TRABAJO DE FIN DE GRADO

Grado en Estadística

Título: Estimación de la prima pura en seguros de automóvil mediante modelos lineales generalizados y su aplicación en Power BI

Autor: Elena Varas Sánchez

Director: Anna Salazar Belver

Departamento: Departamento de Econometria, Estadística y Economia Aplicada

Convocatoria: Enero 2026

UNIVERSITAT DE BARCELONA

UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH
Facultat de Matemàtiques i Estadística

Agradecimientos

A mi tutora, Anna Salazar Belver, por haberme guiado y apoyado desde el primer día.

A mi familia, que siempre creyeron mucho más en mí de lo que yo lo hacía. Gracias por cuidarme, animarme y apoyarme cada día de estos cuatro años.

A mi hermana Ana, por estar siempre ahí a pesar de la distancia. Por guiarme, por aconsejarme y, sobre todo, por saber criticarme cuando hacía falta. Gracias por motivarme a ser mejor cada día y por creer siempre en mí.

A mi abuela, gracias por cada vela en cada examen, trabajo o día importante.

Y, sobre todo, a mis abuelos, que sabían que lo conseguiría, pero no pudieron llegar a verlo.

Resumen

En los seguros de automóvil, la tarificación es un elemento clave para garantizar el equilibrio técnico de la cartera y una adecuada gestión del riesgo. Este proceso requiere comprender y cuantificar adecuadamente la siniestralidad, teniendo en cuenta la elevada heterogeneidad del riesgo presente en este tipo de carteras. Por ello, es necesario disponer de modelos que permitan estimar de forma consistente el coste esperado y analizar la influencia de las características del asegurado y del vehículo. Además, es importante que este análisis sea comprensible y pueda presentarse de forma clara y sencilla a perfiles de gestor, de forma que los resultados sirvan como apoyo a la toma de decisiones.

En este trabajo, se emplean modelos lineales generalizados para estimar la prima pura y analizar de forma diferenciada las dos componentes fundamentales de la siniestralidad: la frecuencia y la severidad de los siniestros. La modelización se aborda mediante un enfoque en dos etapas, que se complementa con una estimación alternativa basada en el modelo Tweedie, permitiendo comparar ambas metodologías. A lo largo del trabajo se discuten las principales decisiones metodológicas adoptadas y se analiza el efecto de las distintas variables explicativas sobre el riesgo esperado.

Una vez estimados los modelos, los resultados ponen de manifiesto la elevada heterogeneidad del riesgo y una mayor capacidad explicativa de la variable en la modelización de la frecuencia que en la severidad. Además, la comparación entre metodologías muestra que el modelo Tweedie tiende a generar estimaciones de prima pura más conservadores, mientras que el enfoque frecuencia-severidad ofrece una mayor interpretabilidad. Por último, los resultados obtenidos se integran en un *dashboard* interactivo en Power BI que permite visualizar el riesgo de forma clara y apoyar la toma de decisiones en la gestión aseguradora.

Palabras clave: Seguros de automóvil, siniestralidad, prima pura, modelos lineales generalizados, tarificación, Power BI.

Abstract

In motor insurance, pricing is a key element to ensure the technical balance of the portfolio and an appropriate management of risk. This process requires a proper understanding and quantification of claims experience, taking into account the high level of risk heterogeneity that characterizes this type of insurance. For this reason, it is necessary to use models that allow a consistent estimation of the expected cost of claims and to analyse how the characteristics of the policyholder and the vehicle influence this cost. In addition, it is particularly important that this analysis is easy to understand and can be presented in a clear and simple way to insurance managers, so that the results can effectively support decision-making.

In this work, Generalized Linear Models are used to estimate the pure premium and to analyse separately the two main components of claims experience: claim frequency and claim severity. The modelling strategy follows a two-stage approach, where frequency and severity are estimated independently. This approach is complemented by an alternative estimation based on the Tweedie model, which provides an integrated estimation of the total claim cost. Throughout the study, the main methodological decisions are discussed, and the effect of different explanatory variables on the expected risk is analysed.

Once the models have been estimated, the results show a high level of risk heterogeneity and a greater explanatory power of the variables in the modelling of claim frequency than in claim severity. Furthermore, the comparison between methodologies indicates that the Tweedie model tends to produce more conservative pure premium estimates, while the frequency–severity approach offers greater interpretability. Finally, the results obtained are integrated into an interactive dashboard developed in Power BI, which allows the risk to be visualised in a clear way and supports decision-making in an insurance management context.

Keywords: Motor insurance, claims experience, pure premium, Generalized Linear Models, pricing, Power BI

Clasificación AMS (MSC 2010)

62-XX STATISTICS

- *62J12 Generalized linear models*
- *62F10 Point estimation*
- *62F35 Hypothesis testing*
- *62P05 Applications to actual sciences and financial mathematics*

Índice

1. Introducción

La tarificación constituye uno de los pilares fundamentales en la gestión técnica de los seguros de automóvil, ya que de ella depende en gran medida el equilibrio financiero de la cartera y la adecuada valoración del riesgo asumido por la entidad aseguradora. En este contexto, la correcta estimación de la siniestralidad resulta esencial para asegurar que el negocio sea sostenible y que las primas reflejen adecuadamente el riesgo de cada asegurado. La elevada heterogeneidad presente en las carteras de automóvil, derivada de la diversidad de conductores y vehículos, hace necesario el uso de herramientas estadísticas capaces de capturar de forma adecuada dicha variabilidad.

Desde un punto de vista actuarial, la siniestralidad se descompone habitualmente en dos componentes: la frecuencia de los siniestros y la severidad. Esta descomposición permite analizar por separado la probabilidad de ocurrencia de eventos y su impacto económico. De esta manera, se pretende facilitar la interpretación de los factores que influyen en el riesgo. En este contexto, los modelos lineales generalizados se utilizan de forma habitual en el ámbito asegurador, ya que permiten trabajar con variables que no siguen una distribución normal.

El estudio se basa en una base de datos procedentes de CASdatasets, un conjunto de datos de referencia en el ámbito actuarial. Esta base de datos está desarrollada por la *Casualty Actuarial Society* [1] con fines formativos y de investigación [2,3]. Recoge información detallada sobre pólizas y siniestros de seguros de automóvil y se utiliza habitualmente en trabajos académicos y ejemplos prácticos de tarificación.

El estudio se centra en evaluar el efecto de distintas características del conductor, del vehículo y de las variables tarifarias sobre la siniestralidad, diferenciando entre la frecuencia de los siniestros y su severidad. Como punto de partida, se considera que la cartera está formada por perfiles de riesgo distintos, lo que da lugar a diferencias en su comportamiento siniestral. Además, se plantea que los factores explicativos pueden afectar de forma distinta a la probabilidad de ocurrencia de siniestros y a su coste medio. Esto justifica la descomposición de la siniestralidad en ambas componentes. De forma complementaria, se utiliza el modelo de Tweedie para obtener una estimación directa del coste esperado por póliza y comparar ambos métodos en la práctica.

El trabajo se estructura de la siguiente manera. En el primer apartado se presenta la metodología empleada y el marco teórico de los modelos utilizados. A continuación, se desarrolla un análisis exploratorio y descriptivo de la cartera y de las principales variables de interés. Posteriormente, se aborda la modelización de la siniestralidad y la estimación de la prima pura mediante los distintos enfoques considerados, así como una comparación de los resultados obtenidos. Finalmente, se presenta una aplicación práctica mediante el desarrollo de un *dashboard* interactivo en Power BI orientado a apoyar la toma de decisiones en un entorno de gestión aseguradora, y se exponen las conclusiones principales del estudio. Con

el objetivo de garantizar la transparencia y la reproducibilidad del análisis, el código desarrollado en R, así como los materiales complementarios utilizados en el trabajo, se encuentran disponibles en un repositorio público en GitHub[1]. En repositorio incluye la estructura del proyecto, los scripts empleados en la modelización y el acceso al dashboard desarrollado en Power BI.

[1] Repositorio GitHub del proyecto: https://github.com/elenavs03/Elena-Varas-Sanchez-TFG

2. Metodología

La siniestralidad, desde el punto de vista actuarial, se descompone en dos dimensiones fundamentales: la frecuencia y la severidad del siniestro. La frecuencia mide el número de siniestros ocurridos por póliza ajustado al tiempo de exposición al riesgo, mientras que la severidad recoge el coste medio asociado a cada siniestro una vez ocurrido. Esta descomposición permite estudiar de forma separada la probabilidad de ocurrencia y el impacto económico.

A partir de la frecuencia y la severidad se define la prima pura o prima técnica de las pólizas. La prima pura se describe como el coste esperado de los siniestros por póliza durante el periodo considerado. Este parámetro se calcula como el producto entre la frecuencia esperada de siniestros y la severidad media esperada. A diferencia de la prima, la prima pura no incluye gastos de gestión, recargos comerciales ni márgenes de beneficio. Su estimación adecuada es fundamental para garantizar el equilibrio técnico del seguro y la sostenibilidad de la cartera.

La estimación del valor de la prima pura se efectúa utilizando Modelos Lineales Generalizados (MLG) con dos objetivos: (i) analizar y entender el efecto de distintas características del asegurado y del vehículo sobre la siniestralidad, y (ii) obtener predicciones de frecuencia, severidad y coste total orientadas a la estimación de la prima pura. En este contexto, el objetivo del trabajo no se centra en maximizar la capacidad predictiva individual, sino en obtener estimaciones estables e interpretables del riesgo medio, adecuadas para el análisis y la tarificación de la cartera.

2.1 Modelos lineal generalizados (MLG)

Los Modelos Lineales Generalizados (MLG) permiten adaptar el modelo estadístico a la naturaleza de variables de interés que no siguen una distribución normal. En el ámbito asegurador, los MLG se utilizan de forma habitual para modelizar la siniestralidad, ya que permiten analizar variables que presentan características como discreción, asimetría o presencia de valores nulos. Este enfoque proporciona un marco flexible y coherente para estudiar el efecto de las características del asegurado y del vehículo sobre el riesgo esperado. De este modo, los MLG constituyen una herramienta fundamental en los procesos de tarificación en seguros no vida [13].

Formalmente, para cada póliza $i = 1, \dots, n$, se modeliza el valor esperado de la variable respuesta Y_i en función de sus covariables x_i, a través de una función de enlace que relaciona dicha media μ_i con un predictor lineal η. Esta relación se expresa mediante el siguiente predictor lineal:

$$g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

El proceso de modelización se estructura en tres partes: (i) análisis de la frecuencia de los siniestros, (ii) estudio de la severidad condicionada a la ocurrencia y, (iii) modelización compuesta de las partes (i) y (ii).

Parte	Modelo	Parámetros	Esperanza	Varianza
Frecuencia (F)	Poisson	μ_i	$E(N_i) = \mu_i$	$Var(N_i) = \mu_i$
	Binomial negativa	μ_i , θ	$E(N_i) = \mu_i$	$Var(N_i) = \mu_i + \frac{\mu_i^2}{\theta}$
Severidad (S)	Gamma	μ_i , ϕ	$E(C_i) = \mu_i$	$Var(C_i) = \phi\mu_i^2$
Modelo compuesto (F + S)	Tweedie	μ_i , ϕ , p	$E(C_i) = \mu_i$	$Var(C_i) = \phi\mu_i^p$

Tabla 1. Modelos de siniestralidad.
Distribuciones para la modelización de la frecuencia, la severidad y el coste total. El término μ_i representa el valor esperado de la variable respuesta, θ controla el grado de sobredispersión, ϕ es el parámetro de sobredispersión y p determina la estructura del modelo Tweedie.

2.2 Modelización de la frecuencia de los siniestros

La modelización de la frecuencia de los siniestros tiene como objetivo describir el número de siniestros N_i. Esta variable es un conteo que registra la ocurrencia de siniestros discretos a lo largo de un periodo de exposición. Dentro del marco de los modelos lineales generalizados, el modelo Poisson está diseñado para modelizar el número de eventos que ocurren en un intervalo fijo cuando los eventos son independientes y ocurren con una intensidad media constante. Este tipo de modelos se utiliza de forma habitual en el análisis de la siniestralidad y de eventos poco frecuentes [5]. Bajo estas hipótesis, el modelo Poisson proporciona la siguiente formulación apropiada para el análisis de la frecuencia siniestral:

$$N_i \sim Poisson\,(\mu_i), \quad g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}\,.$$

Bajo la distribución Poisson se cumple que $Var(N_i|x_i) = E(N_i|x_i)$. Dicha hipótesis puede ser demasiado restrictiva en carteras reales por heterogeneidad no capturada, dependencia entre riesgos o excesos de ceros. En estos casos, al observar sobredispersión (varianza mayor que la media) podemos adoptar un modelo binomial negativo, una alternativa más flexible de los MLG [6].

El modelo binomial negativo es un modelo de conteo que extiende al modelo Poisson incorporando un parámetro adicional de dispersión, lo que permite capturar situaciones en las que la varianza supera a la media. La parametrización adoptada en este trabajo es:

$$N_i \sim BN(\mu_i, \theta), \qquad g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij},$$

con relación varianza-media:

$$Var\,(N_i|x_i) \;=\; \mu_i + \frac{\mu_i^{\,2}}{\theta},$$

donde $\theta > 0$ controla el grado de sobredispersión. Este modelo mantiene la misma estructura sistemática, pero incorpora un parámetro de dispersión adicional, proporcionando una inferencia más estable y un ajuste generalmente más realista en frecuencia siniestral.

2.3 Modelación de la severidad de los siniestros

La severidad se define como el coste medio por siniestro condicionado a que exista al menos un siniestro. A nivel de modelo, trabajamos con la variable S_i, que corresponde al coste medio por siniestro de la póliza i. Dado que S_i es continua, estrictamente positiva y típicamente asimétrica a la derecha, se emplea un GLM con distribución Gamma:

$$S_i|\,(N_i > 0) \;\sim Gamma\,(\mu_i,\,\phi), \qquad g(\mu_i) = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}.$$

La elección Gamma se ajusta a situaciones donde la variabilidad del coste crece con el nivel medio del mismo. Al definir la severidad como un promedio por póliza, su fiabilidad está condicionada por el número de siniestros registrados. Por ello, la estimación puede realizarse mediante máxima verosimilitud incorporando ponderaciones $w_i = N_i$, de modo que las pólizas con más siniestros aporten más información y se reduzca la variabilidad asociada a medias calculadas.

La severidad estimada combinada con la frecuencia de siniestros permiten obtener la prima pura por póliza mediante un enfoque en dos etapas.

2.4 Modelos compuestos y distribución Tweedie

La modelización en dos etapas permite separar mecanismos, pero también es de interés modelizar el coste total anual por póliza. Esta variable presenta una estructura particular al registrar muchas pólizas sin siniestros que tienen un coste total igual a cero, mientras que aquellas que sí presentan siniestros muestran importes positivos con una distribución asimétrica.

La distribución Tweedie [9] permite analizar este tipo de variables en un único modelo, ya que combina ambos comportamientos. En función del valor del parámetro de forma p, el modelo Tweedie puede interpretarse como el coste total anual por póliza C_i como el resultado de la suma de un número aleatorio de siniestros, cada uno de ellos asociado a un importe positivo.

Desde un punto de vista estadístico, el modelo Tweedie se define por una relación entre la varianza y la media, según descrito en la siguiente fórmula:

$$Var(C_i|x_i) = \phi\mu_i^{\,p}, \text{ con } \phi > 0 \text{ y } 1 < p < 2,$$

En el marco de los MLG, el valor esperado del coste total se relaciona con las covariables mediante un predictor lineal:

$$C_i \sim Tweedie\,(\mu_i, \phi, p), \qquad g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij}$$

Este planteamiento permite obtener una estimación directa del coste esperado por póliza ajustada por exposición y evaluar de forma conjunta el efecto de las variables explicativas sobre la siniestralidad. No obstante, el modelo Tweedie permite una estimación integrada del coste, pero sacrifica parte de la interpretabilidad de las componentes de la siniestralidad.

2.5 Validación y criterios de evaluación

2.5.1 Interpretación de los coeficientes del modelo

Los coeficientes de los MLG se interpretan según la función de enlace y la distribución asumida para la variable respuesta. En este trabajo se emplea el enlace logarítmico, de manera que, los coeficientes indican cómo influyen las variables explicativas en el valor esperado de la variable. En términos prácticos, un coeficiente positivo indica que un aumento de la variable explicativa se asocia con un incremento del valor esperado de la variable respuesta, mientras que un coeficiente negativo implica una disminución. Con el objetivo de cuantificar el cambio se aplica una función exponencial a los coeficientes que pasan a interpretarse de forma multiplicativa. En concreto, $exp(\beta_i)$ indica en qué proporción cambia el valor esperado de la variable respuesta cuando la variable x_{ij} aumenta en una unidad, manteniendo el resto de las variables constantes. Esta interpretación permite evaluar directamente cómo las características del riesgo influyen en la frecuencia, la severidad o el coste total esperado de los siniestros.

2.5.2 AIC como criterio de comparación

El criterio de información de Akaike (AIC) se utiliza para comparar distintos modelos estimados sobre el mismo conjunto de datos y con la misma variable respuesta. Esta métrica evalúa el equilibrio entre la calidad del ajuste y la complejidad del modelo. El AIC penaliza la inclusión de un mayor número de parámetros, evitando así modelos excesivamente complejos que puedan sobreajustar los datos. Matemáticamente, se define como:

$$AIC = -2\,l(\hat{\theta}) + 2k,$$

donde $l(\hat{\theta})$ es el logaritmo de la función de verosimilitud evaluado en los parámetros estimados, y *k* es el número de parámetros del modelo.

En la selección entre modelos estimados sobre el mismo conjunto de datos, se prioriza aquel con menor AIC, ya que ofrece un mejor compromiso entre ajuste y parsimonia.

2.5.3 Diagnóstico de los residuos

La calidad del ajuste de los modelos estimados se evalúa a través de un diagnóstico de los residuos con el fin de comprobar si el modelo captura adecuadamente la relación entre las variables explicativas y la variable respuesta. En particular, se revisa que los residuos no presenten patrones sistemáticos cuando se representan frente a los valores ajustados o frente a las variables explicativas. La aparición de tendencias, formas curvas o agrupaciones puede indicar que el modelo no está capturando correctamente la relación entre las variables o que faltan covariables relevantes. También se analiza la heterocedasticidad, que se produce cuando la variabilidad de los residuos no es constante. Este efecto se identifica cuando la dispersión de los residuos aumenta o disminuye a lo largo de los valores ajustados. Por último, se estudia la presencia de observaciones atípicas y observaciones influyentes, que pueden detectarse mediante residuos de gran magnitud, como la distancia de *Cook*, que permite identificar aquellas observaciones que ejercen una influencia desproporcionada sobre la estimación de los parámetros del modelo.

En conjunto, un buen ajuste se caracteriza por residuos distribuidos de forma aproximadamente aleatoria alrededor del cero, con variabilidad constante y sin observaciones individuales que ejerzan una influencia desproporcionada sobre la estimación. Dado el tamaño de la cartera analizada, pequeñas desviaciones pueden resultar estadísticamente significativas sin que ello implique necesariamente una falta de adecuación del modelo desde un punto de vista práctico.

2.6 Metodología de estimación de la prima pura

La estimación de la prima pura se realiza a partir de las predicciones obtenidas por los modelos de siniestralidad ajustados. En este trabajo se emplean dos procedimientos de estimación, que permiten aproximar el coste técnico esperado del riesgo desde perspectivas complementarias: (i) enfoque frecuencia-severidad y (ii) enfoque integrado mediante el modelo Tweedie. Ambos se apoyan en modelos lineales generalizados y permiten obtener una estimación individualizada del coste esperado por póliza.

En el enfoque tradicional para estimar la prima pura de una póliza se define como el producto entre el número de siniestros que se espera que ocurra y el coste medio asociado a cada uno de ellos ($PP_i^{(FS)}$) . Denotando por N_i el número de siniestros por póliza i y por S_i el coste medio de dichos siniestros, el coste total esperado puede expresarse como:

$$\pi_i = E(C_i|x_i) = E(N_i \cdot S_i|x_i).$$

Bajo el supuesto habitual de independencia entre frecuencia y severidad, dicha esperanza puede descomponerse como el producto de las esperanzas marginales de ambas componentes, lo que conduce a:

$$E(C_i|x_i) = E(N_i|x_i) \cdot E(S_i|N_i > 0, x_i)$$

La frecuencia esperada de siniestros y la severidad media esperada se obtienen a partir de los modelos estimados para cada una de las componentes de la siniestralidad. Bajo este planteamiento, la prima pura asociada a la póliza i se define:

$$PP_i^{(FS)} = E(N_i|x_i) \cdot E(S_i|N_i > 0, x_i) = \mu_i^{(F)} \cdot \mu_i^{(S)}.$$

Este razonamiento permite modelizar de forma separada la ocurrencia de siniestros y su impacto económico, manteniendo una interpretación clara y coherente desde un punto de vista actuarial.

Como alternativa a este enfoque, la prima pura puede estimarse de forma directa a partir de un modelo Tweedie ($PP_i^{(T)}$) aplicado al coste total anual por póliza. Denotando bajo C_i el coste total anual por póliza i, la prima pura estimada mediante esta perspectiva se expresa como:

$$PP_i^{(T)} = E(C_i|x_i) = \mu_i^{(T)}$$

Ambos enfoques proporcionan estimaciones coherentes de la prima pura a partir de los valores esperados de la siniestralidad, aunque difieren en la forma en que representan la estructura del riesgo. La comparación entre ambas metodologías permite evaluar sus diferencias prácticas y fundamentar la elección del enfoque más adecuado, análisis que se desarrollará más adelante en profundidad.

3. Análisis exploratorio y descriptivo

En este apartado se realiza un análisis descriptivo de los datos con el objetivo de obtener una visión general de la cartera. Las variables que intervienen en el estudio han sido agrupadas en cinco categorías: (i) exposición y siniestralidad, (ii) variables económicas del riesgo, (iii) características del conductor, (iv) características del vehículo, (v) y variables tarifarias. Los detalles de cada grupo se describen en la Tabla 2.

Categoría	Variable	Descripción	Tipo
Exposición y siniestralidad	Exposición	Tiempo asegurado de la póliza por la compañía en años	Cuantitativa continua
	Número siniestros	Número de reclamaciones realizadas durante el periodo de exposición	Cuantitativa discreta
	Tasa de siniestros	Frecuencia relativa con la que ocurren los siniestros una vez ajustada la duración de cobertura de cada póliza por unidad de exposición	Cuantitativa continua
Variables económicas del riesgo	Importe de siniestros	Suma de los costes indemnizados por cada póliza a lo largo del periodo de exposición	Cuantitativa continua
	Prima pura estimada	Aproximación al coste esperado del riesgo para cada póliza por unidad de exposición	Cuantitativa continua
Características del conductor	Edad del conductor	Años del asegurado	Cuantitativa discreta
Características del vehículo	Edad del vehículo	Antigüedad del vehículo en años	Cuantitativa discreta
	Potencia	Media ordinal de la potencia del motor en orden ascendente marcada por una escala del 1 al 15	Categórica ordinal
	Tipo de combustible	Tipo de carburante utilizado por el vehículo asegurado: diesel o gasolina	Categórica nominal
Variables tarifarias	Bonus-Malus	Índice de bonificación o penalización aplicado a la prima del asegurador en función de su historial de siniestros	Cuantitativa continua

Tabla 2. Descripción y tipología de las variables del conjunto de datos.
Variables empleadas en el análisis clasificadas por categorías. La clasificación de las variables ha sido determinada según su naturaleza.

Este análisis permite identificar patrones básicos, distribuciones características y posibles valores atípicos que pueden influir en la posterior modelización.

3.1 Variables de exposición y siniestralidad

3.1.1 Exposición

La variable exposición representa el tiempo efectivo durante el cual cada póliza permanece cubierta por la aseguradora. Este periodo determina la duración real del riesgo asumido y resulta, por tanto, un elemento esencial en el análisis actuarial de siniestralidad.

En el análisis exploratorio se pone de manifiesto una elevada concentración de pólizas con exposiciones cercanas al año, junto con un conjunto reducido de contratos de duración inferior (Figura 1.A). También se contemplaron algunas observaciones superiores al año, atribuibles a inconsistencias de los datos. Debido a que las pólizas deben pertenecer a un único ejercicio anual consideramos estos valores como inconsistencias, o posibles errores de registro y para corregirlos los convertimos en pólizas anuales (Figura 1.B). Con esta depuración se garantiza la coherencia temporal del conjunto de datos y se evitan distorsiones en la estimación de la frecuencia (Figura 1.C).

3.1.2 Número de siniestros

El número de siniestro es una de las magnitudes principales del análisis, al cuantificar la ocurrencia de reclamaciones a lo largo del periodo de cobertura. Al analizar la variable se evidencia una fuerte concentración de pólizas sin siniestros y una presencia muy limitada de valores positivos (Tabla 3). Esta distribución altamente asimétrica es característica de los datos de seguros, donde la ocurrencia de siniestros es poco frecuente. El comportamiento observado, apunta a la conveniencia de emplear modelo de conteo capaces de capturar adecuadamente la baja probabilidad de aparición de eventos.

Intervalo	Recuento	Porcentaje
0	653047	96.32%
1	23571	3.48%
2	1298	0.19%
3	62	0.01%
4-5	7	0.00%
6-10	3	0.00%
11-20	3	0.00%

Tabla 3. Distribución del número de siniestros por póliza.
El 96% de las pólizas no registran siniestros durante el periodo analizado, mientras que una proporción muy reducida presenta uno o dos eventos. Los recuentos superiores son excepcionales y representan una fracción residual de la cartera.

Con el fin de evitar que observaciones infrecuentes distorsionen la estimación de la frecuencia, se llevó a cabo una depuración de las pólizas con un número de siniestros superior a cuatro. Esta decisión responde al objetivo de no modelizar la cola extrema de la distribución, sino de obtener una representación estable del comportamiento típico de la siniestralidad en la cartera. El estudio detallado de este tipo de casos, aunque relevante desde un punto de vista actuarial, excede el alcance de este trabajo. Tras la depuración, no se observa que la estructura general de la cartera se vea alterada (Figura 1.D), lo que confirma que la eliminación de estas observaciones afecta únicamente a casos marginales que podrían distorsionar el análisis estadístico.

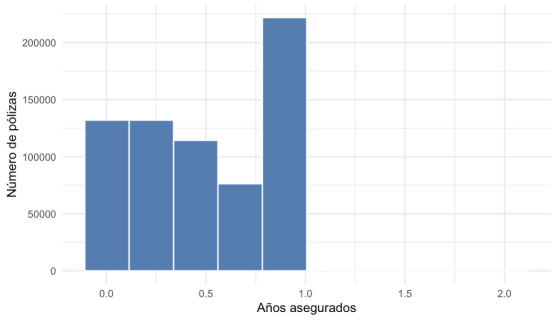
3.1.3 Tasa de siniestros

La tasa de siniestros no se encuentra directamente en la base de datos, sino que ha sido calculada como el cociente entre el número de siniestros y la exposición. Esta nueva variable permite analizar la frecuencia de ocurrencia de siniestros ajustada por el tiempo de exposición, facilitando la comparación entre pólizas con diferentes periodos de vigencia. Desde un punto de vista exploratorio, esta medida proporciona una visión estandarizada del riesgo y resulta especialmente útil para identificar patrones que podrían quedar enmascarados al analizar únicamente el número absoluto de siniestros.

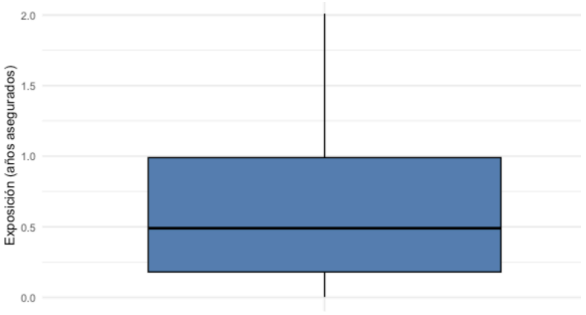
El análisis muestra como era de esperar, que la gran mayoría de las pólizas presentan una tasa nula, consecuencia directa de la elevada proporción de contratos sin siniestros (Figura 1.E). Al centrar la atención en las tasas estrictamente positivas, se observa una distribución asimétrica, con una fuerte concentración en valores próximos a la unidad y una cola derecha alargada. Las tasas más elevadas se asocian principalmente a pólizas con exposiciones reducidas que han registrado uno o más siniestros, generando valores elevados del cociente. (Figura 1.E)

La relación observada entre la tasa de siniestros y la exposición evidencia una asociación inversa entre ambas magnitudes, de modo que las pólizas con periodos de cobertura más reducidos presentan una mayor variabilidad en la tasa estimada (Figura 1.G). Este comportamiento confirma que la tasa de siniestros es sensible a exposiciones cortas, lo que motiva el uso de modelos de conteo con *offset* de exposición en lugar de modelizar directamente la tasa.

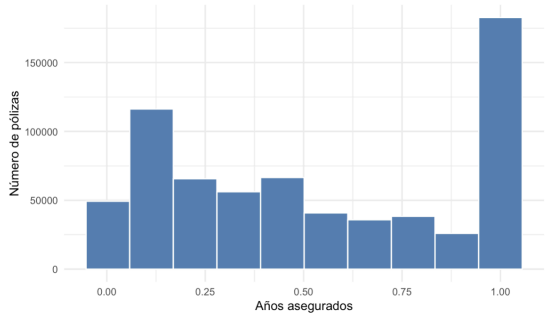
A


Número de pólizas
200000
150000
100000
50000
0
0.0
0.5
1.0
1.5
2.0
Años asegurados

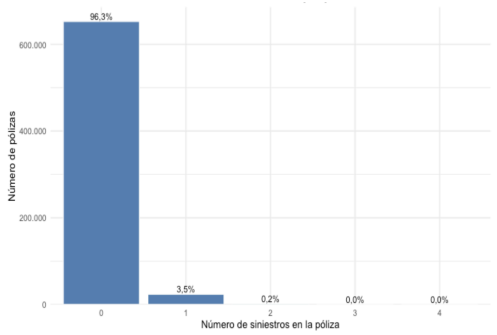
B


Exposición (años asegurados)
2.0
1.5
1.0
0.5
0.0

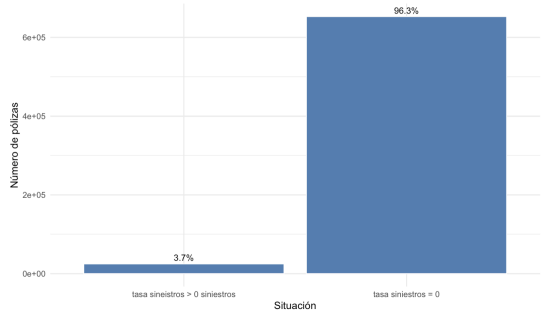
C


Número de pólizas
150000
100000
50000
0
0.00
0.25
0.50
0.75
1.00
Años asegurados

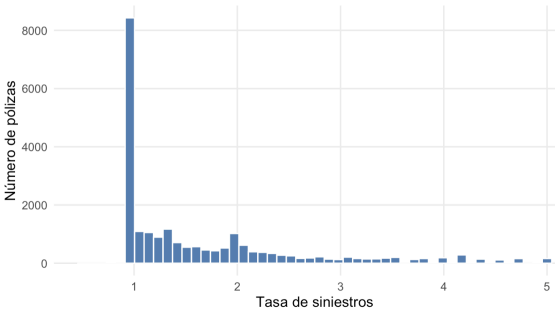
D


96.3%
600.000
200.000
200.000
Número de pólizas
3.5%
0.2%
0.0%
0.0%
0
1
2
3
4
Número de siniestros en la póliza

E


96.3%
6e+05
4e+05
2e+05
0e+00
Número de pólizas
3.7%
tasa sineistros > 0 siniestros
tasa siniestros = 0
Situación

F


8000
6000
4000
2000
0
Número de pólizas
1
2
3
4
5
Tasa de siniestros

G

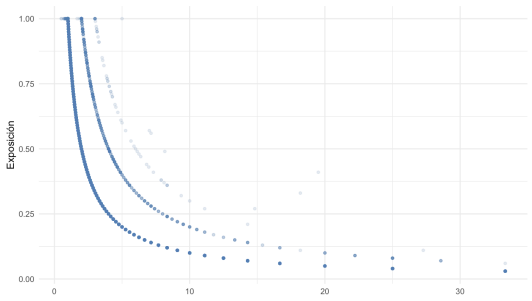

1.00
0.75
0.50
0.25
0.00
Exposición
0
10
20
30

Figura 1. Análisis exploratorio de la exposición, la frecuencia y la tasa de siniestros.
(A)Distribución de la exposición medida en años asegurados, con una elevada concentración de pólizas anuales o cercanas al año. (B) Diagrama de caja de la exposición, que muestra la dispersión de los períodos asegurados y la presencia de algunos valores superiores al año. (C) Distribución de la exposición tras la depuración de registros, con una clara acumulación en el valor anual. (D) Distribución del número de siniestros por póliza, caracterizada por una elevada proporción de pólizas sin siniestros y una frecuencia decreciente a medida que aumenta el número de eventos. (E) Proporción de pólizas con tasa de siniestros nula y positiva, destacando la predominancia de pólizas sin siniestros durante el periodo analizado. (F) Distribución de la tasa de siniestros positiva, con una fuerte concentración en valores próximos a uno y una cola derecha alargada. (G) Relación entre la tasa de siniestros y la exposición, que pone de manifiesto una mayor variabilidad de la tasa en pólizas con exposiciones reducidas.

3.2 Variables económicas del riesgo

3.2.1 Importe medio siniestro

Con el objetivo de aislar la intensidad económica del daño del número de eventos ocurridos, resulta de interés para el estudio analizar la severidad media de los siniestros registrados por póliza. Esta variable se define como el cociente entre el importe total y el número de siniestros de cada póliza, condicionado a que exista al menos un siniestro. De esta manera, podemos analizar la severidad condicionada a la ocurrencia de al menos un siniestro y se reduce la influencia de pólizas con múltiples eventos sobre la variabilidad del coste.

La distribución del importe medio presenta una marcada asimetría a la derecha, característica de los costes siniestrales (Figura 2.A). Desde un punto de vista actuarial, la severidad media constituye una medida especialmente adecuada para analizar el coste esperado por siniestro. Permite aislar la magnitud económica del daño de la frecuencia de ocurrencia. De este modo, proporciona una base coherente para su utilización como variable respuesta en la modelización de la severidad.

Para disminuir la influencia desproporcionada de estos valores extremos y con el fin de obtener estimaciones más robustas, se aplica un trucado superior del importe medio por siniestro. Esta decisión responde al objetivo del trabajo de analizar el comportamiento típico del coste por siniestro en la cartera, y no el de modelizar la cola extrema de la distribución. El truncado permite limitar la contribución de los siniestros más excepcionales sin eliminar observaciones, preservando así la estructura general de la distribución, tal y como se observa en la Tabla 4.

Estadístico	Severidad original	Severidad truncada
Mínimo	1.00	1.00
Primer cuartil	710.56	710.56
Mediana	1172.00	1172.00
Media	2222.25	1683.84
3er cuartil	1228.08	1228.08
Máximo	407500.56	34564.65

Tabla 4. Estadísticos de la severidad antes y después del truncado.
La tabla muestra cómo el truncado afecta principalmente a los valores extremos de la distribución de la severidad. Los estadísticos centrales y de posición (cuartiles y mediana) permanecen inalterados. La media y, especialmente el valor máximo se reducen de forma significativa.

La severidad truncada sigue presentando una distribución asimétrica (Figura 2.B) cuando se analiza en escala original. Este comportamiento es habitual en el ámbito asegurador. Al aplicar una transformación logarítmica, esta asimetría se reduce de forma notable, dando lugar a una distribución más regular y cercana a la simetría (Figura 2.C). De esta manera, se facilita tanto el análisis como la modelización.

En conjunto, el uso de la severidad media, junto con el truncado de valores extremos y la transformación logarítmica, permite reducir la influencia de siniestros excepcionales y obtener una representación más estable del coste por siniestro. Estas decisiones facilitan el análisis de la variable y mejoran las condiciones para su modelización posterior, contribuyendo a estimaciones más robustas y coherentes de la severidad siniestral.

3.2.2 Prima pura empírica

La prima pura empírica se construye como el producto entre la tasa de siniestros y la severidad media por siniestro, proporcionando una aproximación al coste técnico esperado por póliza durante el periodo analizado. Esta variable es relevante en el estudio como herramienta descriptiva. Se construye a partir de realizaciones observadas de la frecuencia y la severidad por lo que no representa una estimación actuarial del coste esperado por póliza. Su uso se limita al análisis exploratorio de la cartera y a servir como referencia para la comparación posterior con la prima pura estimada a partir de los modelos.

La prima pura empírica presenta una elevada proporción de valores nulos, lo que explica que la mediana y los cuartiles se sitúen en cero. Este resultado se debe a que la mayoría de las pólizas no ha registrado siniestros durante el periodo analizado. En cambio, el valor

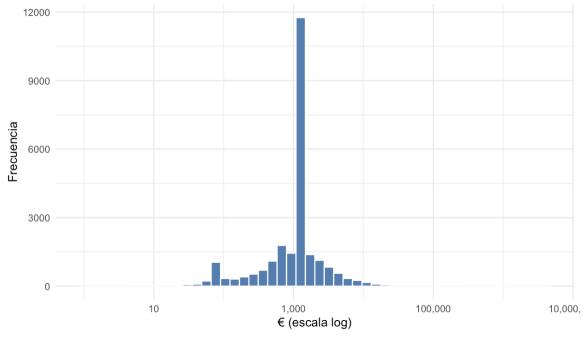
medio está influido por un número reducido de pólizas con importes elevados, que concentran una parte importante del coste total (Tabla 5).

Al analizar únicamente las pólizas con prima pura positiva, se observa una distribución claramente asimétrica, con muchos valores bajos y una cola derecha prolongada asociada a importes más altos. Este comportamiento refleja la diversidad en los costes del riesgo dentro de la cartera y pone de manifiesto la presencia de unos pocos casos con costes elevados (Figura 2.D). Esta estructura, caracterizada por una masa puntual en cero y una parte continua positiva asimétrica motiva el uso posterior de modelos compuestos para la modelización del coste total.

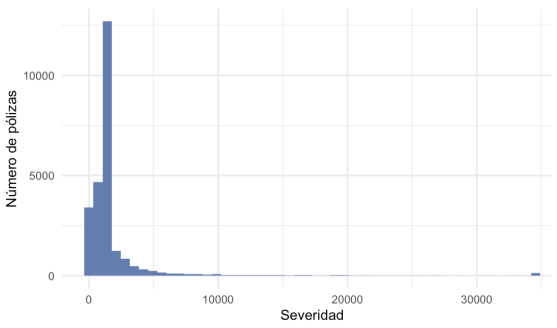
Estadístico	Valor
Valor mínimo	0.0000
Primer cuartil	0.0000
Mediana	0.0000
Media	383.2732
Tercer cuartil	0.0000
Valor máximo	18524548.0000

Tabla 5. Estadísticos descriptivos de la prima pura empírica.
Los estadísticos de posición se sitúan en cero. En contraste, la media y el valor máximo se ven influidos por un número reducido de pólizas con costes elevados.

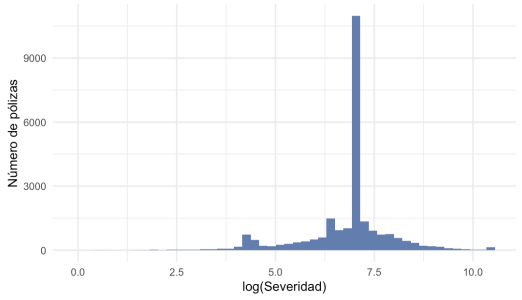
A



B



C



D

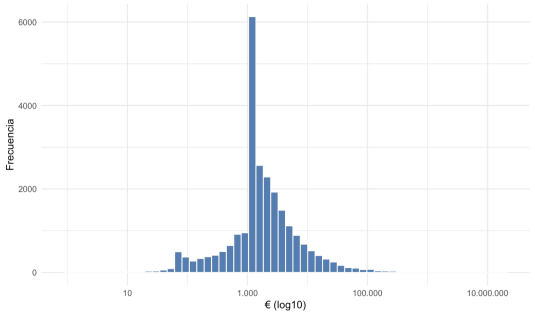


Figura 2. Análisis exploratorio de la severidad de los siniestros.
(A) Histograma del importe medio por siniestro en escala logarítmica. Fuerte concentración de observaciones en torno a valores intermedios y una dispersión creciente hacia importes más elevados. (B) Distribución de la severidad en escala original. Se caracteriza por una marcada asimetría a la derecha, con una elevada frecuencia de siniestros de bajo coste y una cola larga asociada a importes elevados. (C) Distribución de la severidad tras la transformación logarítmica. Reducción de la asimetría de forma notable y valores concentrados de manera más uniforme alrededor del centro de la distribución. (D) Distribución del importe medio por siniestro. Amplitud del rango de valores y concentración predominante de observaciones en un intervalo acotado.

3.3 Variables del conductor

3.3.1 Edad del conductor

La variable edad nos aporta información sobre el año de nacimiento del asegurado principal asociado a cada póliza. Se trata de una de las variables explicativas más relevantes en el estudio de la siniestralidad, ya que la experiencia al volante suele estar estrechamente relacionada con la probabilidad de ocurrencia de siniestros.

En la cartera analizada, la distribución de la edad muestra una mayor concentración de conductores en edades intermedias, especialmente entre los 30 y los 55 años. Por el contrario, los extremos del rango de edad muestran una presencia más limitada (Figura 3.A).

Al analizar conjuntamente la edad y la exposición, se observa que el volumen asumido por la aseguradora es máximo en los tramos de edad intermedia, lo que indica que estos grupos constituyen el núcleo principal de la cartera (Figura 3.B). Por otro lado, la relación entre la edad del conductor y la tasa media de siniestros presenta un patrón no lineal, con niveles más elevados en conductores de menor y mayor edad (Figura 3.C).

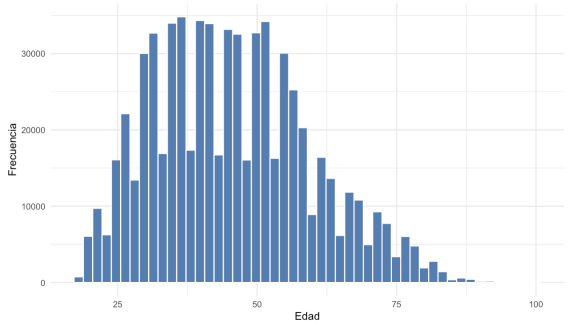
A partir de estos resultados, se propone una categorización en cuatro grupos de edad: (i) Jóvenes ($\leq$ 25 años): grupo con menor exposición y la tasa media de siniestralidad más alta. (ii) Adultos jóvenes (26-40 años): grupo con aumento de exposición y notable descenso del riesgo. (iii) Adultos (41-60 años): tramo de máxima estabilidad tanto en

número de conductores como en siniestralidad. (iv) Mayores (≥ 60 años): exposición decreciente y repunte del riesgo de siniestros.

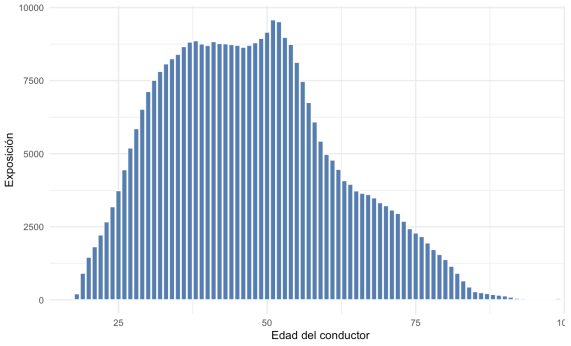
La distribución de la cartera según estas categorías, muestra que los adultos concentran la mayor proporción de pólizas, seguidos de los adultos jóvenes, mientras que los conductores jóvenes y mayores representan una fracción menor del total (Figura 3.D).

Finalmente, para completar el análisis, se estudió la relación entre los grupos de edad y las dos dimensiones del riesgo: la frecuencia y la severidad. Desde el punto de vista de la frecuencia, la tasa media de siniestros muestra una relación decreciente con la edad del conductor, registrando los valores más elevados en los conductores jóvenes y reduciéndose progresivamente en los tramos de mayor edad (Figura 3.E). En cuanto a la severidad, el importe medio por siniestro no presenta diferencias marcadas entre los distintos tramos de edad, lo que sugiere que la edad del conductor tiene un efecto más relevante sobre la frecuencia de ocurrencia de siniestros que sobre su coste medio (Figura 3.F).

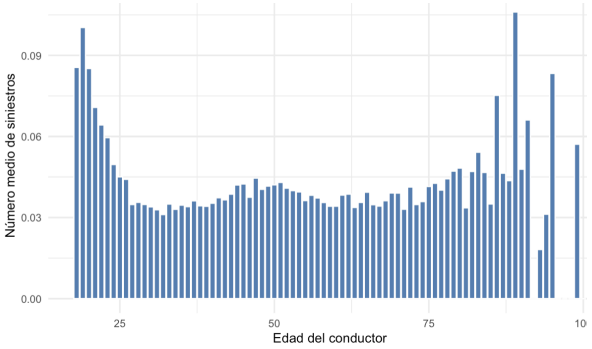
A



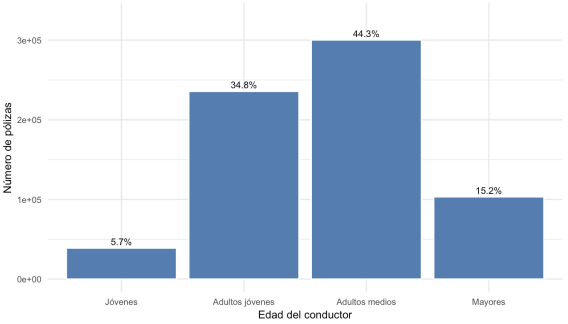
B



C



D



E

F

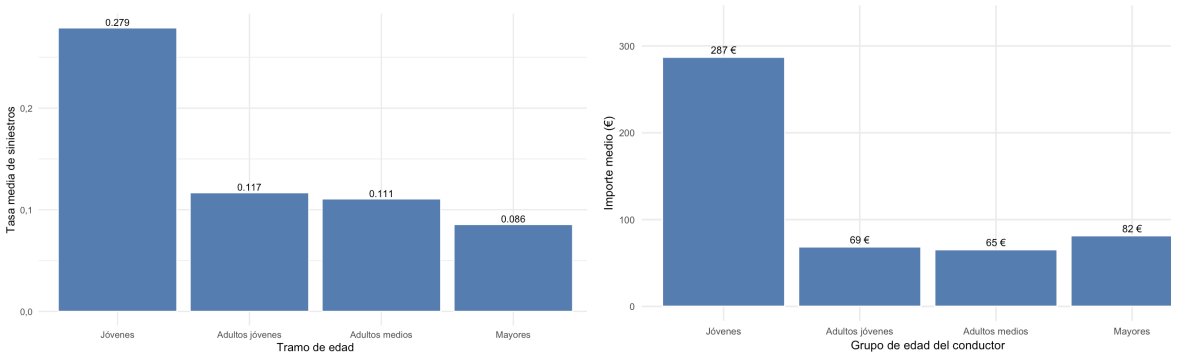


Figura 3. Análisis exploratorio de la variable edad del conductor.
(A) Distribución de la edad del conductor. Se aprecia una mayor concentración de pólizas en edades intermedias y una menor presencia en los tramos más jóvenes y de mayor edad. (B) Distribución de la exposición según la edad del conductor. El volumen de riesgo asumido se concentra principalmente en los tramos de edad intermedia. (C) Relación entre la edad del conductor y el número medio de siniestros. Se observan valores más elevados en edades jóvenes, una reducción progresiva en edades intermedias y un aumento de la variabilidad en edades avanzadas. (D) Distribución del número de pólizas por grupos de edad, destacando la predominancia de los conductores adultos jóvenes y adultos medios en la cartera. (E) Tasa media de siniestros por tramo de edad, con valores claramente más elevados en conductores jóvenes y una disminución progresiva a medida que aumenta la edad. (F) Importe medio por siniestro según el grupo de edad del conductor. Valores similares entre los distintos tramos y un importe más elevado en el grupo de conductores jóvenes.

3.4 Características del vehículo

3.4.1 Edad del vehículo

La antigüedad del vehículo puede influir tanto en la probabilidad de sufrir un siniestro como en el coste, ya que los vehículos más nuevos suelen incorporar sistemas más avanzados, mientras que los más antiguos pueden presentar mayor desgaste mecánico. Es por eso que la variable explicativa edad del vehículo es relevante para nuestro estudio de siniestralidad.

En la base de datos original se identificaron algunas observaciones con edades superiores a 30 años. Estas observaciones se consideraron poco habituales y posibles errores de registro. Para solucionarlo se decidió homogeneizar fijando su edad en 30 años, para evitar así distorsiones en el análisis. La distribución resultante muestra una mayor concentración de pólizas en vehículos entre los 0 y 5 años. A partir de ese punto, el número de pólizas disminuye de forma progresiva conforme aumenta la antigüedad, aunque se observa un pico en los vehículos de 10 y 15 años (Figura 4.A).

Para simplificar el análisis y facilitar la interpretación de los resultados, agrupamos la variable edad del vehículo en tres categorías (Figura 4.B): (i) Nuevos: vehículos entre 0 y 5 años. (ii) Medios: vehículos entre 6 y 12 años. (iii) Viejos: vehículos a partir de los 13 años.

Con el fin de completar el análisis descriptivo, estudiamos la variable edad del vehículo con las dos dimensiones de riesgo que modelizaremos posteriormente. Esta comparación

permite identificar patrones relevantes y comprobar si la antigüedad del vehículo presenta variaciones significativas en el comportamiento siniestral.

Al analizar la relación entre la edad del vehículo y la frecuencia de siniestros, observamos que los vehículos nuevos registran la frecuencia más baja de siniestralidad, mientras que los de antigüedad media presentan el valor más elevado (Figura 4.C). Por otro lado, al relacionar la variable con la severidad, apreciamos una tendencia creciente en el importe medio de siniestro por categorías. Esto indica que, aunque la frecuencia es mayor en los vehículos de antigüedad intermedia, los vehículos más antiguos tienden a generar siniestros más costosos (Figura 4.D).

3.4.2 Potencia del vehículo

La potencia del vehículo es una variable que puede estar relacionada con distintos patrones de conducción y, potencialmente, con el riesgo asegurado. Los vehículos con mayor potencia suelen asociarse a una conducción más deportiva y peligrosa, lo que puede incrementar la probabilidad y el coste de los siniestros. Por el contrario, los vehículos de baja potencia tienden a asociarse con un uso más urbano y familiar.

Al analizar la distribución de la potencia del vehículo se observa una mayor concentración en los niveles con potencia media-baja (4-7). En cambio, los vehículos con potencias más elevadas presentan una frecuencia considerablemente menor (Figura 4.E).

Al analizar la relación entre la potencia del vehículo y la tasa de siniestros, se observa una ligera tendencia creciente en la tasa media de siniestralidad, aunque con oscilaciones entre niveles de potencia que impiden identificar un patrón claramente monotónico (Figura 4.F). En cambio, el número medio de siniestros se mantiene relativamente estable a lo largo de los distintos niveles de potencia, sin mostrar una tendencia definida (Figura 4.G). Estos resultados indican que la potencia del vehículo no constituye por sí sola un factor determinante del riesgo de la siniestralidad.

Para mejorar la interpretación y facilitar su incorporación en la posterior modelización, se ha considerado conveniente categorizar la variable en 4 niveles representativos: (i) Vehículos de baja potencia, menores que 6. (ii) Vehículos de potencia media, entre 6 y 9. (iii) Vehículos de potencia alta, entre 10 y 12. (iv)Vehículos de potencia muy alta, mayores que 12.

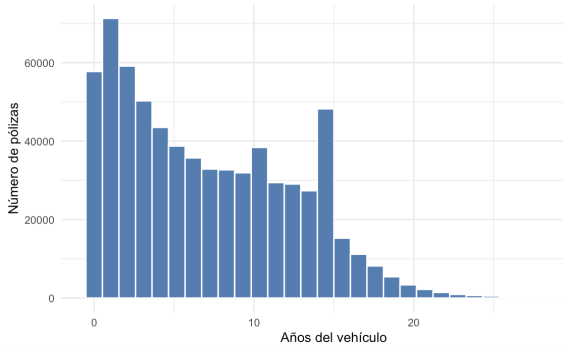
3.4.3 Tipo de combustible

El tipo de combustible del vehículo puede estar relacionado con distintas características técnicas y patrones de uso. En la cartera analizada, los vehículos de gasolina y diésel presentan una distribución muy equilibrada: el 49% de los vehículos utilizan diésel y el 51% restante gasolina.

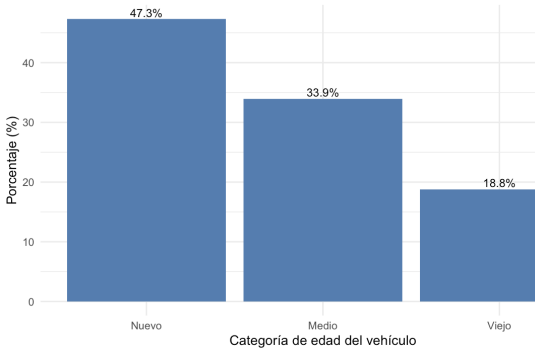
Dado que los vehículos diésel suelen asociarse a un mayor nivel de riesgo, se analiza su relación con la frecuencia y la severidad de los siniestros en la muestra. Los resultados muestran que los vehículos diésel presentan un número medio de siniestros y una tasa media de siniestralidad superior a los de gasolina (Figura 4.H y Figura 4.I). Aunque la diferencia es reducida (y podría estar medida por otras características del riesgo), este patrón sugiere que los vehículos diésel tienden a presentar una mayor probabilidad de ocurrencia de siniestros. En cuanto a la severidad, observamos que el importe medio del siniestro es superior en vehículos de gasolina (Figura 4.J). Esto nos indica que aunque los vehículos de diésel tienen una frecuencia superior, los vehículos de gasolina tienden a ser más costosos.

Por lo tanto, el tipo de combustible puede considerarse una variable explicativa secundaria, que aporta información complementaria al analizar el riesgo.

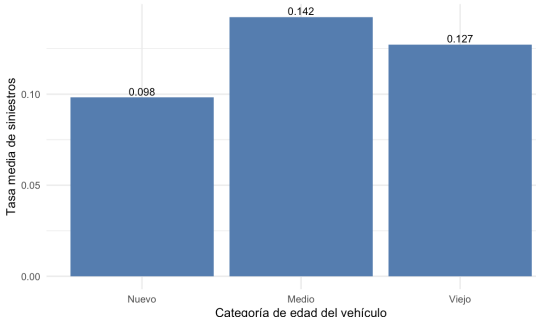
A



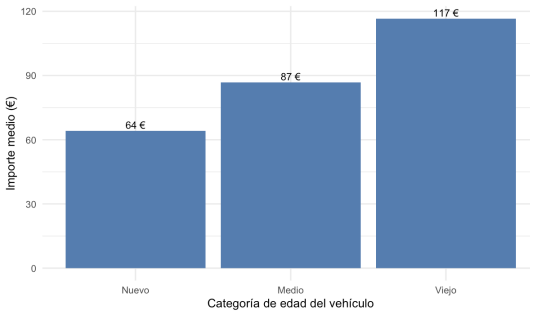
B



C



D



E

F

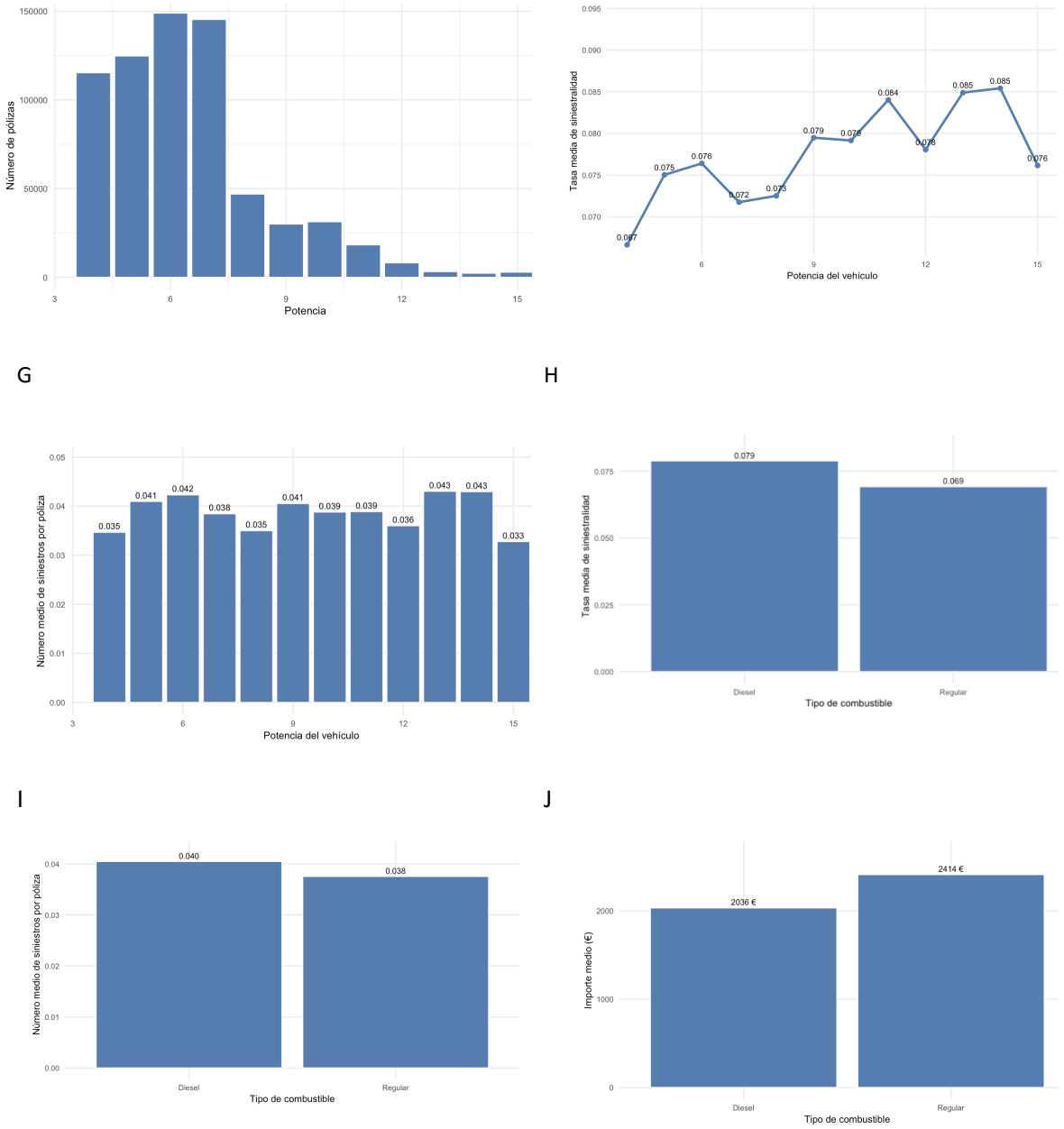


Figura 4. Análisis exploratorio de la edad del vehículo, la potencia y el tipo de combustible.
(A) Distribución de la edad del vehículo, con mayor concentración de pólizas en vehículos más recientes. (B) Distribución porcentual de las pólizas por categoría de edad del vehículo. (C) Tasa media de siniestros según la categoría de edad del vehículo. Valores más elevados en los vehículos de antigüedad media y antigua en comparación con los vehículos nuevos. (D) Importe medio por siniestro por categoría de edad del vehículo. Incremento del coste medio a medida que aumenta la antigüedad del vehículo. (E) Distribución de la potencia del vehículo, concentrada en niveles bajos y medios. (F) Tasa media de siniestralidad en función de la potencia del vehículo, con una tendencia general ligeramente creciente. (G) Número medio de siniestros por póliza según la potencia del vehículo, con variaciones moderadas. (H) Tasa media de siniestralidad por tipo de combustible, con diferencias reducidas entre categorías. (I) Número medio de siniestros por póliza por tipo de combustible. (J) Importe medio por siniestro según el tipo de combustible.

3.5 Variables tarifarias

3.5.1 Bonus-malus

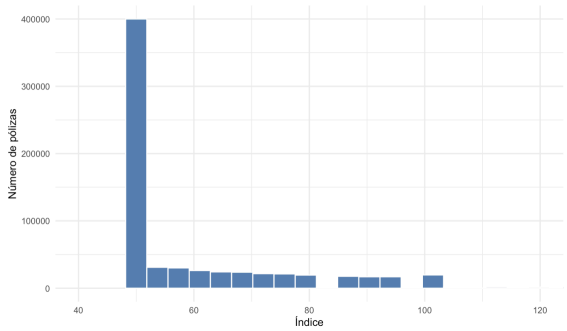
El índice bonus-malus recoge el historial previo de siniestralidad del asegurado y constituye una variable tarifaria fundamental en el análisis del riesgo. Inicialmente, todos los asegurados comienzan con un índice igual a 100, que corresponde a la tarifa base de la póliza. Con el tiempo, si el conductor no registra ningún siniestro, su índice puede descender por debajo de 100, lo que significa que por su buen comportamiento tiene derecho a una reducción de prima. Por el contrario, si el conductor tiene uno o varios accidentes, este índice puede superar el valor de 100, y en vez de tener una bonificación se le aplica un recargo de prima.

El análisis exploratorio muestra una elevada concentración de pólizas en niveles de bonificación, especialmente en los valores más bajos del índice, lo que es coherente con la baja frecuencia de siniestros observada en la cartera (Figura 5.A). Por el contrario, los niveles asociados a penalizaciones presentan una presencia mucho más reducida.

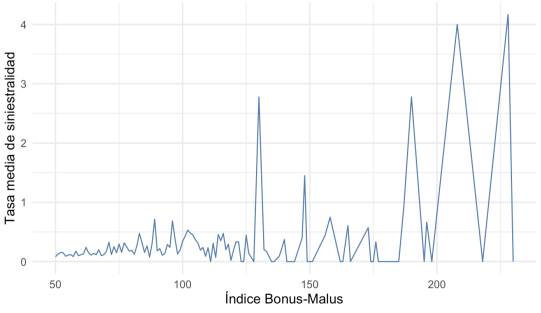
La relación entre el índice bonus-malus y la frecuencia de siniestros muestra una asociación positiva, es decir, a medida que aumenta el valor del índice, también lo hace la tasa media de siniestros (Figura 5.B). En relación con la severidad, el importe medio por siniestro aumenta ligeramente en los niveles más altos del índice, aunque este efecto es menos marcado que el observado en la frecuencia (Figura 5.C). Este resultado indica que el bonus-malus está más relacionado con la probabilidad de que ocurran siniestros que con el coste medio de los mismos. Finalmente, si lo relacionamos con la edad del conductor se pone de manifiesto una relación decreciente, con valores del índice más elevados en conductores jóvenes y una estabilización progresiva a partir de edades intermedias (Figura 5.D)

En conjunto, los resultados observados confirman que el índice bonus-malus constituye una variable clave para explicar el comportamiento siniestral de la cartera, especialmente en términos de frecuencia, y justifican su posterior inclusión en los modelos realizados.

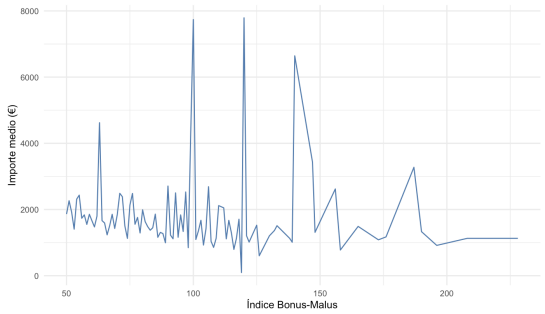
A



B



C



D

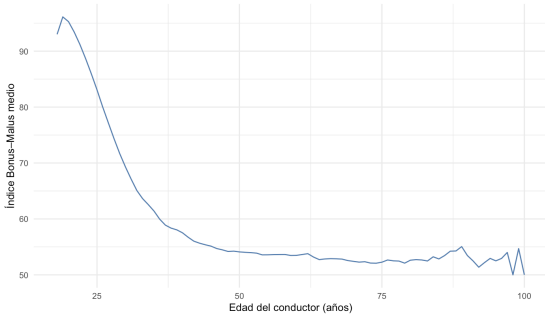


Figura 5. Comportamiento del índice bonus-malus en la cartera.
(A) Distribución del índice bonus-malus en la cartera. Elevada concentración de pólizas en los niveles de bonificación y una presencia decreciente a medida que aumenta el valor del índice. (B) Relación entre el índice bonus-malus y la tasa media de siniestralidad. Aumento general de la tasa conforme se incrementa el índice, con mayor variabilidad en los niveles más elevados. (C) Relación entre el índice bonus-malus y el importe medio por siniestro, caracterizada por una elevada dispersión y la presencia de valores extremos, especialmente en los niveles más altos del índice. (D) Relación entre el índice bonus-malus medio y la edad del conductor, mostrando valores más elevados del índice en edades jóvenes y una estabilización progresiva a partir de edades intermedias.

4. Modelización y resultados

Una vez completado el análisis exploratorio de las variables, pasamos a analizar la relación entre estos factores y la siniestralidad mediante técnicas de modelización. El objetivo es identificar qué características del asegurado y del vehículo influyen en la frecuencia y el coste de los siniestros para obtener estimaciones estables que permitan aproximar la prima pura. Para ello se emplean modelos lineales generalizados que nos permiten trabajar con distribuciones no normales y ajustar adecuadamente la exposición de cada póliza.

4.1 Modelización de la frecuencia siniestral

La modelización de la frecuencia siniestral constituye el primer paso en la estimación de la prima pura. La variable frecuencia es de naturaleza discreta y depende directamente del tiempo asegurado. Consecuentemente, resulta adecuado emplear un modelo de regresión de conteo con un término *offset* que incorpore la exposición. Con este planteamiento podemos estimar la frecuencia ajustada al tiempo en riesgo y analizar cómo las distintas características del asegurado y del vehículo influyen en la siniestralidad.

En el análisis exploratorio observamos que la distribución del número de siniestros presenta una distribución fuertemente asimétrica centrada en cero y una cola hacia valores positivos. Además, la media y la varianza del número de siniestros resultaron ser muy similares, como es de esperar bajo una distribución de Poisson.

En este modelo se utiliza una función enlace logarítmica. Dicho enlace garantiza la positividad de la variable respuesta y permite una interpretación multiplicativa de los efectos de las variables explicativas. Además, facilita la inclusión del tiempo de exposición al riesgo con un término de *offset*, expresando la frecuencia esperada como una tasa ajustada por exposición. Finalmente, la especificación explícita del modelo Poisson adoptado tras incorporar el término *offset* y el enlace logarítmico queda expresado como:

$$N_i \sim Poisson(\mu_i),\ log(\mu_i) = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_p X_{p,i} + log(e_i) + \varepsilon_i$$

Las variables incorporadas han sido seleccionadas según si en el análisis exploratorio mostraban relación relevante con la siniestralidad: (i) edad del conductor, (ii) antigüedad del vehículo, (iii) potencia del motor, (iv) tipo de combustible y, (v) índice bonus-malus. Además, se llevó a cabo un análisis específico para determinar el predictor lineal más adecuado, evaluando distintas combinaciones de variables y posibles transformaciones (Anexo 1). Una vez definidas las covariables seleccionadas y sus correspondientes categorías de referencia, la especificación explícita del modelo Poisson adoptado queda expresada como:

$$\log(\mu_i) = \beta_0 + \beta_1\ 1(\text{Adultos jóvenes}_i) + \beta_2\ 1(\text{Adultos medios}_i) + \beta_3\ 1(\text{Mayores}_i) +$$
$$\beta_4\ 1(\text{Edad del vehículo}_i) + \beta_5\ 1(\text{Edad del vehículo}_i^2) +$$
$$\beta_6\ \text{Potencia}_i + \beta_7\ 1(\text{Combustible regular}_i) + \beta_8\ Potencia \cdot 1(\text{Combustible regular}_i) +$$
$$\beta_9\ 1(\text{bonus-malus}_i) + \epsilon_i$$

El intercepto del modelo (β_0) fija el nivel base de la frecuencia siniestral. Este término corresponde a una póliza de referencia, definida por las categorías base de las variables cualitativas y una exposición unitaria. Su función es establecer el punto de partida a partir del cual las distintas variables explicativas aumentan o reducen la frecuencia esperada. Las variables categóricas incluidas en el modelo se han elegido para representar los perfiles más habituales de la cartera. En concreto, se ha tomado como referencia a los conductores jóvenes y el combustible diesel.

Tras ajustar el modelo Poisson, se evaluó el parámetro de dispersión con valor superior a 1.6, indicando una variabilidad mayor a la permitida bajo una distribución Poisson. Esta sobredispersión puede afectar a la estimación de los errores estándar y a la estabilidad de los coeficientes. Cuando se detecta, el modelo binomial negativo constituye una alternativa más flexible al modelo Poisson para la modelización de la frecuencia de siniestros [6]. Por lo tanto, con el fin de disponer de un modelo más robusto frente a esta variabilidad, se implementó el modelo Binomial Negativo (Tabla 6), representado por la siguiente fórmula:

$$N_i \sim BN(\mu_i, \theta),\ log(\mu_i) = \beta_0 + \beta_1 X_{1,i} + \ldots + \beta_p X_{p,i} + log(e_i)$$

Variable	Beta	Estimación	Error estándar	Valor z	p-valor
(Intercepto)	β_0	-4.4839	0.0547	-82.04	<0.001
Conductores jóvenes adultos	β_1	-0.0784	0.0238	-3.04	<0.001
Conductores adultos	β_2	0.2308	0.0270	8.53	<0.001
Conductores mayores	β_3	0.1325	0.0306	4.34	<0.001
Edad del vehículo	β_4	0.0315	0.0038	8.33	<0.001
Edad del vehículo ^{2}	β_5	-0.0025	0.0002	-12.04	<0.001
Potencia	β_6	0.0184	0.0051	3.64	<0.001
Combustible regular	β_7	-0.2969	0.0438	-6.70	<0.001
Interacción potencia con combustible regular	β_8	0.0251	0.0065	3.89	<0.001
Bonus-malus	β_9	0.0279	0.0004	74.84	<0.001

Tabla 6. Estimaciones del modelo Binomial Negativo para la frecuencia siniestral.
Todas las variables incluidas en el predictor lineal son estadísticamente significativas (p-valor < 0.05). Estos resultados sugieren que el modelo capta adecuadamente la estructura del riesgo en la cartera. La columna *estimación* muestra el valor del coeficiente, el *error estándar* y su precisión, el *valor z* el contraste de significación y el *p-valor* el nivel de significación estadística.

En el análisis descriptivo (Figura 3.D) se observó que los conductores jóvenes (grupo 1) y mayores (grupo 4) presentaban una mayor frecuencia de siniestros. En cambio, en el modelo ajustado (Tabla 6), el patrón cambia al tener en cuenta el efecto conjunto del resto de las variables. Los signos de los coeficientes estimados muestran que los conductores adultos (grupo 3) y mayores (grupo 4) presentan coeficientes positivos, lo que indica un aumento estadísticamente significativo de la frecuencia esperada de siniestros respecto al grupo de referencia. Por el contrario, los conductores jóvenes adultos (grupo 2) presentan un coeficiente negativo, reflejando una ligera reducción de la frecuencia esperada una vez analizado el ajuste. Estas diferencias respecto al análisis descriptivo ponen de manifiesto que parte del mayor riesgo observado en los conductores jóvenes estaba asociado a otras características del riesgo que ahora quedan controladas en el análisis multivariante.

La combinación de las variables explicativas referentes a la edad del vehículo nos sugiere una relación no lineal entre la antigüedad del vehículo y la siniestralidad (Figura 6). El coeficiente positivo estimado para la edad del vehículo indica que la frecuencia esperada de siniestros aumenta inicialmente con la antigüedad. No obstante, el signo negativo que

se estima en el término cuadrático, señala que este efecto se ve atenuado a medida que el vehículo envejece. Este patrón es coherente desde un punto de vista actuarial, ya que los vehículos más nuevos suelen beneficiarse de sistemas más avanzados, los de antigüedad media suelen presentar un mayor uso y exposición, y los más antiguos tienden a circular menos.

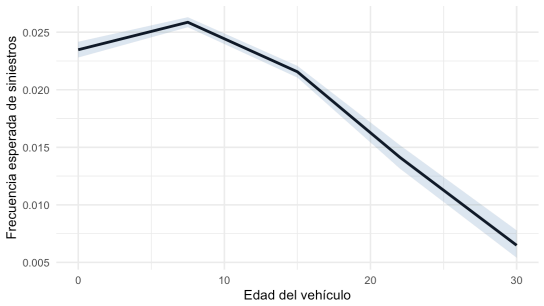


Figura 6. Efecto no lineal de la edad del vehículo sobre la frecuencia siniestral.
La curva representa la frecuencia esperada de siniestros estimada por el modelo Binomial Negativo en función de la antigüedad del vehículo, manteniendo constantes el resto de variables. Se observa un incremento inicial seguido de un descenso progresivo para vehículos de mayor antigüedad.

En cuanto al resto de variables relacionadas con las características del vehículo, la potencia presenta un coeficiente positivo, lo que indica que los vehículos más potentes tienen una mayor frecuencia de siniestros. Por su parte, el uso del combustible regular, se relaciona, en promedio con una menor frecuencia de siniestros respecto al combustible de referencia. No obstante, la interacción entre potencia y tipo de combustible indica que el efecto de la potencia sobre la siniestralidad es más acusado en los vehículos que utilizan combustible regular, lo que pone de manifiesto la conveniencia de interpretar ambos efectos de forma conjunta.

Por último, la variable bonus-malus presenta un coeficiente positivo que indica que un mayor valor de este índice se traduce en un aumento sistemático de la frecuencia esperada de siniestros. Este resultado es coherente, ya que el bonus-malus resume el historial previo de siniestralidad del conductor. Por otro lado, el valor del parámetro de dispersión $\theta = 1.0485$ refleja que el modelo capta adecuadamente la sobredispersión presente en los datos.

Los resultados del diagnóstico confirman la adecuación del modelo. Los residuos no presentan patrones sistemáticos ni indicios de una mala especificación del predictor lineal (Figura 7.A). Se observan algunos residuos con valores elevados que corresponden a pólizas que registran varios siniestros (Figura 7.A y Figura 7.B). Estos casos corresponden a situaciones poco frecuentes dentro de la cartera y sin impacto relevante sobre el ajuste global . El comportamiento de la varianza residual se mantiene estable (Figura 7.C), y no se detectan observaciones con una influencia excesiva en la estimación de los coeficientes (Figura 7.D). Por lo tanto, en conjunto los resultados respaldan la solidez del modelo.

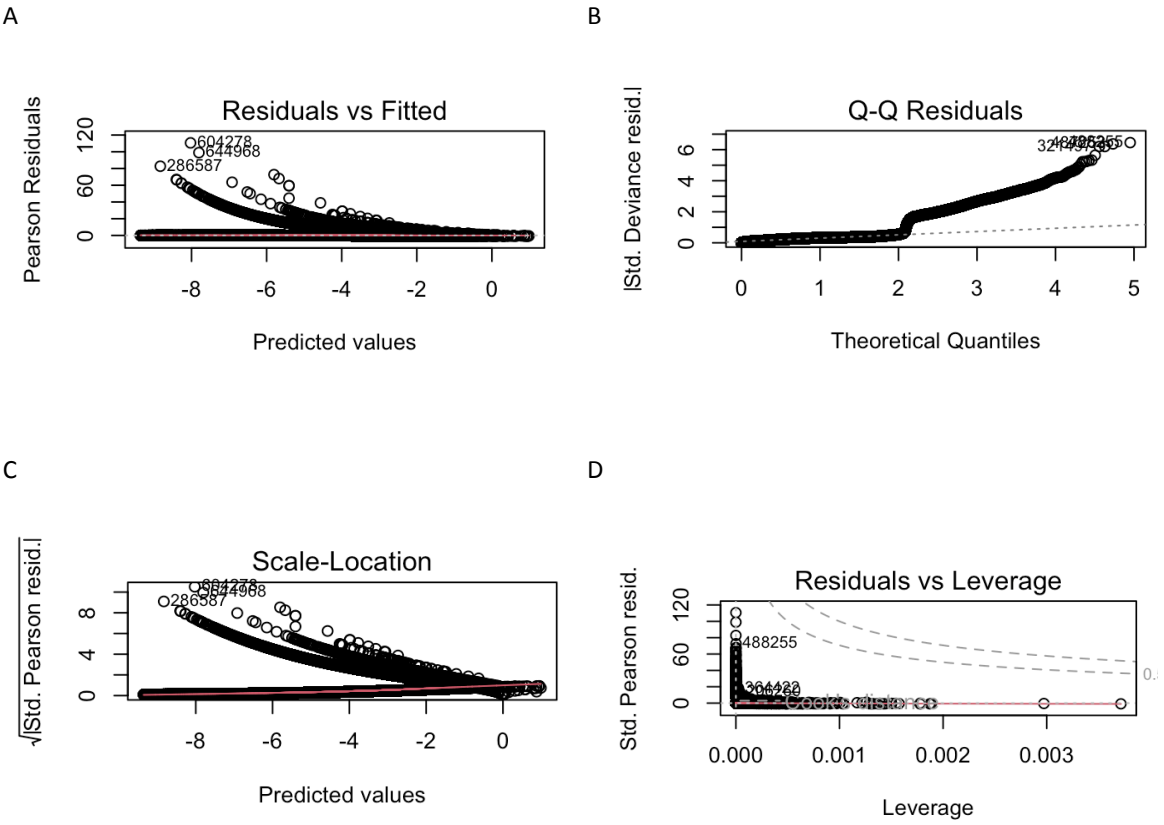


Figura 7. Evaluación gráfica del modelo Binomial Negativo.
(A) *Residuos vs. valores ajustados.* Puntos concentrados alrededor del eje horizontal, con algunos valores extremos asociados a pólizas con varios siniestros. Ausencia de patrones sistemáticos. (B) *Q-Q de residuos*. Cuantiles centrales cerca de la recta teórica. Se observan desviaciones en el cuantil superior, asociados a residuos positivos de mayor magnitud. (C) *Scale-location.* Dispersión de los residuos relativamente homogénea a lo largo de los valores ajustados, sin incrementos sistemáticos de la variabilidad. Por último, (D) *Residuos vs. Leverage*, junto con la distancia de *Cook*. La mayoría de las observaciones se concentran en valores bajos de *leverage*, con algunos puntos aislados de mayor influencia.

Para confirmar la selección del modelo Binomial Negativo frente al modelo Poisson se calcula el valor AIC, siendo de 21102.4 y 21563.3, respectivamente. Esta diferencia entre valores es relevante y resulta indicativo de una mejora en el ajuste del modelo Binomial Negativo, sin introducir una complejidad innecesaria, y que refuerza su idoneidad para describir la frecuencia de la cartera y sustentar el posterior proceso de tarificación.

4.2 Modelización de la severidad siniestral

Una vez analizada la frecuencia de los siniestros, el siguiente paso en la estimación de la prima pura consiste en modelizar la severidad, entendida como el coste medio asociado a los siniestros ocurridos. A diferencia de la frecuencia, la severidad se analiza condicionalmente a la ocurrencia de al menos un siniestro, por lo que el conjunto de datos utilizado se restringe a aquellas pólizas con un número de siniestros positivo.

A partir de las evidencias observadas en el análisis exploratorio, se optó por modelizar la severidad mediante un modelo lineal generalizado con distribución Gamma y enlace logarítmico. El predictor lineal del modelo se especificó incorporando aquellas variables con sentido actuarial claro y que mostraron relación con la siniestralidad en el análisis exploratorio. En particular, se incluyeron la edad del conductor, la edad del vehículo, la potencia del vehículo y el índice bonus-malus.

A diferencia del modelo de frecuencia, en el que se llevó a cabo un análisis exhaustivo para determinar el predictor lineal óptimo. En el modelo de severidad se priorizó una especificación más parsimoniosa del predictor lineal, debido a que al estimarse únicamente sobre pólizas con siniestros, se reduce considerablemente el tamaño muestral. Consecuentemente, se limita la capacidad de detectar relaciones funcionales complejas sin incurrir en sobreajuste. Por este motivo, variables como la edad del conductor, la edad del vehículo o la potencia se mantuvieron en forma categorizada por tramos, tal y como se definieron en el análisis exploratorio. No obstante, sí se realizaron comparaciones puntuales mediante el criterio AIC para verificar que las variables excluidas del modelo realmente no aportan mejoras relevantes en el ajuste.

Finalmente, la especificación explícita del modelo Gamma adoptado queda expresada como:

$$S_i \mid (N_i > 0) \sim Gamma(\mu_i, \phi), \qquad log(\mu_i) = \eta_i$$

$$\begin{aligned}\log(\mu_i) = {} & \beta_0 + \beta_1\ 1(\text{Adultos jóvenes}_i) + \beta_2\ 1(\text{Adultos medios}_i) + \beta_3\ 1(\text{Mayores}_i) + \\ & \beta_4\ 1(\text{Edad del vehículo medio}_i) + \beta_5\ 1(\text{Edad del vehículo viejo}_i) + \\ & \beta_6\ 1(\text{Potencia Baja}_i) + \beta_7\ 1(\text{Potencia Media}_i) + \beta_8\ 1(\text{Potencia muy Alta}_i) + \\ & \beta_9\ 1(\text{bonus-malus}_i) + \epsilon_i\end{aligned}$$

El intercepto del modelo representa la severidad esperada asociada a una póliza de referencia, definida por conductores jóvenes, vehículos nuevos y de potencia alta y un nivel base del índice categorías seleccionadas por ser representativas de la cartera. Bajo esta especificación, se procedió a la estimación de los coeficientes (Tabla 7) mediante máxima verosimilitud ponderada, utilizando el número de siniestros como peso en la estimación.

Variable	Beta	Estimación	Error estándar	Valor z	p-valor
(Intercepto)	β_0	0.0900	0.0900	83.072	<0.001
Conductores jóvenes adultos	β_1	-0.0846	0.0491	-1.725	0.08454
Conductores adultos	β_2	-0.0733	0.0517	-1.418	0.15616
Conductores mayores	β_3	-0.0293	0.0579	-0.506	0.61312
Vehículos de antigüedad media	β_4	-0.0902	0.0266	-3.385	<0.001
Vehículo de mayor antigüedad	β_5	-0.1127	0.0347	-3.251	0.00115
Vehículos de potencia baja	β_6	-0.1150	0.0463	-2.484	0.01298
Vehículos de potencia media	β_7	-0.0502	0.0445	-1.128	0.25922
Vehículos de potencia muy alta	β_8	0.0678	0.1142	0.593	0.55296
Bonus-malus	β_9	0.0023	0.0007	3.191	0.00142

Tabla 7. Estimaciones del modelo Gamma para la severidad siniestral.
No todas las variables incluidas en el predictor lineal presentan un efecto estadísticamente significativo sobre la severidad esperada del siniestro. En particular, el intercepto, los vehículos de antigüedad media y alta, los vehículos de potencia baja y el índice bonus-malus resultan significativas (p-valor<0.05). La columna *estimación* muestra el valor del coeficiente, el *error estándar* su precisión, el *valor z* el contraste de significación y el *p-valor* el nivel de significación estadística.

A diferencia del modelo de frecuencia, el modelo de severidad presenta un menor número de variables con efecto estadísticamente significativo sobre el coste medio por siniestro. Este resultado es coherente con lo observado en el análisis exploratorio y sugiere que la variabilidad del coste está menos explicada por las características del asegurado y del vehículo que la probabilidad de ocurrencia de los siniestros.

Ninguno de los coeficientes referentes a la edad del conductor resultan estadísticamente significativos. Este resultado sugiere que, una vez controlado el resto de variables, la edad del conductor no muestra un efecto estadísticamente significativo en la magnitud económica de los siniestros, a diferencia de lo observado en el modelo de frecuencia, donde su impacto era más relevante.

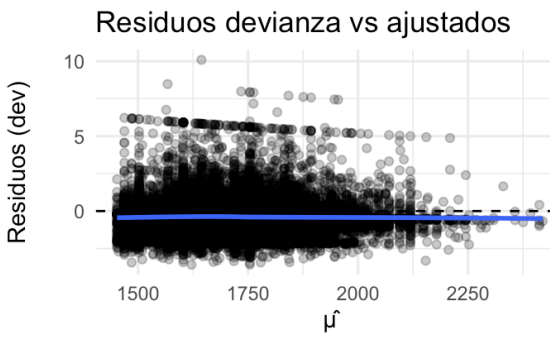
La edad del vehículo en cambio, presenta un efecto claro y estadísticamente significativo sobre la severidad. Tanto los vehículos de antigüedad media como vehículos viejos muestran una reducción significativa del coste medio por siniestro en comparación con el grupo de referencia. Este comportamiento es coherente, ya que los vehículos más antiguos suelen presentar menores valores de mercado y, en consecuencia, coste de reparación e indemnización más reducidos.

Respecto a la potencia del vehículo, los resultados indican que los vehículos de potencia baja presentan una severidad significativamente inferior, mientras que los grupos de potencia media y muy alta no muestran diferencias estadísticamente significativas. Este patrón sugiere que el efecto de la potencia sobre la severidad no es estrictamente creciente y que su influencia puede verse condicionada por otros factores.

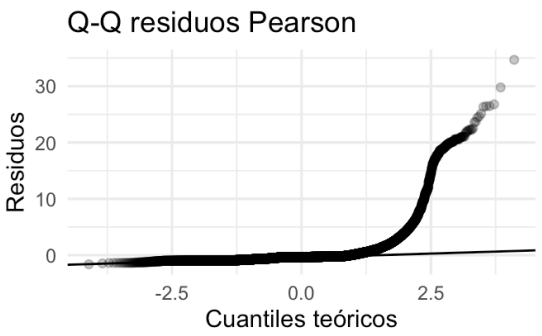
Por último, el índice bonus-malus muestra un efecto positivo y estadísticamente significativo. Un incremento en el nivel bonus-malus se asocia con un aumento del coste medio por siniestro. Este resultado indica que los asegurados con peor historial tienden a generar siniestros de mayor coste.

Los resultados del diagnóstico indican que el modelo presenta un comportamiento globalmente adecuado. No se identifican patrones sistemáticos en los residuos ni indicios de una mala especificación del predictor lineal, lo que respalda la estabilidad del ajuste a lo largo del rango de severidad estimada (Figura 8.A). La presencia de algunos residuos de gran magnitud se asocia a siniestros de coste elevado, poco frecuentes en la cartera, cuya influencia se refleja en las desviaciones observadas en los cuantiles superiores de la distribución de los residuos de Pearson (Figura 8.B). En el análisis de influencia se identifican dos observaciones relativamente más influyentes; no obstante, sus valores no alcanzan niveles que comprometan la estabilidad de las estimaciones (Figura 8.C). Finalmente, la comparación entre valores observados y esperados por deciles muestra una calibración adecuada, con ratios próximos a la unidad en todo el rango de severidad prevista, lo que refuerza la idoneidad del modelo (Figura 8.D).

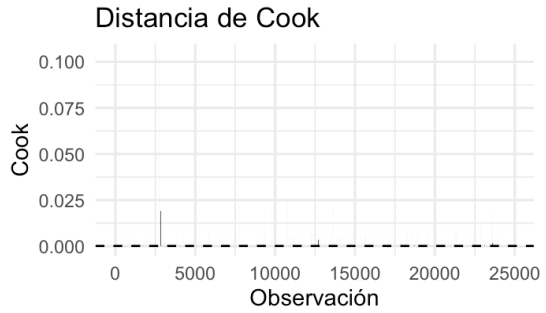
A



B



C



D

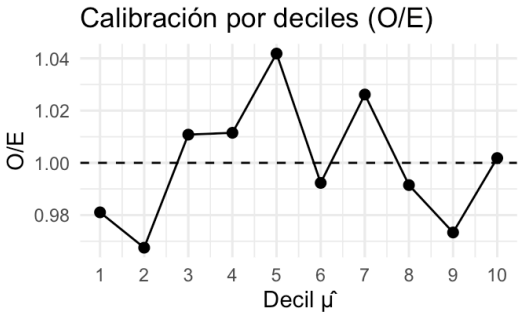


Figura 8. Evaluación gráfica del modelo Gamma.
(A) *Residuos de devianza vs. valores ajustados.* Puntos concentrados alrededor de cero, con mayor densidad en el rango central de μ. La línea suavizada (azul) presenta una leve pendiente descendente, indicando un cambio gradual en el comportamiento promedio de los residuos a lo largo del rango de valores ajustados. (B) *Q-Q de residuos*. Los puntos siguen la recta en la zona central y se separan en los cuantiles superiores, mostrando mayor dispersión en la cola derecha. Este comportamiento es esperable para este tipo de datos. (C) *Distancia de Cook.* Observaciones generalmente cercanas a cero, con algún valor ligeramente superior. Por último, (D) *Calibración por deciles (O/E)*. Los valores oscilan alrededor de la referencia O/E=1, con variaciones moderadas entre deciles.

Estos resultados muestran que el modelo Gamma proporciona una representación adecuada y estable de la severidad siniestral, identificando de forma consistente los factores que influyen en el coste medio de los siniestros. De este modo, el análisis de la severidad completa el estudio de la frecuencia y permite avanzar hacia una estimación coherente de la prima pura.

4.3 Modelo compuesto y distribución de Tweedie

Una vez abordada la modelización de la siniestralidad en dos etapas, estimando por separado la frecuencia de siniestros y la severidad media condicionada a la ocurrencia de al menos un siniestro, pasamos a modelizar ambas de manera conjunta mediante un modelo Tweedie [9].

En línea con las decisiones adoptadas en la modelización de la frecuencia y la severidad, el coste anual por póliza se modeliza con función de enlace logarítmica, que nos permite garantizar la positividad de los valores ajustados y nos facilita la interpretación multiplicativa de los coeficientes. Asimismo, y de forma consistente con el modelo de frecuencia, se incorpora un término *offset*, correspondiente al logaritmo de la exposición temporal de cada póliza. Con este término el modelo estima el coste esperado por unidad de exposición, y no simplemente el coste bruto anual.

Formalmente, la especificación explícita del modelo Tweedie adoptado queda expresada como:

$$log(\mu_i) = \beta_0 + \sum_k \beta_k x_{ik} + log(Exposición_i) + \varepsilon_i,$$

Una vez incorporadas las variables de interés consideradas en el estudio, la especificación anterior puede expresarse de forma más detallada como:

$$\begin{aligned}\log(\mu_i) = {} & \beta_0 + \beta_1\ 1(\text{Adultos jóvenes}_i) + \beta_2\ 1(\text{Adultos medios}_i) + \beta_3\ 1(\text{Mayores}_i) + \\ & \beta_4\ 1(\text{Edad del vehículo medio}_i) + \beta_5\ 1(\text{Edad del vehículo viejo}_i) + \\ & \beta_6\ 1(\text{Potencia Baja}_i) + \beta_7\ 1(\text{Potencia Media}_i) + \beta_8\ 1(\text{Potencia muy Alta}_i) + \\ & \beta_9\ 1(\text{bonus-malus}_i) + \beta_10\ 1(\text{Combustible regular}_i) + \epsilon_i\end{aligned}$$

Tras ajustar el modelo Tweedie, obtenemos los coeficientes estimados (Tabla 8) y un parámetro de potencia estimado de 1.496, valor comprendido entre 1 y 2 que confirma la idoneidad de la distribución Poisson-Gamma compuesta para modelizar el coste total anual por póliza.

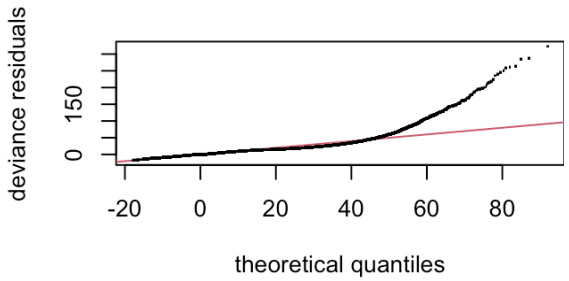
Variable	Beta	Estimación	Error estándar	Valor z	p-valor
(Intercepto)	β_0	3.5631	0.0697	51.145	<0.001
Conductores jóvenes adultos	β_1	-0.2792	0.0376	-7.424	<0.001
Conductores adultos	β_2	-0.0022	0.0399	-0.055	0.956
Conductores mayores	β_3	-0.0786	-0.0445	-1.768	0.077
Vehículos de antigüedad media	β_4	0.0346	0.0197	1.753	0.080
Vehículo de mayor antigüedad	β_5	-0.2491	0.0250	-9.968	<0.001
Vehículos de potencia baja	β_6	-0.2110	0.0343	-6.146	<0.001
Vehículos de potencia media	β_7	-0.0920	0.0329	-2.976	0.005
Vehículos de potencia muy alta	β_8	0.0818	0.0851	0.962	0.336
Bonus-malus	β_9	0.0290	0.0006	49.356	<0.001
Interacción potencia con combustible regular	β_{10}	-0.0847	0.0183	-4.641	<0.001

Tabla 8. Estimaciones del modelo Tweedie para el coste total anual por póliza.
No todas las variables incluidas en el predictor lineal presentan un efecto estadísticamente significativo sobre el coste esperado. Las variables significativas son la edad del conductor, la antigüedad, la potencia del vehículo y el índice bonus-malus (p-valor < 0.005). La columna *estimación* muestra el valor del coeficiente, el *error estándar* su precisión, el *valor z* el contraste de significación y el *p-valor* el nivel de significación estadística.

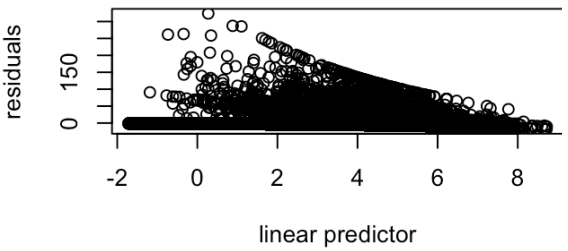
En términos generales, los coeficientes estimados muestran signos y niveles de significación coherentes con los obtenidos en los modelos de frecuencia y severidad. La edad del conductor presenta efectos moderados sobre el coste total esperado, destacando una reducción significativa para el grupo de adultos jóvenes. Comportamiento consistente con la combinación de resultados obtenidos previamente sobre la probabilidad de siniestros y el coste medio por siniestro. Respecto a la edad del vehículo y la potencia, los resultados confirman que vehículos más antiguos y de menor potencia se asocian a un menor coste total esperado, en línea con lo observado especialmente en el modelo de severidad. Por otro lado, el factor bonus-malus destaca nuevamente como la variable con mayor capacidad explicativa, mostrando un efecto positivo y altamente significativo sobre el coste total esperado por la póliza. Finalmente, el tipo de combustible regular presenta un efecto reductor sobre el coste total esperado, consistente con los resultados obtenidos previamente y atribuible al patrón de uso.

El análisis de residuos confirma un comportamiento consecuente con la naturaleza del coste total anual. Se observa una elevada proporción de valores nulos y una marcada asimetría positiva (Figura 9.C). También se aprecia algún importe extremo (Figura 9.A). Los cuantiles superiores de la distribución presentan desviaciones asociadas a pólizas con costes excepcionalmente elevados. Esta variabilidad extrema es inherente a la modelización del coste total (Figura 9.A). La dispersión de los residuos no es constante a lo largo del predictor lineal, lo que indica la presencia de heterocedasticidad, un fenómeno habitual en modelos Tweedie (Figura 9.B). Por último, la comparación entre valores observados y ajustados muestra una elevada dispersión a nivel individual. Este efecto es más acusado en pólizas con bajo coste esperado. En consecuencia, la capacidad predictiva individual es limitada (Figura 9.D). En conjunto, el diagnóstico respalda la adecuación global del modelo.

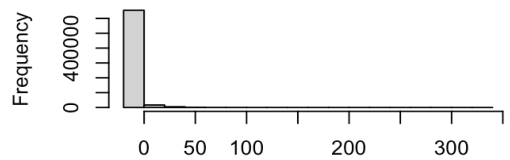
A



B



C



D

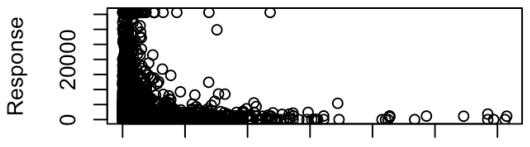


Figura 9. Evaluación gráfica del modelo Tweedie para el coste total anual.
(A) *Q-Q plot de los residuos de devianza.* Los cuantiles centrales se alinean razonablemente con la recta teórica. Se observan desviaciones asociadas a importes elevados y mayor variabilidad en la cola derecha. (B) *Residuos frente al predictor lineal.* Se aprecia una estructura en abanico, con una varianza no constante a lo largo del rango de valores ajustados de los residuos a medida que aumenta el predictor, indicativa de heterocedasticidad. (C) *Histograma de los residuos.* Alta concentración de valores próximos a cero y distribución marcadamente asimétrica a la derecha. (D) *Valores observados frente a valores ajustados*. Elevada dispersión individual, especialmente para pólizas con bajo coste esperado, sin apreciarse patrones sistemáticos en el ajuste global.

5. Estimación de la prima pura

Finalizada la modelización de la siniestralidad, este apartado tiene como objetivo trasladar los resultados obtenidos a la estimación de la prima pura. En el trabajo se aborda la estimación desde dos enfoques complementarios. En primer lugar, se emplea el enfoque clásico en dos etapas, combinando la frecuencia esperada de siniestros, estimada a partir del modelo Binomial Negativo, con la severidad media esperada, obtenida a partir del modelo Gamma. En segundo lugar, se estima la prima pura de manera directa a través del modelo Tweedie.

A partir de ambas estimaciones, se realiza una comparación de los resultados obtenidos. Se analiza sus diferencias en términos de nivel medio, dispersión y coherencia con la estructura de la cartera, utilizando como referencia la variable prima empírica construida a partir de los datos observados. Esta comparación permite valorar las ventajas e inconvenientes de cada metodología y razonar cuál de las dos estimaciones es la más adecuada para su utilización en el caso práctico.

5.1 Estimación de la prima pura: frecuencia-severidad

La estimación de la prima pura mediante el enfoque frecuencia-severidad se ha realizado combinando las predicciones individuales de la frecuencia esperada de siniestros y de la severidad media esperada. Las predicciones se han obtenido de los modelos ajustados en el apartado de modelización. De este modo, para cada póliza de la cartera se dispone de una estimación del coste técnico del riesgo, ajustada a sus características y al tiempo de exposición.

La prima pura resultante muestra el comportamiento habitual de las variables de coste en el ámbito asegurador. La mayoría de las pólizas se concentran en niveles de coste esperado relativamente bajos, con una mediana en torno a 57 unidades monetarias, mientras que la media alcanza aproximadamente 67, reflejando la influencia de valores más elevados. Para facilitar la visualización de esta estructura, la distribución se representa en escala logarítmica, lo que permite apreciar con mayor claridad la distribución central de la cartera y la presencia de valores elevados de prima pura (Figura 10.A).

El análisis por cuantiles muestra que, aunque la prima pura presenta una cola derecha prolongada, la mayoría de las pólizas tienen valores moderados. En concreto, el 75% de las pólizas presentan primas puras inferiores a 93 unidades monetarias y el 95% se sitúan debajo de 165. Estos resultados muestran que los valores más elevados no dominan el comportamiento global del modelo. Esta estructura se refleja también en la distribución acumulada de la prima pura, donde se observa que una fracción limitada de la cartera explica una proporción elevada del coste esperado total (Figura 10.B)

En conjunto, la estimación de la prima pura mediante el enfoque frecuencia-severidad proporciona una medida individualizada, coherente y relativamente estable del coste técnico esperado por póliza. Los resultados obtenidos reflejan adecuadamente la heterogeneidad del riesgo presente en la cartera y permiten identificar la concentración del coste en un número reducido de pólizas, lo que resulta especialmente relevante desde un punto de vista actuarial.

5.2 Estimación de la prima pura: modelo Tweedie

La estimación de la prima pura mediante el modelo Tweedie se ha realizado de forma directa a partir de las predicciones del coste total anual esperado por póliza. De este modo, para cada póliza de la cartera se dispone de una estimación del coste técnico esperado del riesgo, ajustada a sus características y al tiempo de exposición, sin descomponer explícitamente la siniestralidad en frecuencia y severidad.

La prima pura estimada mediante este enfoque presenta también una distribución compatible con la observada en carteras aseguradoras, aunque con una mayor dispersión que la observada en el enfoque anterior. La mayor parte de las pólizas se concentran en niveles intermedios de coste esperado. La mediana se sitúa próxima a 71 unidades monetarias, mientras que la media se sitúa en torno a 82. Estos resultados nos indican que los valores más elevados de prima pura influyen de forma relevante en el valor medio de la cartera, tal y como se observa en la distribución de la prima pura estimada en escala logarítmica (Figura 10.C).

El análisis de por cuantiles confirma este comportamiento, ya que la prima pura estimada mediante el modelo toma valores más elevados en toda la distribución. En concreto, el 75% de las pólizas presenta primas puras inferiores a 115 unidades monetarias y el 95% por debajo de 203. Estos resultados muestran que la distribución se desplaza hacia valores más altos y no solo aumentan los casos extremos. Esta concentración del coste se aprecia claramente al analizar la distribución acumulada de la prima pura estimada mediante el modelo Tweedie, donde se observa que una fracción limitada de la cartera explica una proporción elevada del coste agregado (Figura 10.D).

Estos resultados muestran que el modelo Tweedie también proporciona una estimación coherente del coste esperado por póliza y permite identificar la heterogeneidad y la concentración del riesgo presentes en la cartera.

5.3 Análisis comparativo de las estimaciones de la prima pura

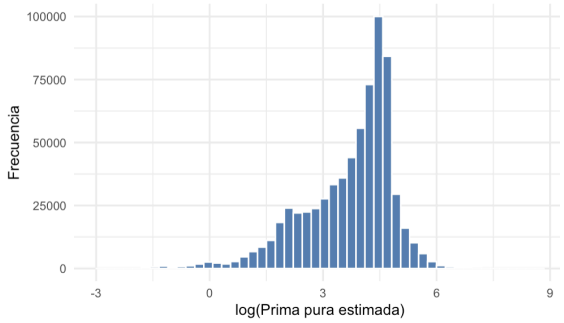
Los resultados obtenidos a partir de ambas estimaciones sobre la prima pura muestran un comportamiento general coherente y que captura adecuadamente la heterogeneidad del riesgo presente en la cartera. Ambos enfoques asignan primas puras más altas a las pólizas con mayor riesgo, lo que muestra que identifican de forma similar los perfiles más y menos riesgosos de la cartera.

No obstante, se observan diferencias relevantes entre ambos enfoques en cuanto a nivel, dispersión y concentración del coste esperado. En términos generales, el modelo Tweedie proporciona estimaciones más elevadas, con valores medios y medianos superiores, lo que refleja un enfoque más conservador en la valoración del riesgo. Además, las primas puras obtenidas con Tweedie presentan una mayor dispersión, asignando más peso a las pólizas con mayor coste esperado. Por el contrario, el enfoque frecuencia-severidad genera estimaciones más moderadas y estables para la mayor parte de la cartera, siendo menos sensible a los valores más elevados.

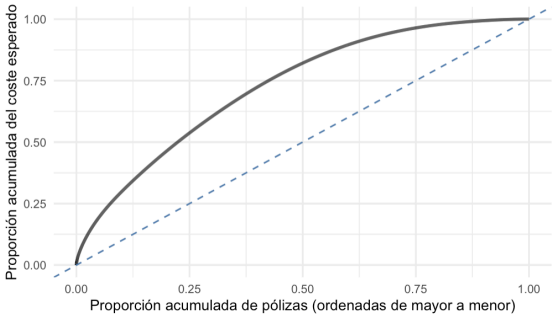
Como referencia adicional, la prima pura empírica analizada en el análisis exploratorio presenta una mayor variabilidad y una mayor concentración del coste, debido a que está directamente afectada por valores extremos. En comparación, las estimaciones obtenidas mediante los modelos presentan un comportamiento más suave y estable.

Desde un punto de vista práctico, el primer enfoque analizado resulta más interpretable y facilita el análisis detallado del riesgo. En cambio, el segundo enfoque proporciona una estimación directa y compacta del coste esperado. La elección entre ambos enfoques depende del objetivo del análisis y del uso que se quiera dar a la prima pura. En nuestro caso, para el caso práctico desarrollado se ha optado por la primera opción, ya que permite tener la información diferenciada sobre la frecuencia de siniestros y la severidad asociada. Información especialmente útil para el análisis y la toma de decisiones en un entorno de gestión.

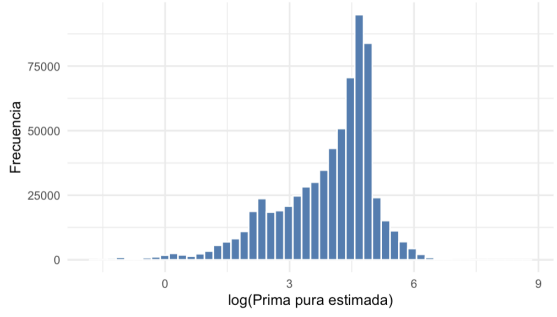
A



B



C



D

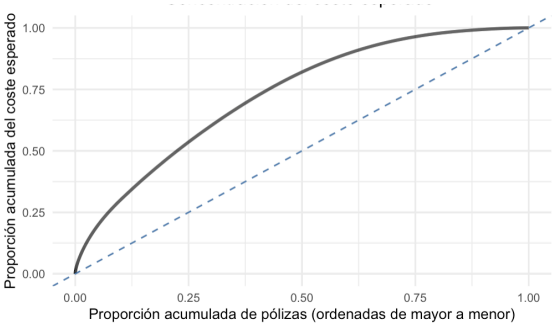


Figura 10. Distribución y concentración de la prima pura estimada según el método de estimación.
(A) Histograma de la prima pura estimada mediante el enfoque frecuencia-severidad en escala logarítmica, con mayor concentración de observaciones en los valores centrales y descenso progresivo en los extremos. (B) Curva de concentración del coste esperado correspondiente al enfoque frecuencia-severidad, donde se observa una acumulación rápida del coste en los primeros tramos y una pendiente decreciente a medida que aumenta la proporción de pólizas. (C) Histograma de la prima pura estimada mediante el modelo Tweedie en escala logarítmica, con una estructura similar y mayor dispersión en los valores elevados. (D) Curva de concentración del coste esperado correspondiente al modelo Tweedie, mostrando una acumulación inicial pronunciada del coste y una pendiente que se suaviza conforme se incorporan pólizas de menor prima pura.

6. Aplicación práctica: *dashboard* de apoyo a la gestión

El objetivo de esta aplicación práctica es trasladar los resultados obtenidos en el estudio de la siniestralidad a una herramienta visual e interactiva que facilite la toma de decisiones en un entorno real de gestión aseguradora. En concreto, se ha desarrollado un *dashboard* en Power BI orientado a perfiles de gestor o vendedor de seguros, con el fin de proporcionar una visión clara y práctica del riesgo, tanto a nivel individual como del conjunto de la cartera. Power BI es una herramienta ampliamente utilizada en entornos profesionales para el análisis y la visualización de datos [12]. El *dashboard* permite explotar la información histórica de pólizas y siniestros, así como las estimaciones de frecuencia, severidad y prima pura obtenidas a partir de los modelos desarrollados en los apartados anteriores. El diseño utilizado responde a la necesidad de presentar la información de forma clara y comprensible para perfiles de gestión. De este modo, se facilita la exploración del riesgo y la identificación de patrones relevantes en la cartera, en línea con los principios de la analítica visual [11]

La herramienta se estructura en dos vistas principales: (i) una vista orientada a la evaluación del perfil de riesgo individual y (ii) una vista centrada en la evaluación del perfil de riesgo de la cartera en su conjunto. Ambas vistas se alimentan por la misma base de datos histórica y son coherentes con el marco metodológico desarrollado en el trabajo.

Por un lado, la vista orientada a la evaluación de riesgo individual está diseñada como herramienta de apoyo directo al gestor o vendedor del seguro. En esta vista, el usuario puede seleccionar de forma interactiva las principales características del cliente. En concreto, se considera la categoría de edad del asegurado, la categoría de edad del vehículo, el tipo de combustible, la potencia del vehículo y la situación del asegurado dentro del índice bonus-malus, distinguiendo entre rangos de bonificación y penalización. A partir de estas selecciones, el *dashboard* muestra las estimaciones de la frecuencia esperada de siniestros, la severidad media esperada y la prima pura esperada asociadas a perfiles con dichas características. Estos resultados están calculados a partir de las predicciones obtenidas sobre la base de datos, sin que el propio entorno de Power BI realice una nueva estimación de los modelos.

La visualización conjunta de estas magnitudes permite al gestor comprender no solo el coste técnico esperado del riesgo, sino también su descomposición en término de probabilidad de ocurrencia e impacto económico. Esta información resulta especialmente útil como apoyo a la toma de decisiones, ya que ayuda a identificar el origen del riesgo. Por ese motivo, en el desarrollo del caso práctico se ha optado por utilizar la estimación de la prima pura basada en el enfoque frecuencia-severidad. Este planteamiento la interpretabilidad y la utilidad práctica frente a enfoques más integrados y conservadores, como el modelo de Tweedie, que proporcionan una estimación global del coste pero dificultan este tipo de análisis desagregado.

Por otro lado, la vista centrada en la evaluación del perfil de riesgo de la cartera proporciona una visión agregada del negocio asegurado, permitiendo al gestor analizar la composición y el comportamiento siniestral del conjunto de pólizas. En esta vista se presentan indicadores globales que resumen el volumen y la evolución de la cartera, como el número total de pólizas y siniestros, el coste total de los siniestros, la frecuencia media, la severidad media y la prima pura media. Asimismo, se incluyen representaciones gráficas que permiten analizar la distribución del riesgo según las principales variables explicativas consideradas en el estudio, así como el comportamiento del índice bonus-malus y su relación con la prima pura media. La interactividad de la vista permite filtrar información por distintos criterios, facilitando un análisis dinámico y adaptado a las necesidades del usuario.

En conjunto, esta aplicación práctica muestra cómo los resultados obtenidos a partir de la modelización de la siniestralidad pueden integrarse en una herramienta de apoyo a la gestión que combina rigor técnico y utilidad operativa. El *dashboard* desarrollado permite trasladar las estimaciones actuariales a un entorno visual e interactivo, facilitando tanto la evaluación individual del riesgo como el análisis agregado de la cartera. No obstante, la herramienta presentada debe entenderse como una primera aproximación, ya que el potencial de este tipo de aplicaciones es amplio y podría explotarse en mayor profundidad mediante la incorporación de nuevas funcionalidades, métricas adicionales o un mayor nivel de desagregación de la información. En este sentido, el *dashboard* constituye una base inicial sobre la que podrían desarrollarse aplicaciones más avanzadas orientadas a la gestión y tarificación del riesgo en un entorno profesional.

7. Conclusiones

En este trabajo se ha analizado la siniestralidad de una cartera de seguros de automóvil mediante modelos lineales generalizados. El objetivo ha sido estimar la prima pura y estudiar el efecto de distintas características de riesgo sobre su valor esperado. La descomposición de la siniestralidad en frecuencia y severidad ha permitido analizar de forma diferenciada los factores que explican la probabilidad de ocurrencia de siniestros y su impacto económico. De este modo, se proporciona una visión clara y coherente desde un punto de vista actuarial.

Los resultados obtenidos muestran que las variables consideradas presentan una mayor capacidad explicativa en la modelización de la frecuencia que en la severidad. En particular, el índice bonus-malus destaca como el factor más relevante para explicar la probabilidad de siniestros. Le siguen variables relacionadas con las características del vehículo, como su antigüedad o potencia. En cambio, la severidad media de los siniestros resulta menos sensible a las características del asegurado y del vehículo. Esto pone de manifiesto la elevada variabilidad del coste de los siniestros y la dificultad de explicarlo únicamente a partir de la información disponible a nivel de póliza.

La estimación alternativa basada en el modelo Tweedie proporciona resultados coherentes con los obtenidos mediante la descomposición frecuencia-severidad. No obstante, tiende a generar primas puras más elevadas y con mayor variabilidad, especialmente en las pólizas con mayor coste esperado. Este comportamiento indica una valoración más conservadora.

No obstante, el análisis pone de manifiesto varias limitaciones. La cartera presenta una elevada heterogeneidad del riesgo y la existencia de valores extremos dificulta la capacidad predictiva a nivel individual, especialmente en la estimación del coste total. En este trabajo se ha optado por una depuración de los datos con el objetivo de no modelizar la cola extrema, sino de obtener una estimación más estable del comportamiento típico del coste por siniestro. Aun así, el análisis de la cola extrema es relevante desde un punto de vista actuarial y debería abordarse en trabajos futuros mediante metodologías específicas orientadas al estudio de estos eventos de alto impacto. Además, el enfoque adoptado se basa en supuestos habituales en la práctica actuarial, como en la independencia entre frecuencia y severidad en la estimación de la prima pura, y en el uso de modelos estáticos que no incorporan información temporal ni comportamiento dinámico de los asegurados. Estas simplificaciones se ajustan con los objetivos del trabajo y a la información disponible, pero implican ciertas limitaciones en los resultados.

Finalmente, el *dashboard* desarrollado en Power BI permite trasladar los resultados obtenidos en la modelización a una herramienta visual e interactiva de apoyo a la toma de decisiones. De esta manera, se facilita tanto el análisis individual del riesgo como la

evaluación agregada de la cartera. No obstante, esta aplicación debe entenderse como una primera versión, basada en modelos previamente estimados, y no como un sistema de tarificación dinámico o predictivo. El *dashboard* desarrollado permite comunicar y utilizar los resultados actuariales de manera clara y práctica, dejando abierto el desarrollo de aplicaciones más avanzadas.

En conjunto, este trabajo ofrece una base sólida y realista para el análisis actuarial de la siniestralidad en cartera de seguros de automóvil, con resultados interpretables y de utilidad práctica. Quedan además claramente identificadas sus principales limitaciones y posibles mejoras.

8. Bibliografía

[1] Causalty Actuarial Society (CAS)| CAS https://www.casact.org/

[2] CASdatasets - Casualty Actuarial Society Data Sets |Université du Québec à Montréal (UQAM) https://cas.uqam.ca/

[3] CASdatasets: User Manual | Casualty Actuarial Society & Université https://cas.uqam.ca/pub/web/CASdatasets-manual.pdf

[4] Wüthrich, M. V. (2018). Non-Life Insurance Pricing with Machine Learning | SSRN https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3164764

[5] Lord, D., Washington, S. P., & Ivan, J. N. (2005). Poisson, Poisson-Gamma and Zero-Inflated Regression Models of Motor Vehicle Crashes | Accident Analysis & Prevention https://search.lib.virginia.edu/sources/uva_library/items/u8089883

[6] Getting Started with Negative Binomial Regression Modeling | University of Virginia Library https://library.virginia.edu/data/articles/getting-started-with-negative-binomial-regression-modeling

[7] Venables, W. N., & Ripley, B. D. (2002). Modern Applied Statistics with S | MASS package https://cran.r-project.org/web/packages/MASS/index.html

[8] Smyth, G. K. et al. statmod: Statistical Modeling | R package https://cran.r-project.org/web/packages/statmod/index.html

[9] Getting Started with Tweedie Models | University of Virginia Library https://library.virginia.edu/data/articles/getting-started-tweedie-models-0

[10] Dunn, P. K. tweedie: Tweedie Exponential Family Models | R package https://cran.r-project.org/web/packages/tweedie/index.html

[11] Keim, D. A., Mansmann, F., Schneidewind, J., & Ziegler, H. (2008). Visual Analytics: Scope and Challenges https://faculty.cc.gatech.edu/~stasko/7450/Papers/keim-va08.pdf

[12] Why Power BI? | Microsoft https://learn.microsoft.com/en-us/power-bi/connect-data/service-publish-from-excel

[13] Generalized Linear Models | Casualty Actuarial Society (CAS) https://www.casact.org/sites/default/files/2021-03/8_GLM.pdf

9. Anexos

Anexo 1. Selección del predictor lineal en el modelo de frecuencia.

Para no sobrecargar el texto principal, el estudio detallado del proceso de selección del predictor lineal se incorpora como Anexo, donde se documenta la evaluación de distintas formas funcionales de las variables explicativas y los criterios utilizados para su elección. Este proceso se llevó a cabo empleando el modelo Poisson como referencia, utilizando el criterio AIC, contrastes de razón de verosimilitud y el análisis del parámetro de dispersión como herramientas principales de comparación, manteniendo constante el término *offset* de exposición.

Como punto de partida, se definió un modelo base (M0) que incluye las variables explicativas seleccionadas para el análisis, utilizando las categorías definidas en el análisis exploratorio en aquellos casos en los que las variables pasaron a ser discretas. En particular, se incluyeron las características del conductor, del vehículo y las variables tarifarias. A partir de este modelo base, se introdujeron de forma iterativa distintas modificaciones en la especificación funcional de las variables. Cada nueva especificación se comparó con el modelo de referencia con las herramientas comentadas (Tabla 9). Cuando una nueva forma funcional aportaba una mejora relevante del ajuste y mantenía coherencia actuarial, se adopta como nueva referencia y el proceso continuaba sobre dicha especificación.

Bloque analizado	Modelo	Especificación evaluada	AIC	Dispersión	Decisión
Modelo base	M0	Edad conductor categorizada, edad vehículo categorizada, potencia categorizada, combustible, bonus-malus	215721	1.66	Referencia
Edad del conductor	M1	Edad conductor continua (lineal)	215985	1.67	Rechazada
	M2	Edad conductor continúa (cuadrática)	215890	1.67	Rechazada
	M3	Edad conductor con spline	215544	1.66	Rechazada
Edad del vehículo	M4	Edad de vehículo continua (lineal)	215840	1.66	Rechazada
	M5	Edad del vehículo continúa (cuadrática)	215676	1.66	Aceptada
	M6	Edad del vehículo con spline	215661	1.66	Rechazada
Potencia	M7	Potencia lineal	215696	1.66	Aceptada
	M8	Potencia cuadrática	215698	1.66	Rechazada
	M9	Potencia logarítmica	215696	1.67	Rechazada
Combustible	M10	Exclusión de la variable combustible	215808	1.66	Rechazada
Interacción	M11	Interacción potencia x combustible	215635	1.66	Aceptada

Tabla 9. Proceso de selección del predictor lineal en el modelo de frecuencia.
Comparación de distintas especificaciones funcionales de las variables explicativas mediante el criterio AIC y el parámetro de dispersión. La columna "Decisión" indica si la especificación es aceptada o rechazada en función de su mejora del ajuste y su coherencia actuarial.

A partir del modelo base (M0), se evaluaron distintas especificaciones funcionales para la variable edad del conductor, considerando tanto su tratamiento continuo como formas más flexibles. Si bien el uso de splines (M3) permitió mejorar ligeramente el ajuste del modelo, estas especificaciones introducían una mayor complejidad y dificultan la interpretación de los resultados. Por lo tanto, como uno de los objetivos del trabajo está orientado a la comprensión y explicación del efecto de las variables sobre la frecuencia siniestral, se optó por mantener la categorización por tramos.

Para la edad del vehículo, se esperaba una relación no lineal asociada a la antigüedad y al desgaste del automóvil. La inclusión de un término cuadrático (M5) permitió capturar adecuadamente este comportamiento y mejorar el ajuste del modelo. La especificación mediante *splines* para la edad del vehículo (M6), aunque ofrecía una ligera mejora adicional, fue descartada por los mismos motivos expuestos en el caso de la edad del conductor, al introducir una mayor complejidad interpretativa sin aportar ventajas relevantes para los objetivos del trabajo.

En el caso de la potencia del vehículo, se comparó su tratamiento en tanto forma categórica como en variable continua (M7), así como mediante transformaciones cuadráticas y logarítmicas. Los resultados mostraron que la especificación continua lineal ofrecía un mejor ajuste sin incrementar la complejidad del modelo. La variable tipo de combustible se evaluó comprobando su exclusión del modelo (M10), lo que produjo un empeoramiento del ajuste, justificando su permanencia. Finalmente, la inclusión de una interacción entre la potencia y el tipo de combustible (M11) aportó una mejora relevante, indicando que el efecto de la potencia sobre la frecuencia siniestral difiere según la motorización del vehículo. En conjunto, el proceso de selección del predictor lineal, permite definir una estructura que equilibra adecuando calidad del ajuste, parsimonia e interpretabilidad actuarial.

Índice de figuras

Índice de tablas