TARIFICADOR DE UN SEGURO DE RESPONSABILIDAD CIVIL DE AUTOMÓVILES

Universidad de Alcalá
Master en Ciencias Actuariales y Financieras
Facultad de Ciencias Económicas y Empresariales

Trabajo realizado por Carlos Polo Melero
Dirigido por Alejandro Sánchez Sustaeta
Y Juan José Lucio Fernández

Septiembre 2022

Índice

Resumen 2
Palabras clave 3
1. Introducción 4
2. Modelos Lineales generalizados 6
2.1. Familia exponencial 6
2.2. Elementos principales 7
2.2.1. Componente aleatoria 7
2.2.2. Componente sistemática 8
2.2.3. Función Link o enlace 8
2.3. Análisis y contrastes 9
2.3.1. Deviance 10
2.3.2. AIC y BIC 10
3. Base de datos y variables 11
3.1. Base de datos final 12
3.1.1. Edad 13
3.1.2. Antigüedad del vehículo 14
3.1.3. Valor del vehículo 16
3.1.4. Sexo 17
3.1.5. Bonus 18
4. Modelo de frecuencias 20
4.1. Modelo de frecuencias de 2 variables 21
4.2. Modelo de frecuencias de 3 variables 24
5. Modelo de costes medios 27
5.1. Modelo de costes medios de 2 variables 28
5.2. Modelo de costes medios de 3 variables 30
6. Prima de riesgo y tarifa final 33
6.1. Prima de riesgo bonificada 33
6.2. Tarifa final 35
6. Conclusiones 38
7. Bibliografía 39
8. Anexos 40

Resumen

El objetivo del estudio consiste en tarificar la garantía de RC a partir de una cartera de automóviles de una entidad aseguradora, para lo cual se utilizan técnicas de GLM obteniendo a partir de ellas un modelo de frecuencias y otro de costes medios.

Posteriormente, se obtiene la prima de riesgo del seguro, con la que se elabora la prima comercial y la tarifa final del seguro.

Algunas de las variables más relevantes para la obtención de esta tarifa son la edad del asegurado, el sexo de este, la antigüedad del vehículo y su valor. Además, gracias a los modelos obtenidos se puede observar el peso que tiene cada una de estas categorías en la frecuencia de los siniestros y su coste y por lo tanto en la prima de riesgo del seguro.

En cuanto a la organización del estudio, en el primer punto se presenta la introducción mostrando las motivaciones y explicando brevemente el proceso de tarificación. Posteriormente se explican los GLM utilizados en el trabajo.

También se presenta la base de datos utilizada para llevar a cabo los cálculos del estudio, se nombra como se ha limpiado y tratado, y se muestran las variables definitivas una a una y las categorías que la forman.

Uno de los puntos más importantes es el que muestra los procedimientos seguidos para la obtención de los modelos de frecuencias y de costes medios. De forma final, se ha decidido realizar dos tipos de modelos ya que siendo ambos significativos y coherentes, esto permite ver las diferencias entre ellos. En uno de ellos, se tendrá en cuenta la variable del sexo del asegurado, a pesar de que en la realidad esto está prohibido, sin embargo, de forma académica se puede observar el efecto que puede llegar a tener en la tarificación de seguros.

Para finalizar, se calcula la prima de riesgo para la tarifa final y se crea una calculadora a través de la herramienta Excel para la tarificación de cualquier seguro de RC según el tipo de asegurado y vehículo, terminando el trabajo con las conclusiones obtenidas de todo el estudio en su conjunto.

Palabras clave

- Tarificación o *pricing*.

- Vehículo

- GLM (Generalized Linear Models o Modelos Lineales Generalizados).

- Variable dependiente o respuesta.

- Variables independientes o explicativas.

- RC (Responsabilidad Civil).

- Frecuencia

- Costes medios

- Prima de riesgo

1. Introducción

El cálculo actuarial dedicado a la tarificación, más comúnmente denominado *pricing*, es de los más llamativos dentro de esta ciencia. Forma parte de los principales retos a los que se enfrentan todas las empresas del sector y se mantiene siempre en el centro de atención de todas ellas, ya que este cálculo afecta e interviene directamente en el negocio asegurador, pues la rentabilidad, el futuro número de clientes o el tipo de estos que tenga la aseguradora dependen estrechamente de ello.

La tarificación de seguros de No Vida sirve para adecuar los precios del producto a los objetivos del negocio, pudiendo ser estos por ejemplo una mayor rentabilidad o el crecimiento de la cartera. Esto explica, que además de calcular de forma correcta los riesgos agregados del seguro sea importante orientar las técnicas de *pricing* a los objetivos de las compañías que las elaboran.

A la hora de tarificar, como se ha mencionado, la principal labor consiste en localizar los riesgos agregados, para lo que se modelizan a través de GLM. Esto conlleva una gran complejidad, ya que las distribuciones tanto de las frecuencias como de los costes medios de los siniestros en el caso de los seguros de automóviles son muy distintas para las diferentes garantías que se cubren. En este caso, el cálculo se va a realizar para tarificar la responsabilidad civil, sin embargo, existen otras garantías como las de daños materiales en las que se incluyen las lunas, el robo o el incendio, los daños propios cubiertos en los seguros de todo riesgo o también otros costes cubiertos en algunos casos como la defensa jurídica y la reclamación de daños entre muchos otros.

Mediante los nombrados GLM (Modelos Lineales Generalizados), en este caso se realizarán dos modelos, uno de frecuencias y otro de costes medios para el posterior cálculo de la prima de riesgo. Una vez hechos estos cálculos el siguiente paso consiste en obtener la prima comercial a partir de la prima de riesgo, pero para ello, entra en juego la estrategia a seguir por parte de la entidad aseguradora, ya que según a esta le interese un mayor volumen de negocio o un mayor beneficio se deben hacer estos cálculos de una forma u otra.

Tal y como se menciona, el actual estudio trata de analizar una base de datos de pólizas y siniestros reales de una empresa aseguradora, tratando de modelizar el comportamiento de estas a partir de GLM según sus características. Con ello se trata de obtener una tarifa de las garantías mínimas que cubre un seguro de automóviles, en este caso de la RC (Responsabilidad Civil).

Esta garantía mínima, la responsabilidad civil del seguro de automóviles, cubre todos los daños tanto corporales como materiales causados en un accidente, exceptuando los daños personales del conductor asegurado. Estos daños pueden ser tanto de otros vehículos o pasajeros de estos, como de mobiliario urbano o peatones del exterior. Eso sí, siempre que el cliente de la entidad aseguradora en cuestión sea culpable, en caso contrario, la RC no se encargaría de cubrir todos estos costes.

2. Modelos Lineales generalizados

El estudio, como se ha mencionado anteriormente, trata de tarificar un seguro de automóviles a través de una cartera de pólizas para la garantía de responsabilidad civil localizando los riesgos y modelizándolos a través de GLM (Generalized Linear Models).

Estos Modelos lineales generalizados se utilizan para modelizar el comportamiento de una variable dependiente o respuesta a partir de un conjunto de variables independientes o explicativas. Este método se diferencia de una regresión lineal habitual en que existe una relación lineal entre las variables independientes citadas anteriormente y una transformación de la media de la variable dependiente. Además, la variable respuesta no sigue necesariamente una distribución normal, sino que habitualmente sigue alguna de las distribuciones encuadradas en la familia exponencial.

2.1. Familia exponencial

La familia exponencial de distribuciones es una de las construcciones clave en los GLM, su expresión general sigue la forma

$$f(y,\theta,\phi) = exp\left\{\frac{y(\theta) - b(\theta)}{a(\phi)} + c(y,\phi)\right\}$$

Donde θ es el parámetro natural o canónico y ϕ es un parámetro adicional de escala o dispersión, si este es conocido es un modelo de la familia exponencial lineal, si es desconocido de dispersión exponencial. Además, $a(\cdot), b(\cdot)\ y\ c(\cdot)$ son funciones específicas.

En su caso, es interesante conocer la relación existente entre la media y la varianza, para lo cual, se calculan las primeras derivadas con respecto de θ del neperiano de la función de verosimilitud, la función score, debiéndose cumplir que:

- El valor esperado de la función score es 0
- La varianza de la función score debe ser igual al valor esperado de la primera derivada de la función score con el signo cambiado.

Dentro de la familia exponencial, las distribuciones más frecuentes son las siguientes:

- Normal o Gaussiana
- Bernoulli
- Binomial
- Poisson
- Gamma
- Geométrica
- Binomial Negativa
- Inversa Gaussiana

Posteriormente, tanto para el modelo de frecuencias como para el modelo de costes medios, la variable dependiente de cada uno de estos análisis seguirá una de las distribuciones de la familia exponencial.

2.2. Elementos principales

Los GLM tienen tres elementos principales, estos son:

- Componente aleatoria
- Componente sistemática
- Función Link o enlace

2.2.1. Componente aleatoria

En un GLM la componente aleatoria consiste en una variable aleatoria Y, llamada variable respuesta, con un numero *n* de observaciones independientes $(y_1, y_2, y_3, \dots, y_n)$ todas y cada una de ellas con la misma distribución.

En algunas aplicaciones, las observaciones de Y son binarias, estas se identifican en este caso como éxito y fracaso, en otros casos cada observación corresponde a un recuento y en otros, estas observaciones pueden ser continuas, aunque de forma más habitual, cada Y_i indica el número de éxitos de entre un número fijo de ensayos.

2.2.2. Componente sistemática

La componente sistemática de un GLM especifica las variables independientes o explicativas, que entran en forma de efectos fijos en un modelo lineal, por lo que las variables x_j se relacionan mediante

$$\alpha + \beta_i x_i + \cdots + \beta_j x_j$$

Esta combinación lineal de variables independientes se denomina predictor lineal. También se puede expresar como un vector $(\eta_1, \ldots, \eta_n)$ tal que

$$\eta_i = \sum_j \beta_j X_{ij}$$

Donde x_{ij} es el valor del j-ésimo predictor en el i-ésimo individuo. En cuanto al término independiente α se obtendría en este caso haciendo que todos los x_{ij} sean igual a 1 para todos los i.

2.2.3. Función Link o enlace

Denotado el valor esperado de Y como $\mu = E(Y)$, entonces la función enlace especifica una función g(·) que relaciona μ con el predictor lineal de la siguiente forma:

$$g(\mu) = \alpha + \beta_i x_i + \cdots + \beta_j x_j$$

Así, la función enlace g(·) relaciona las componentes aleatoria y sistemática.

La función g más simple es la identidad, g(μ) = μ, que da lugar al modelo de regresión lineal clásico, sin embargo, hay más tipos de función enlace, como pueden ser la función log, la logit, la Raíz cuadrada; las más frecuentes se detallarán en la siguiente tabla, en la que se nombrará el tipo de función Link con su g(μ) y se relacionará con el tipo de distribución para la que es canónica cada una de estas.

Tabla 1: Links frecuentes.

Función	g(μ)	Canónica para
Identidad	μ	Normal o Gaussiana
Log	$\ln(\mu)$	Poisson
Potencia	μ^p	Gamma (p=-1) Inversa Gaussiana (p=-2)
Raíz cuadrada	$\mu^{1/2}$	…
Logit	$\ln(p/1-p)$	Binomial

Fuente: Elaboración propia con base en datos Simulación y modelos para el análisis de datos actuariales, TEMA 4, Pablo J. Alonso-Gonzalez.

A pesar de que cada una de las funciones enlace son canónicas para distintas distribuciones, no son solo las únicas que pueden actuar como link para cada una de ellas. Muchas de las distribuciones que se van a nombrar posteriormente y utilizar para los modelos de este estudio son ejemplo de ello.

La familia Gaussiana acepta las funciones enlace identidad, log e inversa; la familia binomial el link log (complementariamente log-log); la familia gamma los links inversa, identidad y log; la familia poisson el link log.

2.3. Análisis y contrastes

Para el siguiente estudio se desarrollarán algunos de los análisis más importantes para determinar el ajuste del modelo estimado y, de esta forma poder elegir entre unos u otros modelos.

Para estos cálculos se debe nombrar previamente el método de obtención de las betas o coeficientes en este tipo de modelos, las cuales, serán aquellas que maximicen el logaritmo neperiano de verosimilitud del modelo estimado, cálculo que nos será útil también para obtener los resultados de los siguientes análisis.

2.3.1. Deviance

El análisis de la Deviance es una generalización del análisis de la varianza para los Modelos Lineales Generalizados obtenido para una secuencia de modelos anidados.

Su expresión general es la siguiente:

$$S = -2ln\lambda = -2[lnf(y;\beta;\phi) - lnf(y;\beta_{sat};\phi)]$$

Donde el subíndice *sat* indica que corresponde al modelo saturado.

Para obtener el valor de la Deviance se debe calcular en un primer lugar el logaritmo neperiano de la verosimilitud del modelo estimado, y calcular a continuación el logaritmo neperiano de la verosimilitud del modelo saturado, para finalmente multiplicar la diferencia de estos dos modelos por (-2).

Una vez obtenido este valor en cada modelo, para comparar entre ellos y observar cual tiene un mejor ajuste, se deberá seguir el criterio de elegir aquel que menor Deviance tenga.

2.3.2. AIC y BIC

Otros estadísticos de comparación entre modelos importantes para el estudio son el AIC (Akaike Information Criterion, Akaike 1973) y el BIC (Bayesian Information Criterion, Schwarz, 1978).

El AIC y el BIC son estimadores de error de predicción y por lo tanto de la calidad relativa de cada modelo estadístico para un conjunto de datos dado, están estrechamente relacionados y se basan, en parte, en la función de verosimilitud.

Para ambos casos un menor valor nos indica un mejor ajuste del modelo. Se definen de la siguiente forma:

$$AIC = -2ln[M_k] + 2p$$

$$BIC = -2lnL + pln(n)$$

Donde ln[M_k] = neperiano de la verosimilitud del modelo M_k, y p = número de regresores estimados.

3. Base de datos y variables

La base de datos utilizada para este estudio está extraída de una entidad aseguradora, por lo que son datos reales, sin embargo, estos han sido trabajados y limpiados previamente y cambiados de formato mediante varios filtros en SAS y luego exportados a Excel para finalmente realizar los modelos en RStudio.

En su inicio la base de datos contenía más variables de las que posteriormente se muestran en los resultados finales, sin embargo, debido a que alguna de ellas no es representativa para este estudio se ha decidido excluirlas del mismo.

La base de datos inicial está dividida en las siguientes variables:

- *Npoliza:* Variable que indica el número de la póliza dentro de la cartera
- *Exposición:* Cantidad de días que el automóvil ha contratado la póliza, en este caso, debido a la normativa de la aseguradora en todos los casos toma el valor de un año.
- *Siniestro:* Indica si se ha sucedido algún siniestro en la póliza en cuestión, en este caso toma el valor de 1 si ha tenido uno o más siniestros o el valor de 0 si no ha tenido siniestros.
- *Coste:* Esta variable toma el valor del coste total de los siniestros se han sucedido en una póliza (Solamente teniendo en cuenta daños materiales, nunca corporales).
- *Sexo:* Variable que toma como valor M en el caso de ser un hombre el tomador o F en caso de ser mujer.
- *Tipovehículo:* Tipo de coche que está asegurado en la póliza.
- *Vehiaños:* Antigüedad del vehículo asegurado en la póliza.
- *Zona:* Variable que indica en qué tipo de área reside el tomador de la póliza.
- *Edad:* Variable que toma como valor la edad del tomador de la póliza.
- *Vehivalor:* Valor que toma el coche asegurado en el momento de contratación de la póliza.
- *Bonus:* Valor del 1 al 5 con el que relacionan las aseguradoras a los asegurados según lo buenos conductores que sean. En este caso toma el valor 1 para los peores asegurados y el valor 5 para los mejores. Esta variable se mide principalmente según los siniestros

que haya tenido el asegurado los últimos años dependiendo de la cantidad de tiempo de la que se tengan referencias de él como conductor.

De las variables mostradas anteriormente las que no se han utilizado para el trabajo son *Tipovehículo* y *Zona* debido a que ninguna de ellas ha resultado representativa para ninguno de los modelos estudiados. En ambos casos parece razonable que estas variables puedan ser significativas a la hora de medir la frecuencia o el coste de los siniestros de un asegurado, sin embargo, no en este estudio. Esto se puede deber a que la frecuencia o el coste exclusivamente de la garantía de RC no se vea reflejada con cualquier cambio en alguna de estas variables. Otra posible explicación es la forma utilizada para separar en tramos o grupos las variables, según los tipos de vehículos en este caso, berlina, compacto, utilitario y resto y según la zona en 6 tipos distintos de zonas. También puede deberse a que alguna de ellas tenga una alta correlación con otra de las variables utilizadas, como puede ser el tipo de vehículo con el valor de este.

3.1. Base de datos final

Partiendo de los datos y variables nombradas anteriormente, estos han sido trabajados y limpiados de modo que quedasen definitivamente una cantidad de 65.535 pólizas, todas ellas de coches, excluyendo motos y vehículos de empresa y utilizadas cada una para el Modelo de Frecuencias.

A su vez, de estas pólizas quedan 4.454 para el modelo de Costes Medios, ya que son todas las que han tenido al menos un siniestro en el año de exposición estudiado y por lo tanto son las únicas que pueden servir a la hora de medir y analizar como se comportan los costes.

Todos los siniestros que forman la base de datos son aquellos definidos como culpables y que han aportado coste a la garantía de RC (responsabilidad civil).

Debido a que todos los registros de siniestros tomados son culpables, cabe destacar que no se han dado costes con importe negativo en nuestra base de datos, ya que estos costes derivados de los siniestros en cuestión han sido indemnizados o reparados por la aseguradora, sin poder de este modo recibir indemnizaciones por parte de otras entidades.

La segmentación de los datos de las variables independientes en grupos o tramos se ha realizado siempre con el fin de facilitar el trabajo con ellas y encontrar una significatividad de cara al estudio. Estas agrupaciones serán detalladas a continuación al explicar a fondo las variables utilizadas. Tanto estas agrupaciones como la limpieza de datos previa se han realizado a través de la herramienta SAS.

3.1.1. Edad

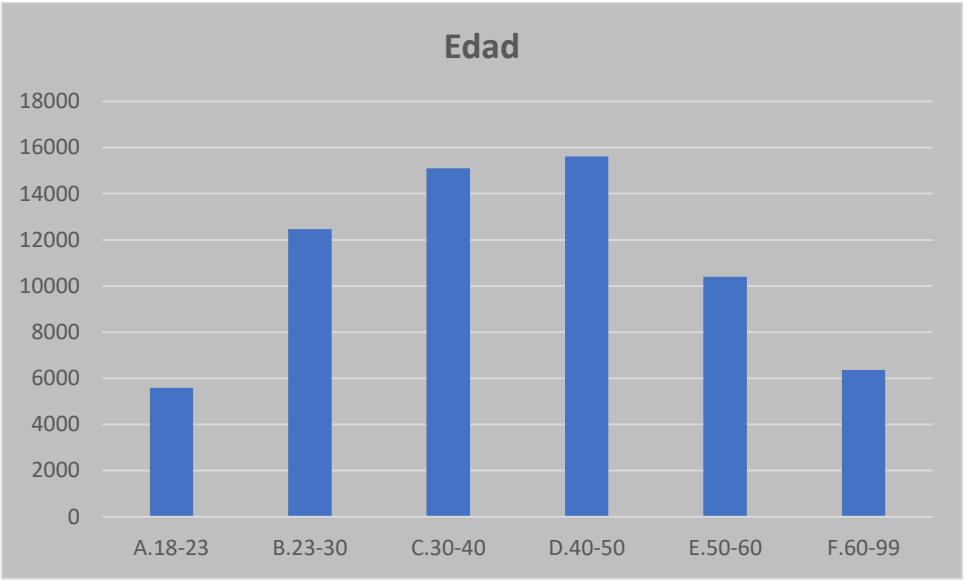
La variable *Edad* en el estudio definitivo está separada en 6 tramos distintos:

- A.18-23: Pólizas en los que el asegurado tiene menos de 23 años (sin incluirlos).
- B.23-30: Pólizas en los que el asegurado tiene más de 23 años (incluyéndolos) y menos de 30 (sin incluirlos).
- C.30-40: Pólizas en los que el asegurado tiene más de 30 años y menos de 40.
- D.40-50: Pólizas en los que el asegurado tiene más de 40 años y menos de 50.
- E.50-60: Pólizas en los que el asegurado tiene más de 50 años y menos de 60.
- F.60-99: Pólizas en los que el asegurado tiene más de 60 años (incluyéndolos).

Para separar en distintos grupos las edades de los asegurados se ha analizado la distribución de la cartera según estos. Agrupándolos cada 10 años se obtienen grupos de un tamaño parecido, exceptuando el de mayores de 60 años, por lo que estos últimos se han reunido en el mismo conjunto.

Cabe destacar que no es lo óptimo mantener el grupo de asegurados de entre 18 y 30 años, a pesar de tener un tamaño parecido al resto de categorías. Esto es así, debido a las diferencias en los cálculos que podían llegar a tener asegurados de 18 o 19 años con los asegurados más experimentados como conductores de este tramo de edades, por lo que se separaron en 2 subgrupos, uno de asegurados entre 18 y 23 años de edad y otro de asegurados entre 23 y 30 años de edad. De esta forma se consigue reunir a los asegurados según sus edades de una forma que facilita los futuros cálculos de la mejor forma posible.

Gráfico 1: Distribución de la cartera por edades de los asegurados



Fuente: Elaboración propia con datos de una entidad aseguradora.

Como se puede observar, inicialmente el grupo más numeroso de asegurados según su edad era el comprendido entre los 18 y 30 años, sin embargo, una vez separada esta categoría, el conjunto más numeroso es el que comprende las edades de 40 y 50 años, con 15.620 asegurados y sus correspondientes pólizas, seguido del C.30-40 que tiene 15.105 asegurados, mientras que el grupo menos numeroso es el de asegurados mayores de 60 años, con 6.366 pólizas.

3.1.2. Antigüedad del vehículo

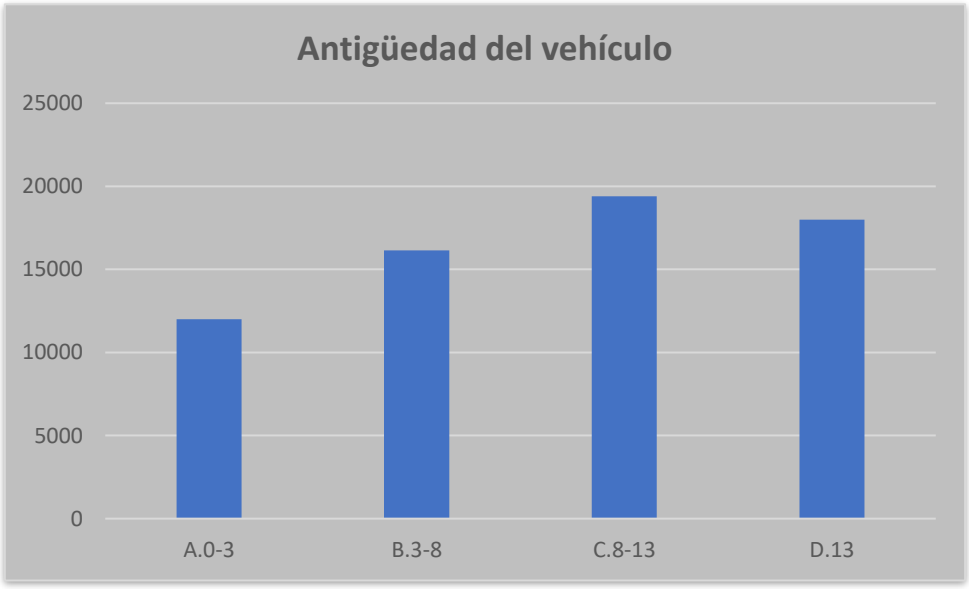
La antigüedad del vehículo recogida en la variable *Vehiaños*, se ha separado en 4 grandes grupos no equitativos ni en cuanto a la edad del vehículo asegurado, ni en cuanto a la magnitud de estos en número de pólizas. Esto es debido a que los vehículos más antiguos son coches con más de 30 años, pero muy poco significativos entre la masa.

Además, se han agrupado de forma que cada una de las partes fraccionadas sea más característica de los costes de sus correspondientes siniestros comparándose con otras formas de segmentar esta variable:

- A.0-3: Vehículos asegurados con una antigüedad igual o inferior a 3 años (hasta 3 años y 1 día).
- B.3-8: Vehículos asegurados con una antigüedad entre los 3 y los 8 años.
- C.8-13: Vehículos asegurados con una antigüedad entre los 8 y los 13 años.
- D.13: Vehículos con una antigüedad mayor de 13 años (a partir de 13 años y 1 día).

Tal y como se cita anteriormente, las divisiones de los grupos no se han realizado intentando que sean lo más equitativas posibles en número, si no que sean lo más representativas posibles de los costes de sus siniestros, por lo que por un lado se separa un conjunto de vehículos nuevos de entre 0 y 3 años que según la teoría los conductores los cuidan más, una gran masa de vehículos que forman la mayor parte de la cartera de entre 3 y 13 años dividida en 2 formando un subgrupo de entre 3 y 8 años y otro de entre 8 y 13 años, y por último, los vehículos más antiguos de la cartera con más de 13 años.

Gráfico 2: Distribución de la cartera por antigüedad del vehículo



Fuente: Elaboración propia con datos de una entidad aseguradora.

Se observa en el Grafico 2 que el grupo menos numeroso es de los vehículos nuevos con un total de 11.996, mientras que la cartera tiene 17.991 pólizas con los más antiguos. El grupo del

estudio más numeroso en la cartera es el de los vehículos comprendidos entre 8 y 13 años de antigüedad con 19.400.

3.1.3. Valor del vehículo

La variable que identifica el valor del vehículo la denominamos *vehivalor,* cabe destacar sobre ella, que está denominada en euros.

El abanico de valores de los vehículos de la cartera es muy amplio, desde 1.800 euros el de menor valor, hasta 190.000 euros el de mayor. Esta gran variedad dificulta la agrupación en tramos de la variable, ya que si se divide equitativamente en número de pólizas encontraremos grupos con mucha diferencia de valores entre unos vehículos y otros, lo que puede distorsionar los cálculos. En cambio, si se divide repartiendo tramos de valores de forma equitativa daría como resultado algún intervalo, sobre todo central, que agrupe demasiados vehículos y entre los que menos y más valor tengan de estos pueda haber diferencias significativas que alteren los resultados.

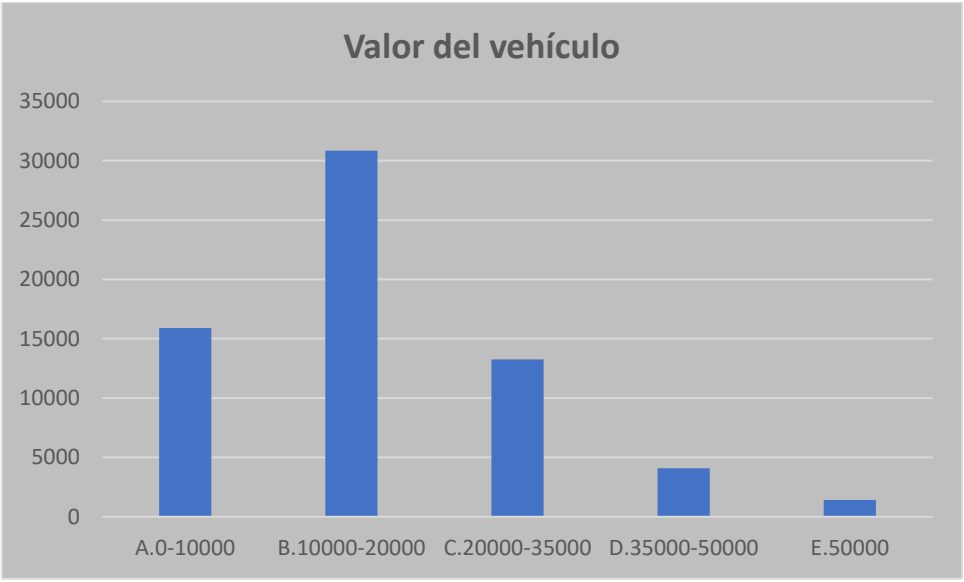
Con todo ello, la segmentación final de la variable es la siguiente:

- A.0-10000: En este grupo se añaden todas las pólizas con vehículos de valor igual o menor de 10.000 euros.
- B.10000-20000: Pólizas con vehículos de valor mayor de 10.000 euros y menores o iguales de 20.000 euros.
- C.20000-35000: Pólizas con vehículos de valor mayor de 20.000 euros y menores o iguales de 35.000 euros.
- D.35000-50000: Pólizas con vehículos de valor mayor de 35.000 euros y menores o iguales de 50.000 euros.
- E.50000: Pólizas con vehículos de valor mayor de 50.000 euros.

La decisión de separar los grupos de vehículos según este criterio se debe a la intención de agruparlos de la forma más homogénea posible, aunque el resultado final nos lleve a categorías con un número mucho mayor que otras.

La distribución de la cartera según el valor de los vehículos quedaría de la siguiente forma:

Gráfico 3: Distribución de la cartera por el valor del vehículo



Fuente: Elaboración propia con datos de una entidad aseguradora.

Tal y como se aprecia en el Grafico 3, la distribución no es para nada equitativa, el grupo más abundante es el que tiene vehículos entre 10.000 y 20.000 euros, que suman 30.842, mientras que los menos numerosos, el comprendido entre 35.000 y 50.000 euros y el de más de 50.000 tienen 4.102 y 1.410 vehículos respectivamente.

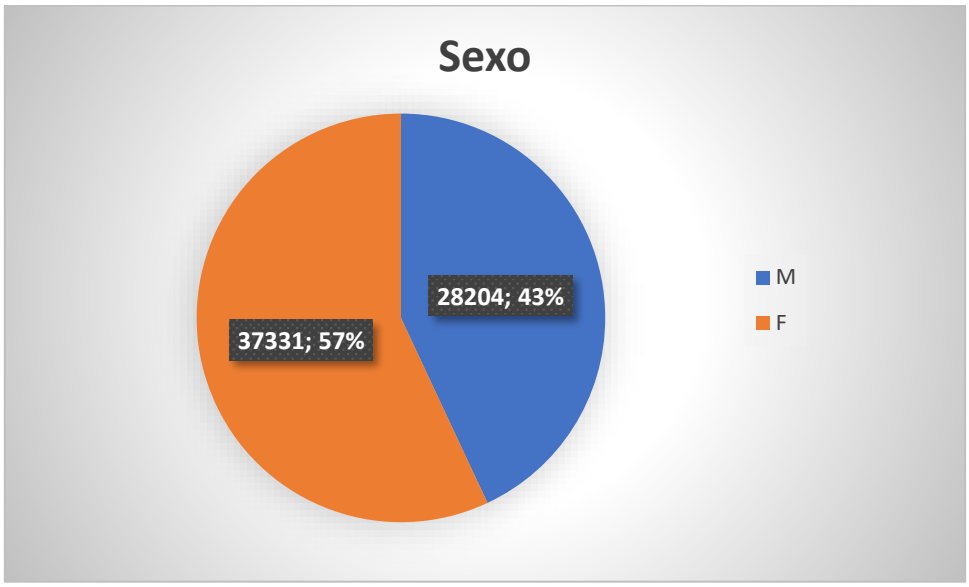
3.1.4. Sexo

En la cartera se puede separar también cada una de las pólizas según el sexo biológico del asegurado, esta variable llamada *sexo*, está separada entre hombres, M y mujeres, F.

Esta variable puede ser muy significativa para el estudio, sin embargo, actualmente la ley prohíbe segmentar el precio del seguro por esta categoría, por lo que posteriormente modelizaremos el comportamiento de las primas de riesgo de ambas formas. Una sin tener en cuenta la variable como dicta la ley, y otra, ya que el estudio es meramente académico, observando los resultados con ella y comprobando si es o no significativa y si afecta a la frecuencia de los siniestros o al coste medio de estos.

A continuación, en el Gráfico 4 se puede observar la distribución de la cartera entre hombres y mujeres:

Gráfico 4: Distribución de la cartera por el sexo del asegurado



Fuente: Elaboración propia con datos de una entidad aseguradora.

Se aprecia, que la cantidad de mujeres aseguradas es ligeramente superior que la de los hombres, un 57% frente a un 43%.

3.1.5. Bonus

La variable *Bonus* es un baremo establecido por las entidades aseguradoras en el que los buenos conductores son premiados mediante importantes descuentos en el precio del seguro, mientras que los malos son castigados con aumentos en el mismo.

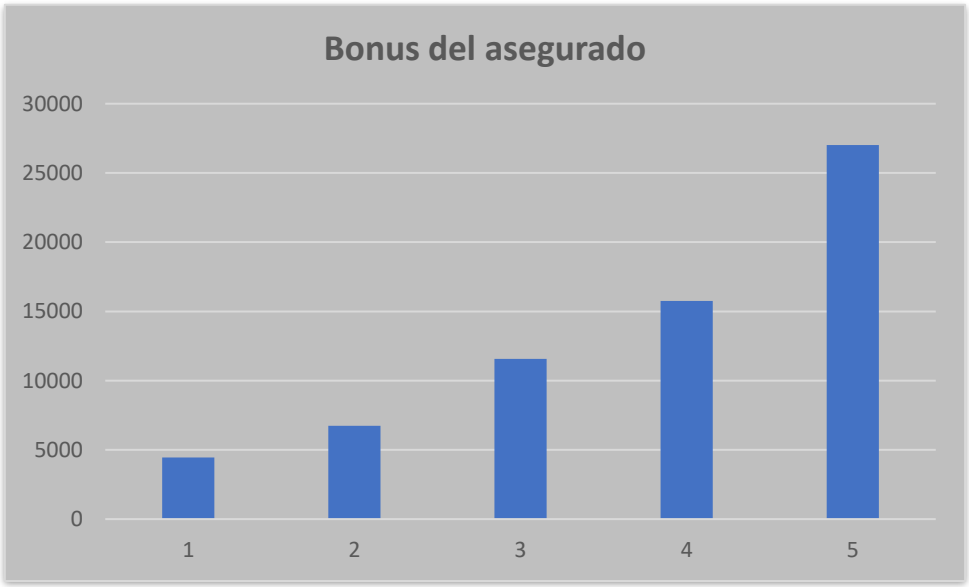
Los datos de siniestralidad son compartidos por la mayoría de las compañías aseguradoras para facilitar el cálculo de la prima y la tarificación de clientes de nuevo negocio, esto es así mediante un fichero informatizado denominado SINCO. Datos con los que las entidades posteriormente realizan ajustes de precio en sus primas, una estrategia importante para una de las principales claves en la actualidad, que es la auditoría de suscripción, imprescindible para tener buenos clientes y, por lo tanto, para los beneficios de la aseguradora.

Como se ha mencionado anteriormente, el valor de la variable está dividido en 5 grupos, siendo el 5 el de los conductores que tienen pocos o ningún siniestro a lo largo del tiempo y el 1 el que corresponde a los peores conductores.

Esta variable en el estudio actual no será utilizada a la hora de calcular las primas de riesgo base, si no que se utilizará posterior al cálculo de esta para realizar un descuento en caso de que el cliente nos interese como asegurado o una sobreprima en caso contrario. Estos porcentajes de descuentos se citarán posteriormente en el estudio.

La distribución de la cartera según el bonus de cada una de las pólizas es el siguiente:

Gráfico 5: Distribución de la cartera por el bonus del asegurado



Fuente: Elaboración propia con datos de una entidad aseguradora.

En la distribución correspondiente, el grupo más numeroso es el de los buenos conductores con bonus 5, que tiene un total de 27.020 pólizas, mientras que los peores conductores con un nivel de bonus de 1 son 4.454, un número de pólizas mucho menor, pero que corresponde a asegurados que tienen constantemente siniestros, por lo que la penalización que conllevará en el cálculo de la prima definitiva deberá ser significativo, evitando así este tipo de conductores. De igual forma, para intentar mantener el grueso de pólizas con asegurados que tienen una buena siniestralidad, se aplicará una bonificación o descuento sobre el cálculo de la prima de riesgo base.

4. Modelo de frecuencias

Una vez preparada la base de datos y agrupados los valores de las variables aptas para ser utilizadas en el modelo se comienza con el modelo de frecuencias. Para este, se utiliza la técnica GLM explicada brevemente en el comienzo del estudio, realizando los correspondientes cálculos en el programa estadístico RStudio para explicar el comportamiento de la variable dependiente *Siniestro*.

Para comparar los resultados de ambos, se ha decidido realizar dos modelos distintos, uno con más variables que el otro, ambos a partir de la base a datos final de 65.535 pólizas. En el GLM utilizado para los modelos de frecuencias daremos por hecho que el número de siniestros se distribuye como una función de distribución Poisson, ya que es la adecuada para variables de respuesta de conteos de acontecimientos como ocurre en este caso.

La función de enlace canónica para la distribución de Poisson en los GLM es la función *log*, por lo que será la utilizada para el cálculo de estos modelos.

Se añadirán variables independientes para modelizar la frecuencia y mediante los análisis y contrastes mencionados anteriormente, como la comparación de la Deviance y el AIC, obtendremos el modelo final con las variables optimas que expliquen el comportamiento de la frecuencia de los siniestros.

Antes de realizar los contrastes entre los modelos con diferentes variables para ver cual explica mejor el comportamiento de la frecuencia, es importante comprobar que las variables explicativas sean significativas dentro del modelo, ya que en caso contrario este no sería de utilidad. También hay que tener en cuenta que añadir más variables al modelo no tiene por qué hacerlo mejor, esto se puede ver posteriormente, ya que muchos de los modelos observados de 2 variables son mejores modelos que los de 3, al igual que algunos de 3 explican mejor el comportamiento de la variable dependiente que un modelo con las 4 variables independientes que tenemos, *edad, vehiaños, sexo y vehivalor,* ya que, por ejemplo en el caso de este último modelo, la variable *sexo* no es significativa para explicar la variable dependiente.

Los 2 modelos de frecuencias definitivos para el cálculo posterior de la prima de riesgo son un modelo de 2 variables y otro de 3, por lo que a continuación, se detallarán cada uno de ellos y se explicará el motivo por el cual se eligen unas variables u otras en cada caso.

Los resultados de las salidas expuestas en todos los modelos del estudio de este y los siguientes apartados serán obtenidas de la herramienta RStudio, además, todos los códigos utilizados en el programa serán expuestos en el apartado Anexos al final del trabajo.

4.1. Modelo de frecuencias de 2 variables

El modelo de frecuencias de 2 variables elegido para el cálculo de la prima de riesgo se ha determinado de entre todos ellos observando aquel que más reduce la Deviance y menor AIC tiene, siempre y cuando, las variables sean estadísticamente significativas, es decir, que su p-valor sea lo suficientemente pequeño como para considerar que la variable independiente explica el comportamiento de la variable dependiente. Sabiendo esto, el modelo óptimo de 2 variables sería el siguiente:

```
> MFQ3=glm(siniestro~edad+vehivalor,family=poisson(link="log"),data=Cartera)
> summary(MFQ3)

Call:
glm(formula = siniestro ~ edad + vehivalor, family = poisson(link = "log"),
    data = Cartera)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4411  -0.3821  -0.3720  -0.3419   2.0356

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.61707    0.05479 -47.768  < 2e-16 ***
edadB.23-30           -0.17229    0.05667  -3.040 0.002364 **
edadC.30-40           -0.19597    0.05513  -3.555 0.000378 ***
edadD.40-50           -0.21416    0.05507  -3.889 0.000101 ***
edadE.50-60           -0.40622    0.06161  -6.593 4.30e-11 ***
edadF.60-99           -0.40244    0.07020  -5.733 9.90e-09 ***
vehivalorB.10000-20000  0.16051    0.03955   4.059 4.93e-05 ***
vehivalorC.20000-35000  0.25582    0.04577   5.590 2.28e-08 ***
vehivalorD.35000-50000  0.28692    0.06497   4.416 1.00e-05 ***
vehivalorE.50000        0.18344    0.10585   1.733 0.083098 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 23952  on 65534  degrees of freedom
Residual deviance: 23850  on 65525  degrees of freedom
AIC: 32778

Number of Fisher Scoring iterations: 6
```

Observando los diferentes modelos, se aprecia que la variable que se debe tomar en primer lugar en el modelo es el grupo de edad de los asegurados, mientras que en segundo lugar el valor del vehículo es de la que se obtiene mejor resultado.

Como puede verse en la salida de RStudio, los p-valores asociados a todas las variables son mucho menores del 5%. Los grupos de variables incluidos en el modelo sin ser significativos lo están por formar parte de una variable que en su totalidad sí que lo es, ya que alguno de los grupos que la componen son significativos, como así ocurre con la variable del valor del vehículo.

Otro buen modelo es el que tiene como variables explicativas la edad del asegurado y la antigüedad del vehículo, sin embargo, como se aprecia a continuación no lo es tanto como el mostrado anteriormente:

```
> MFQ1=glm(siniestro~edad+vehiaños,family=poisson(link="log"),data=Cartera)
> summary(MFQ1)

Call:
glm(formula = siniestro ~ edad + vehiaños, family = poisson(link = "log"),
    data = Cartera)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4346  -0.3775  -0.3676  -0.3526   2.0085

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -2.478811   0.054920 -45.135  < 2e-16 ***
edadB.23-30    -0.162793   0.056699  -2.871 0.004089 **
edadC.30-40    -0.185522   0.055154  -3.364 0.000769 ***
edadD.40-50    -0.214072   0.055092  -3.886 0.000102 ***
edadE.50-60    -0.404223   0.061623  -6.560 5.39e-11 ***
edadF.60-99    -0.418414   0.070153  -5.964 2.46e-09 ***
vehiañosB.3-8   0.119013   0.045355   2.624 0.008689 **
vehiañosC.8-13 -0.001948   0.044815  -0.043 0.965333
vehiañosD.13   -0.068235   0.046192  -1.477 0.139621
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 23952  on 65534  degrees of freedom
Residual deviance: 23867  on 65526  degrees of freedom
AIC: 32793

Number of Fisher Scoring iterations: 6
```

Para comprobar que el modelo de 2 variables inicial es el óptimo se puede realizar una comparativa tanto de la Deviance como del AIC:

```
> AIC(MFQ3)
[1] 32778.33
> AIC(MFQ1)
[1] 32793.43
> anova(MFQ3,MFQ1,test="Chisq")
Analysis of Deviance Table

Model 1: siniestro ~ edad + vehivalor
Model 2: siniestro ~ edad + vehiaños
  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1     65525      23850
2     65526      23867 -1  -17.092 3.562e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Tal y como se aprecia en la salida de RStudio, el AIC del modelo con las variables *edad* + *vehivalor* es menor que el de las variables *edad* + *vehiaños*, 32.778 frente a 32.793, además de que la Deviance de este modelo también es menor al segundo, por lo que podemos asumir que el primer modelo es el óptimo.

Una vez confirmado el modelo de frecuencias para el cálculo de la prima de riesgo, es importante exportar las *rates* de este a Excel y así facilitar la comprensión de los resultados de las variables.

En la tabla siguiente se muestran los resultados exportados del cálculo de frecuencia, según el modelo utilizado y sus variables:

Tabla 2: Rates del modelo de frecuencias de 2 variables

Variable	Rates
(Intercept)	7,30%
A.18-23	1,0000
B.23-30	0,8417
C.30-40	0,8220
D.40-50	0,8072
E.50-60	0,6662
F.60-99	0,6687
A.0-10000	1,0000
B.10000-20000	1,1741
C.20000-35000	1,2915
D.35000-50000	1,3323
E.50000	1,2013

Fuente: Elaboración propia.

Se aprecia que la constante de la que parte el modelo es un 7.30%, lo que significa que un asegurado de entre 18 y 23 años y con un vehículo valorado entre 0 y 10.000 euros tiene un 7.30% de posibilidades de media, de tener al menos un siniestro.

Observando la variable edad, el ser más joven supone un coeficiente multiplicativo mayor, por lo que cuanto más joven, según nuestros resultados, mayor tendencia a tener algún siniestro. En cuanto al valor del vehículo el grupo que mayor tendencia tiene son los comprendidos entre 35.000 y 50.000 euros con un coeficiente multiplicativo de 1,3323 mientras que los que menos tendencia tienen son los vehículos valorados en menos de 10.000 euros.

4.2. Modelo de frecuencias de 3 variables

Siguiendo los mismos pasos que con el modelo de 2 variables podemos desarrollar el modelo óptimo de frecuencias de 3 variables:

```
> MFQ8=glm(siniestro~edad+vehiaños+vehivalor,family=poisson(link="log"),data=Cartera)
> summary(MFQ8)

Call:
glm(formula = siniestro ~ edad + vehiaños + vehivalor, family = poisson(link = "log"),
    data = Cartera)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.4571  -0.3838  -0.3692  -0.3436   2.0428

Coefficients:
                      Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.70389    0.07287 -37.105  < 2e-16 ***
edadB.23-30           -0.17693    0.05677  -3.116 0.001831 **
edadC.30-40           -0.19995    0.05524  -3.620 0.000295 ***
edadD.40-50           -0.21665    0.05510  -3.932 8.42e-05 ***
edadE.50-60           -0.40893    0.06164  -6.634 3.26e-11 ***
edadF.60-99           -0.40476    0.07021  -5.765 8.16e-09 ***
vehiañosB.3-8          0.14287    0.04635   3.082 0.002053 **
vehiañosC.8-13         0.07426    0.04802   1.546 0.122005
vehiañosD.13           0.09546    0.05647   1.691 0.090927 .
vehivalorB.10000-20000 0.16049    0.04596   3.492 0.000480 ***
vehivalorC.20000-35000 0.27239    0.05558   4.901 9.56e-07 ***
vehivalorD.35000-50000 0.30228    0.07529   4.015 5.94e-05 ***
vehivalorE.50000       0.23570    0.11394   2.069 0.038573 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 23952  on 65534  degrees of freedom
Residual deviance: 23841  on 65522  degrees of freedom
AIC: 32775

Number of Fisher Scoring iterations: 6
```

En este caso, al modelo óptimo de 2 variables con *edad* + *vehivalor* se le añade la variable *vehiaños,* de esta forma se obtiene el modelo óptimo, ya que otro de los modelos que da buen resultado es aquel formado por las variables *edad* + *vehivalor* + *sexo,* sin embargo, este como se puede apreciar a continuación no es tan bueno como el primero.

```
> AIC(MFQ8)
[1] 32774.51
> AIC(MFQ9)
[1] 32779.83
> anova(MFQ8,MFQ9,test="Chisq")
Analysis of Deviance Table

Model 1: siniestro ~ edad + vehiaños + vehivalor
Model 2: siniestro ~ edad + sexo + vehivalor
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1     65522      23841
2     65524      23850 -2  -9.3121 0.009504 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Con un menor AIC y una menor Deviance podemos asumir que el modelo óptimo de 3 variables es el formado por las variables *edad* + *vehivalor* + *vehiaños,* además, comparando con los modelos anteriores de 2 variables, también tiene mejores resultados que el formado por las variables *edad* + *vehivalor.*

Una vez confirmado el modelo de frecuencias de 3 variables para el cálculo de la prima de riesgo, tal y como se ha realizado anteriormente es importante exportar las *rates* a Excel y exponer los resultados de las variables.

En este caso la constante de la que parte el modelo es un 6.69%, lo que significa que un asegurado de entre 18 y 23 años y con un vehículo valorado entre 0 y 10.000 euros y con menos de 3 años de antigüedad tiene un 6.69% de posibilidades de media de tener al menos un siniestro.

En la tabla siguiente se muestran los resultados exportados:

Tabla 3: Rates del modelo de frecuencias de 3 variables

Variables	Rates
(Intercept)	6,69%
A.18-23	1,0000
B.23-30	0,8378
C.30-40	0,8188
D.40-50	0,8052
E.50-60	0,6644
F.60-99	0,6671
A.0-3	1,0000
B.3-8	1,1536
C.8-13	1,0771
D.13	1,1002
A.0-10000	1,0000
B.10000-20000	1,1741
C.20000-35000	1,3131
D.35000-50000	1,3529
E.50000	1,2658

Fuente: Elaboración propia.

Al igual que para el modelo anterior, la edad influye de la misma forma en el comportamiento de la variable dependiente, teniendo más tendencia a ocasionar un siniestro cuanto más joven sea el asegurado, ya que el tramo de edad entre los 18 y los 23 años es el que mayor coeficiente multiplicativo tiene, mientras que el tramo que menos, es el formado por los asegurados de entre 50 y 60 años.

Por otro lado, se observa que los vehículos con más tendencia a tener siniestros son los valorados entre los 35.000 y 50.000 euros y los que menos los valorados en menos de 10.000 euros.

También se aprecia según la antigüedad del vehículo que los nuevos, aquellos con menos de 3 años de antigüedad son los que menor número de accidentes tienen, mientras que los vehículos comprendidos entre los 3 y los 8 años son los que más.

5. Modelo de costes medios

Al igual que en los modelos de frecuencias, la herramienta utilizada para los de costes medios será RStudio, con la que se tratará de explicar el comportamiento de la variable dependente *coste* en función de aquellas variables que más información aporten al modelo.

La base de datos utilizada para ello no es la misma que la utilizada en el modelo de frecuencias como se ha mencionado, si no que para este apartado se utilizará la de 4.454 pólizas de la cartera que hayan tenido 1 o más siniestros y por lo tanto tengan un coste, para así poder modelizarlo.

En el GLM utilizado para los modelos de costes medios daremos por hecho que estos costes se distribuyen como una función de distribución Gamma, ya que es la distribución adecuada para modelizar el comportamiento de variables aleatorias continuas con una asimetría positiva como ocurre en este caso.

La función de enlace canónica para la distribución gamma es la función inversa, sin embargo, debido a que los parámetros de este tipo de función son difíciles de interpretar, la función *log* se considera más útil, sobre todo en el ámbito de seguros de vehículos, tal y como explican en su obra, *Generalized Linear Models for Insurance Data,* De Jong, P. y Heller, G. Z**.** Por lo que para los siguientes modelos se utilizará como función Link la función *log*.

Se añadirán las variables independientes que puedan ser útiles y mediante los correspondientes contrastes y análisis, comparando la Deviance y el AIC obtendremos los modelos finales con las variables óptimas que mejor expliquen el comportamiento de los costes medios.

Para mejorar el modelo también se ha comparado entre meter más o menos variables, sin embargo, los resultados son muy parecidos a los de las frecuencias, por lo que introducir un número mayor de ellas no tiene necesariamente por qué proporcionarnos un mejor modelo. En este caso, el modelo de 4 variables tiene 2 variables que no resultan significativas para explicar el comportamiento de los costes, *vehivalor* y *vehiaños*, además de que tiene un AIC mayor que los modelos de 3 variables que mejor explican los costes.

Sabiendo esto, de forma final se han realizado 2 modelos de costes medios, uno de ellos con 2 variables y el otro con 3.

5.1. Modelo de costes medios de 2 variables

Al igual que en el modelo de frecuencias, la variable *edad* es la primera que debe considerarse en el de costes medios, ya que es la que mejor explica el comportamiento de estos. En cuanto a la segunda, la variable *sexo* es la que mejores resultados muestra, sin embargo, esta será añadida en el modelo de 3 variables para diferenciar el cálculo de la prima de riesgo por un lado teniendo en cuenta dicha variable, y por otro sin ella, utilizando en este último caso para ello el modelo de 2 variables de frecuencias y de costes medios.

Sabiendo esto, la variable elegida para el modelo será *vehiaños*, comparando también los resultados con la variable *vehivalor* para confirmar que el modelo elegido es el óptimo.

```
> MCM1=glm(coste~edad+vehiaños,family=Gamma(link="log"),data=Coste)
> summary(MCM1)

Call:
glm(formula = coste ~ edad + vehiaños, family = Gamma(link = "log"),
    data = Coste)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8698  -1.3427  -0.8103   0.0550   6.7701

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     7.78576    0.09718  80.113  < 2e-16 ***
edadB.23-30    -0.21377    0.09981  -2.142 0.032265 *
edadC.30-40    -0.32321    0.09714  -3.327 0.000884 ***
edadD.40-50    -0.30273    0.09711  -3.117 0.001837 **
edadE.50-60    -0.47357    0.10851  -4.364  1.3e-05 ***
edadF.60-99    -0.32479    0.12349  -2.630 0.008568 **
vehiañosB.3-8   0.05573    0.07988   0.698 0.485428
vehiañosC.8-13  0.07483    0.07896   0.948 0.343336
vehiañosD.13    0.19212    0.08136   2.361 0.018256 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 3.101403)

    Null deviance: 7103.3  on 4453  degrees of freedom
Residual deviance: 7019.9  on 4445  degrees of freedom
AIC: 76380

Number of Fisher Scoring iterations: 6
```

Tal y como se explicaba anteriormente, las categorías de *vehiaños* que están en el modelo sin ser significativas se mantienen debido a que forman parte de una variable que en su conjunto sí que lo es.

```
> AIC(MCM1)
[1] 76379.55
> AIC(MCM3)
[1] 76386.59
> anova(MCM1,MCM3,test="Chisq")
Analysis of Deviance Table

Model 1: coste ~ edad + vehiaños
Model 2: coste ~ edad + vehivalor
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4445     7019.9
2      4444     7026.4  1  -6.4833
```

En la comparación de los modelos realizada mediante el test ANOVA se puede apreciar que el modelo 1, el formado por las variables explicativas *edad* y *vehiaños* explica mejor el comportamiento de la variable dependiente ya que su Deviance es menor, además de que observando al AIC de cada uno de los modelos también se aprecian diferencias, por lo que podemos asumir que este modelo es el óptimo.

Una vez confirmado el modelo para el cálculo de los costes, se muestran los resultados exportados al programa Excel.

Tabla 4: Rates del modelo de costes medios de 2 variables

Variables	Rates
(Intercept)	2.406,08
A.18-23	1,0000
B.23-30	0,8075
C.30-40	0,7238
D.40-50	0,7388
E.50-60	0,6228
F.60-99	0,7227
A.0-3	1,0000
B.3-8	1,0573
C.8-13	1,0777
D.13	1,2118

Fuente: Elaboración propia.

En la Tabla 4 se puede apreciar que cuanto más joven es el asegurado, mayores son los costes de los siniestros, mientras que el rango de edades donde menores son los costes de estos es entre los 50 y los 60 años de edad.

Por otro lado, teniendo en cuenta la antigüedad del vehículo, los coches nuevos con menos de 3 años tienen un coste menor en sus siniestros que el resto, mientras que los más antiguos son los más costosos de cubrir de media. Esto podría ser explicado principalmente por los nuevos sistemas de seguridad implementados en los vehículos actuales, lo que puede ayudar a reducir el impacto de estos siniestros.

5.2. Modelo de costes medios de 3 variables

Siguiendo los pasos de los análisis previos, al modelo de 2 variables con *edad* y *vehiaños* se le añade la variable *sexo* para formar el modelo de costes medios óptimo de 3 variables, ya que explica bastante bien el comportamiento de la variable dependiente como se puede apreciar a continuación.

```
> MCM7=glm(coste~edad+vehiaños+sexo,family=Gamma(link="log"),data=Coste)
> summary(MCM7)

Call:
glm(formula = coste ~ edad + vehiaños + sexo, family = Gamma(link = "log"),
    data = Coste)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9009  -1.3571  -0.8103   0.0690   6.9583

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    7.69535    0.09868  77.982  < 2e-16 ***
edadB.23-30   -0.19866    0.09853  -2.016  0.04383 *
edadC.30-40   -0.29197    0.09596  -3.043  0.00236 **
edadD.40-50   -0.28321    0.09587  -2.954  0.00315 **
edadE.50-60   -0.45309    0.10712  -4.230 2.39e-05 ***
edadF.60-99   -0.32639    0.12194  -2.677  0.00746 **
vehiañosB.3-8  0.06305    0.07887   0.799  0.42406
vehiañosC.8-13 0.07711    0.07794   0.989  0.32251
vehiañosD.13   0.18389    0.08038   2.288  0.02220 *
sexoM          0.16131    0.05293   3.048  0.00232 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 3.021865)

    Null deviance: 7103.3  on 4453  degrees of freedom
Residual deviance: 6992.0  on 4444  degrees of freedom
AIC: 76360

Number of Fisher Scoring iterations: 6
```

Analizando el modelo, se observa que la variable *sexo* es significativa y el modelo tiene unos resultados bastante buenos, lo que podemos confirmar comparándolo con otro de los que mejores resultados tiene.

```
> AIC(MCM7)
[1] 76359.79
> AIC(MCM9)
[1] 76361.3
> anova(MCM7,MCM9,test="Chisq")
Analysis of Deviance Table

Model 1: coste ~ edad + vehiaños + sexo
Model 2: coste ~ edad + sexo + vehivalor
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1      4444     6992.0
2      4443     6991.3  1  0.62594   0.6464
```

Viendo la salida de RStudio del Test ANOVA, se puede ver que ambos modelos, tanto el escogido para el análisis del coste medio mostrado anteriormente, como la opción con las variables *edad + sexo + vehivalor* tienen una Deviance muy similar, por lo que para decidir el modelo utilizado para el posterior cálculo de la prima de riesgo se escoge aquel con el AIC menor, en este caso el modelo con las variables *edad + vehiaños + sexo.*

Con el modelo de costes medios de 3 variables óptimo, se exportan los datos a Excel, mostrándose las rates de las variables y sus categorías en la Tabla 5.

Tabla 5: Rates del modelo de costes medios de 3 variables

Variables	Rates
(Intercept)	2.198,10
A.18-23	1,0000
B.23-30	0,8198
C.30-40	0,7468
D.40-50	0,7534
E.50-60	0,6357
F.60-99	0,7215
A.0-3	1,0000
B.3-8	1,0651
C.8-13	1,0802
D.13	1,2019
F	1,0000
M	1,1751

Fuente: Elaboración propia.

La edad del asegurado y la antigüedad del vehículo se comportan de igual forma que en el modelo de 2 variables de costes medios, los grupos que menos coste tienen en sus siniestros son los asegurados de entre 50 y 60 años y los vehículos nuevos de menos de 3 años, y a su vez, los que más tienen son los jóvenes de menos de 23 años y los coches de más de 12.

En cuanto al sexo, es interesante ver que además de ser una variable significativa para analizar el comportamiento de la variable dependiente, se aprecia como los hombres de media tienen un coste en sus siniestros mayor que las mujeres, por lo que a pesar de ser una variable que no se puede tener en cuenta a la hora de tarificar, es atrayente ver el efecto que puede tener en el cálculo de la prima de riesgo y el posterior coste final del seguro.

6. Prima de riesgo y tarifa final

El cálculo para obtener la prima de riesgo del seguro de automóviles de una póliza es el producto de la frecuencia y el coste medio de los siniestros de esta:

$$PR = FQ * CM$$

En el estudio actual se han elegido dos modelos para cada una de estas variables dependientes, por lo que en este caso se han obtenido dos primas de riesgo, una mediante el producto de la frecuencia y el coste medio de los modelos de 2 variables y otra mediante el producto de la frecuencia y el coste medio de los modelos de 3 variables.

En el primero de los casos la prima de riesgo depende de 3 variables, *edad, vehiaños y vehivalor*, ya que el modelo de frecuencias se compone de *edad + vehivalor* y el de costes medios de *edad + vehiaños*.

En el segundo de los casos la prima de riesgo depende de 4 variables, *edad, sexo, vehiaños y vehivalor*, ya que el modelo de frecuencias se compone de *edad + vehiaños + vehivalor* y el de costes medios de *edad + vehiaños + sexo.* En este cálculo se puede estudiar de forma académica el efecto del sexo del asegurado en la prima de riesgo, lo que puede ser bastante interesante, ya que como se ha explicado anteriormente, está prohibido tenerlo en cuenta para tarificar cualquier tipo de seguro.

6.1. Prima de riesgo bonificada

Aplicándole a la prima de riesgo un descuento o sobreprima dependiendo de la bonificación del asegurado se obtiene la prima de riesgo bonificada. La variable *bonus* nos indica los asegurados que más o que menos siniestros tienen históricamente según el valor de esta variable sea mas bajo o más alto respectivamente.

Desde el punto de vista de una entidad aseguradora, la mejor opción es aplicar una sanción en forma de sobreprima a los conductores con más siniestros mientras que se premia con descuentos a los que menos tengan, sin embargo, en la realidad, cada empresa actúa según su propia experiencia, ya que algunas no aplican descuentos y se guardan ese margen para negociar

en el caso de que los clientes estén descontentos o quieran contratar el seguro con otra entidad por un motivo económico. También sucede que en determinadas compañías una mala bonificación supone la no renovación del seguro al término de este. Cada aseguradora lo gestiona según cree oportuno, sin embargo, en este caso a la hora de tarificar se ha decidido recurrir a un descuento o sobreprima de forma directa sobre el precio del seguro.

Anteriormente se ha adjuntado en el Gráfico 5 la distribución de la cartera según el *bonus*, mostrándose que la mayoría de las pólizas tienen una buena bonificación. El nivel de bonus 5 está por encima del 40% y si juntamos los bonus 4 y 5 por encima del 65%, por lo que se entiende que lo normal es tener una bonificación alta. Estos datos dan a entender que aplicar un descuento muy grande a todos los asegurados con buena bonificación puede suponer un coste muy alto, por lo que lo importante será penalizar en gran medida los peores casos para compensar los pequeños descuentos que se realizarán a la gran masa de buenos conductores con un bonus alto.

En la siguiente tabla se puede observar tanto la proporción de las pólizas de la cartera según el bonus como la bonificación a aplicar según este.

Tabla 6: Distribución por bonus y su correspondiente bonificación

Coeficientes Bonificación			
BONUS	Polizas	%	Bonificación
1	4.454	6,80%	1,5
2	6.735	10,28%	1,15
3	11.565	17,65%	1
4	15.761	24,05%	0,95
5	27.020	41,23%	0,85

Fuente: Elaboración propia

En el caso de los mejores conductores la prima de riesgo se multiplica por 0,85 para obtener la prima de riesgo bonificada, mientras que los que tienen un bonus de 4 por 0,95. Los asegurados con un bonus de 3, calificado como neutro, mantienen el valor de la prima de riesgo, mientras que a los peores conductores con 2 y 1 de bonus se les aplican sobreprimas en el precio del seguro, los primeros verían multiplicado el valor en 1,15 mientras que los segundos en 1,5.

6.2. Tarifa final

Una vez obtenida la prima de riesgo bonificada multiplicando la frecuencia por el coste medio y aplicándole la correspondiente bonificación, el paso siguiente para llegar a la tarifa final del seguro es elaborar la prima de inventario.

En este caso la prima de inventario se compone de la prima de riesgo bonificada, el coeficiente de seguridad y los gastos de gestión interna (G.I.).

El coeficiente de seguridad es un porcentaje del valor de la prima de riesgo bonificada que se suma a este como margen para evitar riesgos, garantizando que el seguro sea solvente y la entidad tenga la capacidad adecuada incluso en casos extremos de cumplir sus obligaciones con sus clientes.

Dentro de los gastos de gestión interna de la compañía se sitúan los sueldos de los trabajadores, los gastos en equipos, material necesario, alquiler o compra de oficinas, seguridad social, etc. En este caso también toma como valor un pequeño porcentaje del total de la prima de riesgo bonificada.

En este caso, los porcentajes adecuados para cada uno de estos valores han sido situar el coeficiente de seguridad en el 5% y los gastos de gestión interna en el 10%.

Con la prima de inventario obtenida a partir de la suma de los valores antes descritos, el último paso es disponer de la prima comercial, el valor con el que se comercializa el seguro al público y por lo tanto la tarifa final de este.

La prima comercial está formada por la prima de inventario y el gasto comercial, que en este caso supondrá un porcentaje del total de esta. Este gasto lo forman las comisiones que la entidad aseguradora paga a los comerciales o empresas contratadas para la labor de comercializar el seguro. En el caso de este estudio este porcentaje se sitúa en el 10%.

Con la tarifa final ya planteada, se desarrolla una calculadora a partir de las rates conseguidas de los modelos de frecuencias y costes medios y la bonificación aplicada posteriormente, a la que sumándole primero el coeficiente de seguridad y los gastos internos y posteriormente los gastos comerciales se obtiene la tarifa final.

La calculadora está diseñada en un fichero Excel con el nombre de **Calculadora tarifa RC.xlsx** tanto para el cálculo de la prima que depende de 3 variables como para la de 4. En este fichero

también se muestran los resultados de los modelos ya exportados a Excel además de la prima de riesgo que tendrían todas y cada una de las pólizas de la cartera definitiva que ha servido en este estudio.

La calculadora sirve tanto para la renovación de clientes como para clientes de nuevo negocio, sin embargo, para estos últimos será necesario obtener los datos del asegurado referentes a las variables del estudio, para poder tarificar de forma adecuada, lo que no debería ser un inconveniente en la mayoría de los casos.

Uno de los posibles problemas de esto es la bonificación, ya que algunas compañías no comparten con otras los datos de siniestralidad de sus asegurados y esto complica el proceso, por lo que en los casos en los que se pueda obtener el dato de bonus de los asegurados de nuevo negocio se aplicará la bonificación correspondiente y en caso contrario, se actúa como si el asegurado tuviese un nivel de bonus neutro, igual a 3, ya que de esta forma se mantiene la prima de riesgo inicial.

A continuación, se pueden ver algunos ejemplos de la calculadora diseñada en Excel comparando en la Tabla 7 los casos más baratos de seguro y en la Tabla 8 los más caros, observando la prima calculada a partir de 3 variables y la calculada a partir de 4.

Tabla 7: Comparativa de las primas de los seguros más baratos

EDAD	**E.50-60**
ANTIGÜEDAD DEL VEHÍCULO	**A.0-3**
VALOR DEL VEHÍCULO	**A.0-10000**
BONUS	**5**
Prima de riesgo	**72,89 €**
Prima de riesgo bonificada	**61,95 €**
Coeficiente de seguridad (5%)	**3,10 €**
Gastos de G.I. (10%)	**6,20 €**
Prima de inventario	**71,25 €**
Gasto comercial (10%)	**7,12 €**
Prima comercial	**78,37 €**

EDAD	**E.50-60**
ANTIGÜEDAD DEL VEHÍCULO	**A.0-3**
VALOR DEL VEHÍCULO	**A.0-10000**
SEXO	**F**
BONUS	**5**
Prima de riesgo	**62,14 €**
Prima de riesgo bonificada	**52,82 €**
Coeficiente de seguridad (5%)	**2,64 €**
Gastos de G.I. (10%)	**5,28 €**
Prima de inventario	**60,74 €**
Gasto comercial (10%)	**6,07 €**
Prima comercial	**66,82 €**

Fuente: Elaboración propia

Tabla 8: Comparativa de las primas de los seguros más caros

EDAD	A.18-23
ANTIGÜEDAD DEL VEHÍCULO	D.13
VALOR DEL VEHÍCULO	D.35000-50000
BONUS	1

Prima de riesgo	283,65 €
Prima de riesgo bonificada	425,47 €
Coeficiente de seguridad (5%)	21,27 €
Gastos de G.I. (10%)	42,55 €
Prima de inventario	489,29 €
Gasto comercial (10%)	48,93 €
Prima comercial	538,22 €

EDAD	A.18-23
ANTIGÜEDAD DEL VEHÍCULO	D.13
VALOR DEL VEHÍCULO	D.35000-50000
SEXO	M
BONUS	1

Prima de riesgo	309,33 €
Prima de riesgo bonificada	463,99 €
Coeficiente de seguridad (5%)	23,20 €
Gastos de G.I. (10%)	46,40 €
Prima de inventario	533,59 €
Gasto comercial (10%)	53,36 €
Prima comercial	586,95 €

Fuente: Elaboración propia

Como se puede apreciar en la tabla 7 de la prima que depende de 4 variables, se obtiene un precio final del seguro mucho menor que el de la prima que depende de 3.

También se observa que en el caso opuesto de la Tabla 8, con los seguros más caros sucede al contrario, siendo la prima comercial bastante más alta teniendo en cuenta la variable sexo.

Analizando las tablas anteriores, se puede afirmar que, en el caso de este estudio, meter una variable más para el cálculo de la prima de riesgo supone en los casos más baratos una bajada del precio de la tarifa y en los casos más caros un aumento de esta.

También se puede afirmar que la variable *sexo* influye bastante en la prima del seguro, siendo la tarifa de las mujeres bastante más barata que la de los hombres.

Además, queda mencionar que la prima de riesgo formada a partir de 4 variables es preferible a la de 3, ya que tanto el modelo de frecuencias como el de costes medios desarrollados para ella tienen mejor deviance y AIC que los desarrollados para la prima de riesgo que depende de 3 variables.

6. Conclusiones

Se confirma a partir de los datos analizados la relevancia de algunas variables a la hora de tarificar, ya que son bastante significativas explicando el comportamiento de la frecuencia o de los costes medios de los siniestros.

Cabe destacar principalmente el efecto de la variable *edad*, que ha sido la más determinante para el estudio.

También se han obtenido buenas conclusiones de la variable *sexo*, prohibida para tarificar en cualquier seguro, sea de Vida o No Vida, pero que ha resultado ser útil a la hora de desarrollar el precio de la garantía de responsabilidad civil en este caso. Incluso se podría llegar a afirmar que, en caso de no poseer tal prohibición, se tendría en cuenta para el cálculo de la prima en la mayoría de los tipos de seguros.

Observando los resultados finales, se aprecia que las primas comerciales se acercan bastante a la realidad, sin embargo, es difícil confirmar este hecho, ya que pocas compañías comercializan un seguro de responsabilidad civil en solitario, sino que suelen incluir como coberturas mínimas asistencia en viaje, seguro médico o defensa jurídica y reclamación de daños entre otras. Sin embargo, los resultados parecen ser coherentes con los datos analizados, por lo que parece razonable confirmar el buen desenlace del estudio.

El único punto negativo o a mejorar del trabajo, es que solo ha funcionado a partir de modelos con pocas variables, ya que, al introducir más, estas no aportaban lo suficiente para ser significativas y que el modelo fuese eficaz.

Las compañías en la realidad utilizan bastantes más variables a la hora de tarificar, sin embargo, en el caso de un trabajo académico como este, es complejo obtener una base de datos completa o lo suficientemente grande debido a la negativa de las entidades aseguradoras a proporcionarla.

7. Bibliografía

- **Alonso-González, P. J. (2020):** *Tema 4: Introducción a los modelos GLM.* Simulación y modelos para el análisis de datos actuariales. Universidad de Alcalá.
- **Alonso-González, P. J. (2021):** *Modelo Lineal General (GLM).* Programación con R. Universidad de Alcalá.
- **Arroyo, I., Bravo, L. C., Llinás, H. y Muñoz, F. L. (2014):** *Poisson and Gamma Distributions: A Discrete and Continuous Relationship.* Vol. 12, Nº.1, enero – junio de 2021, págs. 99-107, Universidad del Norte.
- **De Jong, P. y Heller, G. Z. (2008):** *Generalized Linear Models for Insurance Data.* Cambridge.
- **Dobson, A.J. (1990):** *An Introduction to Generalized Linear Models.* Second Edition. Chapman and Hall.
- **Engelhardt, A. (2012):** *R code for Chapter 2 of Non-Life Insurance Pricing with GLM.*
- **Haberman, S. y Renshaw, A. E. (1996):** *Generalized Linear Models and Actuarial Science*. Journal of the Royal Statistical Society, series D, vol 45, nº 4, pp. 407-436.
- **Harbin, J.W. y Hilbe, J.M. (2018):** *Generalized Linear Models and Extensions.* Fourth edition. StataPress.
- **López-Gonzalez, E. y Ruiz-Soler, M. (2011):** *Análisis de datos con el Modelo Lineal Generalizado. Una aplicación con R.* Universidad de Malaga.
- **Marín, J. M. (2018):** *Tema 3: Modelos lineales generalizados.* Análisis de datos categóricos. Universidad Carlos III de Madrid.
- **McCullagh, P. y Nelder, J. A. (1989):** *Generalized Linear Models.* Second edition. Chapman and Hall.
- **Ohlsson, E y Johansson, B. (2010):** *Non-Life Insurance Pricing with Generalized Linear Models.* First Edition. EAA Series.

8. Anexos

Códigos utilizados en RStudio para el modelo de frecuencias:

```
require(openxlsx)
(Cartera=read.xlsx("Cartera.xlsx"))
require(stats)

MFQ1=glm(siniestro~edad+vehiaños,family=poisson(link="log"),data=Cartera)
summary(MFQ1)

MFQ2=glm(siniestro~edad+sexo,family=poisson(link="log"),data=Cartera)
summary(MFQ2)

MFQ3=glm(siniestro~edad+vehivalor,family=poisson(link="log"),data=Cartera)
summary(MFQ3)

MFQ4=glm(siniestro~vehiaños+sexo,family=poisson(link="log"),data=Cartera)
summary(MFQ4)

MFQ5=glm(siniestro~vehiaños+vehivalor,family=poisson(link="log"),data=Cartera)
summary(MFQ5)
```

```
MFQ6=glm(siniestro~sexo+vehivalor,family=poisson(link="log"),data=Cartera)
summary(MFQ6)

MFQ7=glm(siniestro~edad+vehiaños+sexo,family=poisson(link="log"),data=Cartera)
summary(MFQ7)

MFQ8=glm(siniestro~edad+vehiaños+vehivalor,family=poisson(link="log"),data=Cartera)
summary(MFQ8)

MFQ9=glm(siniestro~edad+sexo+vehivalor,family=poisson(link="log"),data=Cartera)
summary(MFQ9)

MFQ10=glm(siniestro~vehiaños+sexo+vehivalor,family=poisson(link="log"),data=Cartera)
summary(MFQ10)

MFQ11=glm(siniestro~edad+vehiaños+sexo+vehivalor,family=poisson(link="log"),data=Cartera)
summary(MFQ11)

AIC(MFQ3)
AIC(MFQ1)
anova(MFQ3,MFQ1,test="Chisq")
```

```
AIC(MFQ8)

AIC(MFQ9)

anova(MFQ8,MFQ9,test="Chisq")


write.csv(MFQ3$coefficients, " FQ2.csv")

write.xlsx(MFQ3$coefficients, " FQ2.xlsx")


write.csv(MFQ8$coefficients, " FQ3.csv")

write.xlsx(MFQ8$coefficients, " FQ3.xlsx")
```

Códigos utilizados en RStudio para el modelo de costes medios:

```
require(openxlsx)

(Coste=read.xlsx("Coste.xlsx"))

require(stats)


MCM1=glm(coste~edad+vehiaños,family=Gamma(link="log"),data=Coste)

summary(MCM1)


MCM2=glm(coste~edad+sexo,family=Gamma(link="log"),data=Coste)

summary(MCM2)


MCM3=glm(coste~edad+vehivalor,family=Gamma(link="log"),data=Coste)
```

```
summary(MCM3)

MCM4=glm(coste~vehiaños+sexo,family=Gamma(link="log"),data=Coste)
summary(MCM4)

MCM5=glm(coste~vehiaños+vehivalor,family=Gamma(link="log"),data=Coste)
summary(MCM5)

MCM6=glm(coste~sexo+vehivalor,family=Gamma(link="log"),data=Coste)
summary(MCM6)

MCM7=glm(coste~edad+vehiaños+sexo,family=Gamma(link="log"),data=Coste)
summary(MCM7)

MCM8=glm(coste~edad+vehiaños+vehivalor,family=Gamma(link="log"),data=Coste)
summary(MCM8)

MCM9=glm(coste~edad+sexo+vehivalor,family=Gamma(link="log"),data=Coste)
summary(MCM9)

MCM10=glm(coste~vehiaños+sexo+vehivalor,family=Gamma(link="log"),data=Coste)
summary(MCM10)
```

```
MCM11=glm(coste~edad+vehiaños+sexo+vehivalor,family=Gamma(link="log"),data=Coste
)

summary(MCM11)


AIC(MCM1)

AIC(MCM3)

anova(MCM1,MCM3,test="Chisq")


AIC(MCM7)

AIC(MCM9)

anova(MCM7,MCM9,test="Chisq")


write.csv(MCM1$coefficients, " CM2.csv")

write.xlsx(MCM1$coefficients, " CM2.xlsx")


write.csv(MCM7$coefficients, " CM3.csv")

write.xlsx(MCM7$coefficients, " CM3.xlsx")
```