva posterior, respondiendo así por la solvencia de la compañía de seguros con este método de tarificación.

Para lograr lo anterior, y debido a que *Alfa Seguros* tiene permiso de operar exclusivamente la operación de Accidentes y Enfermedades, se consideró que el proceso de Ruina explicaba de manera adecuada la solvencia de la compañía. Se encontró que, si bien la prima justa no garantiza la solvencia, una prima con un recargo de seguridad sí lo hace. A través de simulación, se determinó el porcentaje necesario que garantiza lo propuesto por Solvencia II: que la ruina ocurra, en promedio, una vez cada 200 años.

Finalmente, al nivelar la prima, se presentó el problema de pronosticar la tasa libre de riesgo y la inflación médica para el periodo

de nivelación de prima de tres años. Para la selección de la tasa libre de riesgo, se consultó la Encuesta sobre las Expectativas de los Especialistas en Economía del Sector Privado: Junio 2023 publicado por BANXICO, obteniendo las tasas a los distintos plazos necesarios. Por otro lado, la inflación médica se pronosticó mediante un modelo de series de tiempo estacionarias (SARIMA), permitiendo hacer pronósticos más estables y prolongados que los de un proceso ARIMA tradicional.

En resumen, esta investigación ha cubierto la descripción adecuada de la muestra de la población contenida en *Alfa Seguros*, la elección de modelos adecuados para la cantidad y calidad de los datos obtenidos, la propuesta de soluciones ante la falta de madurez en el mercado, y la tarificación con un enfoque tanto social como de negocio. Además, se ha logrado pronosticar la inflación médica, promoviendo la certeza de la prima durante un periodo corto de tiempo sin perder de vista el interés de los empresarios.

6.2. Conclusiones e investigación futura

Hasta el momento, en el mercado asegurador mexicano, los seguros de gastos médicos son cobrados a prima natural. Esta tesis aborda con éxito la construcción de un producto de gastos médicos a través de una prima nivelada por tres años. Lo anterior podría atraer a más clientes a adquirir este tipo de productos financieros debido a que se traslada la incertidumbre sobre inflación y tasas (junto con otros factores de riesgo que aumentan la prima del seguro) hacia las aseguradoras.

Además, esta investigación propone el tratamiento que podrían darle las aseguradoras que no cuentan con un volumen de cartera suficiente o que cuentan con poca experiencia en el mercado

asegurador para tarificar seguros de gastos médicos con experiencia propia. De esta forma, las compañías de seguros podrían dejar de usar el método estatutario y apoyarse de la información de mercado para que sus precios reflejen el comportamiento de la cartera y no del mercado en general. Lo anterior, podría ser complementado por muchas aseguradoras no solo por información de mercado, sino información histórica.

Al mismo tiempo, los resultados obtenidos, cumplen con los criterios propuestos por Solvencia II. Las técnicas usadas de simulación a tiempo infinito (1,000 años) fueron útiles para demostrar que con probabilidad 0.005 la compañía de seguros sería capaz de pagar sus obligaciones a lo largo de ese horizonte. Sin embargo, debido al proceso ocupado (Poisson) es posible interpretar que Alfa Seguros sería solvente con un tamaño de cartera 10 veces más grande por 100 años. Lo anterior es de gran utilidad debido a que la cartera de la aseguradora que se está estudiando sigue creciendo a tasas crecientes. Por lo tanto, la prima propuesta sería suficiente aún si crecen 10 veces su cartera. Este resultado produce la tranquilidad que aún con movimientos económicos muy fuertes en tasas de interés, mortalidad u otros factores de riesgo, la prima seguiría alcanzando para pagar los siniestros de los asegurados.

Al abordar factores económicos como la inflación médica, se propuso un modelo econométrico que describe de manera efectiva el comportamiento de la serie de tiempo asociada. Este modelo resulta fundamental para el desarrollo de la prima nivelada, ya que uno de los mayores desafíos al tarificar productos de gastos médicos radica en la incertidumbre sobre la evolución de la inflación médica. Mediante un pronóstico adecuado de esta variable, es posible calcular una prima justa que refleje con precisión los costos proyectados para los años

futuros.

A pesar de haber sido satisfactoria la tarificación del producto, la investigación puede complementarse ampliamente en un futuro. Una de las principales preguntas que quedan pendientes por responder es: ¿este producto es competitivo en el mercado? Esta pregunta es fundamental para la implementación de este producto. Sin embargo, debido a no contar con un estudio profundo de gastos de administración, adquisición y utilidad es imposible comparar las primas del ejercicio presentado en este estudio contra las primas de mercado.

Por otro lado, podrían explorarse métodos de muestreo más específicos para aquellas distribuciones altamente correlacionadas que no fueron consideradas en esta investigación. Esta representa una gran oportunidad, ya que, aunque el muestreo de estas distribuciones sea más desafiante, podrían ofrecer una descripción más precisa de la población de la cartera.

Asimismo, es posible que los resultados generados en las pruebas de suficiencia mediante procesos de ruina tengan una alta variabilidad. Lo anterior debido a que, como se observó a lo largo de la sección 3.1., los siniestros de gastos médicos son de cola pesada. Lo anterior provoca "brincos"que pueden observarse la figura 4.2. y en las figuras del anexo D. En este caso fue usado un número suficientemente grande de simulaciones para reducir la varianza del estimador. Sin embargo, podría investigarse sobre métodos que sean más efectivos al momento de simular estos procesos.

Finalmente, el cuadro 4.5. podría dejar a algunas personas inconformes debido a que, para algunos grupos, el recargo realizado para lograr que la probabilidad de ruina fuera a lo más 0.005 es de más del 100 %. Es decir, se le estaría cobrando más del doble de la prima para lograr la solvencia. Para resolver esta inconformidad, podría

plantearse una bonificación al final de la vida del seguro donde los asegurados reciban una proporción de la utilidad técnica generada durante el plazo de su seguro. Es decir, si la cartera presenta menores reclamaciones a las esperadas, podría devolverse un porcentaje al asegurado de la prima pagada.

Referencias

AMIS, “El seguro de gastos médicos en méxico. relevancia y tendencias.”, julio de 2023, https://amisprensa.org/public/documentos/seguro-de-gastos-medicos-36.pdf.

Asmussen, Soren y Albrecher Hansjorg. Ruin Probabilities (2nd Edition). World Scientific Publishing Co Pte Ltd, 2010.

BANXICO. “Encuesta sobre las expectativas de los especialistas en economía del sector privado: Junio de 2023”. Banxico, banco central, Banco de México, 3 de julio de 2023. https://www.banxico.org.mx/publicaciones-y-prensa/encuestas-sobre-las-expectativas-de-los-especialis/ %7B384D2689-12A4-02F7-8B3A-D18F36B3281A %7D.pdf.

BANXICO. “Resumen ejecutivo del informe trimestral abril - junio 2023”. Banxico, banco central, Banco de México, 30 de agosto de 2023. https://www.banxico.org.mx/publicaciones-y-prensa/informes-trimestrales/ %7BC930F98F-2AEB-ACCF-B5CB-FAEBB4D7387A %7D.pdf.

BBVA. “¿Qué es la tasa de referencia?” Consultado el 30 de junio de 2024. https://www.bbva.mx/educacion-financiera/t/tasa_de_referencia.html.

Brooks, Chris. Introductory Econometrics for Finance. University of Cambridge ESOL Examinations, 2019.

Comisión Nacional de Seguros y Fianzas. Circular única de seguros y fianzas. CUSF. 2024. https://www.gob.mx/cms/uploads/attachment/file/882077/02_Circular_única_de_Seguros_y_Fianzas_compulsada_sin_Anexos_05-06-2024.pdf.

Comisión Nacional de Seguros y Fianzas.. Circular única de seguros y fianzas anexo 5.3.3-A. CUSF. 2016. https://www.gob.mx/cms/uploads/attachment/file/73530/ANEXO_5.3.3-a.pdf.

Dobson, Annette y Barnett Adrian. An Introduction to Generalized Linear Models. Chapman and Hall/CRC, 2008.

Esquivel, Gerardo. ”Los impactos económicos de la pandemia en México”. Banxico, julio de 2020. https://www.banxico.org.mx/publicaciones-y-prensa/articulos-y-otras-publicaciones/ %7BD442A596-6F43-D1B5-6686-64A2CF2F371B %7D.pdf.

Gelman, Andrew, Carlin John B., Hal S. Stern, David B. Dunson, Aki Vehtari y Donald B. Rubin. Bayesian Data Analysis. Chapman and Hall/CRC, 2013.

Hassani, Hossein y Yeganegi Mohammad. “Sum of Squared ACF and the Ljung–Box Statistics”. Physica A: Statistical Mechanics and Its Applications 520 (abril de 2019): 81–86.

Hastings W. K.,“Monte Carlo Sampling Methods Using Markov Chains and Their Applications”, Biometrika 57, n.º 1 (1 de abril de 1970), https://doi.org/10.1093/biomet/57.1.97.

Hyndman, Rob, Athanasopoulos George, Christoph Bergmeir, Gabriel Caceres, Leanne Chhay, Kirill Kuroptev, Mitchell O’Hara-Wild et al. “Package ‘Forecast’”. The Comprehensive R Archive Network, 20 de junio de 2024. https://cran.r-project.org/web/packages/forecast/forecast.pdf.

Klugman, Stuart A., Panjer Harry H. y Willmot Gordon E. Loss Models. Hoboken, NJ, USA: John Wiley & Sons, Inc., 2013.

Kruschke, John. Doing Bayesian Data Analysis: A Tutorial With R, JAGS, and Stan. Elsevier Science & Technology Books, 2014.

McCullagh, P. y J. A. Nelder. Generalized Linear Models. Boston, MA: Springer US, 1989.

Mendoza Escamilla, Viridiana. “Los seguros en México después de la pandemia”. Forbes México, 23 de junio de 2021. https://www.forbes.com.mx/nuestra-revista-los-seguros-en-mexico-despues-de-la-pandemia/.

Metropolis Nicholas et al.,“Equation of State Calculations by Fast Computing Machines”, Journal of Chemical Physics 21, n.º 6 (junio de 1953), https://doi.org/10.1063/1.1699114.

Nagler, Thomas, Ulf Schepsmeier, Jakob Stoeber, Eike Christian Brechmann, Benedikt Graeler, Tobias Erhardt, Carlos Almeida et al. “Statistical Inference of Vine Copulas”. The Comprehensive R Archive Network, 10 de julio de 2023. https://cran.r-project.org/web/packages/VineCopula/VineCopula.pdf.

Nelsen, Roger B. Introduction to Copulas. Springer London, Limited, 2007.

R. Core Team. “Stats Package - RDocumentation”. Home - RDocumentation, 31 de diciembre de 1969. https://rdocumentation.org/packages/stats/versions/3.6.2.

Ross, Sheldon. First Course in Probability. Pearson Education, Limited, 2013.

Tsay, Ruey S. Analysis of Financial Time Series. Wiley & Sons, Incorporated, John, 2010.

World Health Organization (WHO), “Trastornos congénitos”, 17 de febrero de 2023, https://www.who.int/es/news-room/fact-sheets/detail/birth-defects.

WTW, “Mayor inflación y alta demanda de servicios médicos impulsan costos globales de beneficios médicos hasta 10 % en 2023”, WTW, 5 de diciembre de 2022, https://www.wtwco.com/es-mx/news/2022/10/mayor-inflacion-y-alta-demanda-de-servicios-medicos-impulsan-costos-globales-de-beneficios-medicos.

Anexos

La consulta de los anexos puede realizarse escaneando el siguiente código QR.

Credibilidad Bayesiana para el
cálculo de una prima
nivelada de gastos médicos
escrito por Daniel Maximiliano Guerrero Meneses,
se terminó de imprimir en diciembre de 2024
en los talleres de Tesis Martínez.
República de Cuba 99, colonia Centro,
Ciudad de México.