Máster en Ciencias Actuariales y Financieras
UNIVERSIDAD CARLOS III DE MADRID

ANÁLISIS E INCLUSIÓN DE VARIABLES EXÓGENAS EN LA TARIFICACIÓN DE AUTOS MEDIANTE MODELIZACIÓN POR GLM

Alumno: Juan Alfonso Martín Cabello.

Tutores de la tesis: Dr. D. José Miguel Rodríguez - Pardo del Castillo.
Dr. D. Jesús Ramón Simón del Potro.

Madrid, Lunes 29 de Junio de 2015.

Universidad
Carlos III de Madrid

AGRADECIMIENTOS

A mis tutores, por su tiempo y conocimiento para el desarrollo de la tesis

A mi familia, por aportar cada aliento

A mi compañera de viaje, por su confianza y apoyo incondicional

A la Universidad Carlos III de Madrid, por permitirme desarrollar toda mi formación académica y conocer a tan variedad de personas

A Liberty Seguros, por toda la formación facilitada

Gracias.

RESUMEN

La creciente necesidad de conocer cada vez mejor los riesgos que atañen a las pólizas de seguro de una compañía, hace que sea necesario progresar en la idea de buscar y obtener cada vez más información referente al comportamiento humano, ámbito social y económico y posibles riesgos que puedan estar correlacionados con la siniestralidad observada. Por ello, las compañías de seguros a la hora de elaborar sus tarifas de autos estudian la posibilidad de incluir nuevas variables que aporten a las ya tradicionales de este tipo de seguro.
En este trabajo queremos ver que aporte e influencia pueden tener estas variables exógenas y la posibilidad de incluirlas en las tarifas de auto.

Palabras Clave: Seguros de Autos, Tarificación, Variables Exógenas, Modelos GLM, Siniestralidad.

ABSTRACT

The promptly demand to know ever better the risks regarding to the policies of an insurance company makes it necessary to advance the idea of looking for and getting more and more information related to human behavior, social and economic aspects and potential risks that could be correlated with claims.
Therefore, insurance companies are developing and studying the possibility of including new variables that can contribute to the traditional variables of a car insurance.
In this work we want to see the contribution and influence that can have these exogenous variables and the possibility of including them in the pricing of auto insurance policies.

Key Words: Auto Insurance, Pricing, Exogenous Variables, GLM Models, Claims.

ÍNDICE

INTRODUCCIÓN

El presente Trabajo Fin de Máster, pretende llevar a cabo un análisis acerca de los seguros de auto, la tarificación empleada en ellos y su aplicación.
Dentro de este trabajo y su análisis, haremos hincapié fundamentalmente en la elaboración práctica de una tarifa de los seguros de automóvil, a través de un modelo predictivo que nos dé una imagen lo más representativa y predictiva posible.

Hoy en día, la mayoría de las compañías aseguradoras llevan a cabo técnicas de predicción a la hora de tarificar los seguros de auto, fundamentalmente utilizan como factores de riesgo las variables características y endógenas de una póliza. En esta exposición queremos ver cómo pueden influir en la tarificación otras variables exógenas, tales como variables sociales, meteorológicas, demográficas, etc. y qué aplicación podrían tener en la tarificación de autos.
Por lo tanto, la finalidad y el objetivo fundamental del presente trabajo es realizar un estudio e identificar qué factores de riesgo establecen fuertes relaciones con la frecuencia y el coste medio observado, sobre una base de datos real de una aseguradora y desarrollar un modelo predictivo GLM, a través del cual podamos determinar y cuantificar el riesgo en términos de prima pura. Así mismo, analizar qué impacto pueden llegar a tener variables exógenas en una tarifa de autos.

El presente trabajo sigue los siguientes pasos:

En primer lugar, desde un punto de vista teórico trataremos el ámbito del seguro y concretamente de los seguros no vida, tipologías y características principales con un enfoque legal.
Por otro lado, definiremos las generalidades y modalidades de los seguros de auto, el concepto y ámbito en el que se desarrollan, los aspectos principales de la tarificación y los modelos lineales generalizados (GLM).

En segundo lugar, llevaremos a estudio un caso práctico en el que se desarrolla la idea fundamental del Trabajo, el estudio de las variables y su asociación con Frecuencia y Coste Medio, y la aplicación de variables exógenas que se consideren relevantes. En dicho caso práctico, se explicará la hipótesis de partida y cada uno de los análisis realizados.

Por último, evaluaremos los resultados obtenidos acerca de nuestro estudio práctico y estableceremos las conclusiones oportunas sobre la viabilidad o no de incluir factores de riesgo externos en las tarifas de autos.

Dentro del ámbito de Pricing y Tarificación actuarial en el sector seguros de no vida, crece la necesidad de disponer cada vez de más información y conocimiento de patrones sociales y conductas, así como el entorno, que permitan conocer bien qué tipo de clientes entran a formar parte de una cartera de seguros.

CAPITULO 1. GENERALIDADES DEL SEGURO

1.1. Introducción

El contrato de seguro es un servicio de seguridad por el que el asegurador se obliga, mediante el cobro de una prima y para el caso de que se produzca el evento cuyo riesgo es objeto de cobertura a indemnizar, dentro de los límites pactados, el daño producido al asegurado o a satisfacer un capital, una renta u otras prestaciones convenidas.
El riesgo generalmente viene definido como la vulnerabilidad ante un potencial daño o perjuicio incierto, independientemente de la voluntad de las partes y cuyo hecho implica un efecto negativo para el asegurado.
Por lo tanto, el contrato de seguro implica que mediante el pago de una prima, el asegurado queda cubierto ante cualquier riesgo declarado en la póliza del seguro.

Esta prima queda establecida por la compañía de seguros, calculada sobre la base de los cálculos actuariales y estadísticos teniendo en cuenta la frecuencia y coste medio en la ocurrencia de eventos, y el histórico de eventos ocurridos al cliente.
Salvo pacto contrario, si no se ha pagado la prima antes de producirse el siniestro, el asegurador se libera de la obligación establecida en el contrato. También, salvo pacto en contrario, la prima es pagada en dinero; su pago es de carácter obligatorio para el tomador o contratante según las condiciones establecidas en la póliza de seguros.
Por otro lado, debido a que el tomador ha pagado la prima correspondiente, éste aspira a que el asegurador asuma el riesgo y cumpla con pagar la indemnización en caso de acaecimiento de un siniestro.

1.2. Seguros No Vida

Los seguros de automóvil se encuadran dentro del ámbito de los Seguros No Vida, los cuáles distan bastante de los seguros de Vida, tanto por su aplicación teórica como por su aplicación práctica.
Teniendo en cuenta el marco legal, por la Ley 30/1995 de 8 de Noviembre sobre Ordenación y Supervisión de Seguros Privados (LOSSP), los seguros se clasifican en tres grupos:

- **Seguros de Vida Directo**: Ramo que cubre los riesgos inherentes a las personas, comprende todos los riesgos que puedan afectar a la existencia, integridad corporal o salud del asegurado (incapacidades).
- **Reaseguro**: Técnica mediante la cual una aseguradora cede parte de los riesgos que asume con el fin de reducir el monto de sus posibles pérdidas.
- **Seguro directo distinto del seguro de vida (No vida)**: Enmarca todos aquellos ramos, que no quedan recogidos dentro del ramo de vida, y que exponemos a continuación, tal como cita el artículo 6 de la Ley Ordenación y Supervisión de Seguros Privados.

Dicho artículo, nos muestra una clasificación de los riesgos por ramos dentro de los seguros directos distintos al seguro de vida (No Vida), éstos son:

1. Accidentes.
2. Enfermedad (Salud).
3. Vehículos terrestres (no ferroviarios).
Incluye todo daño sufrido por vehículos terrestres, sean o no automóviles, salvo los ferroviarios.
4. Vehículos ferroviarios.
5. Vehículos aéreos.
6. Vehículos marítimos, lacustres y fluviales.
7. Mercancías transportadas (comprendidos los equipajes y demás bienes transportados).
8. Incendio y elementos naturales.
Incluye todo daño sufrido por los bienes (distinto de los comprendidos en los ramos 3, 4, 5, 6 y 7) causado por incendio, explosión, tormenta, elementos naturales distintos de la tempestad, energía nuclear y hundimiento de terreno.
9. Otros daños a los bienes.
Incluye todo daño sufrido por los bienes (distinto de los comprendidos en los ramos 3, 4, 5, 6 y 7) causado por el granizo o la helada, así como por robo u otros sucesos distintos de los incluidos en el ramo 8.
10. Responsabilidad civil en vehículos terrestres automóviles (comprendida la responsabilidad del transportista).
11. Responsabilidad civil en vehículos aéreos (comprendida la responsabilidad del transportista).
12. Responsabilidad civil en vehículos marítimos, lacustres y fluviales (comprendida la responsabilidad civil del transportista).
13. Responsabilidad civil en general. (Comprende toda responsabilidad distinta de las mencionadas en los ramos 10, 11 y 12).
14. Crédito. (Comprende insolvencia general, venta a plazos, crédito a la exportación, crédito hipotecario y crédito agrícola).
15. Caución (directa e indirecta).
16. Pérdidas pecuniarias diversas.
17. Defensa jurídica.
18. Asistencia.
Asistencia a las personas que se encuentren en dificultades durante desplazamientos o ausencias de su domicilio o de su lugar de residencia permanente.
19. Decesos.
Garantizan prestaciones únicamente en caso de fallecimiento.

En el caso que a la entidad aseguradora, le sea concedida la autorización simultánea para varios ramos, la denominación será:

1. A los ramos 1 y 2, se dará con la denominación "Accidentes y enfermedad".
2. A la cobertura de ocupantes de vehículos del ramo 1 y a los ramos 3, 7 y 10, se dará con la denominación "Seguro de automóvil".
3. A la cobertura de ocupantes de vehículos del ramo 1 y a los ramos 4, 6, 7 y 12, se dará con la denominación "Seguro marítimo y de transporte".
4. A la cobertura de ocupantes de vehículos del ramo 1 y a los ramos 5, 7 y 11, se dará con la denominación "Seguro de aviación".
5. A los ramos 8 y 9, se dará con la denominación "Incendio y otros daños a los bienes".
6. A los ramos 10, 11, 12 y 13, se dará con la denominación "Responsabilidad civil".
7. A los ramos 14 y 15, se dará con la denominación "Crédito y caución".
8. A todos los ramos, se dará con la denominación "Seguros generales".

Partiendo y considerando la estructura y clasificación de los seguros no vida anteriormente expuesta, tendremos en cuenta ciertas características por las que son comúnmente reconocidos.

En primer lugar, y con carácter general son seguros con una periodicidad corta, normalmente un año, donde las primas cubren el riesgo por el período establecido.

En segundo lugar, las indemnizaciones van en proporción a la cuantía del daño, donde se puede presentar la casuística de infraseguro o sobreseguro cuando la suma asegurada no coincide con el valor de interés asegurado.

En tercer lugar, existe cierta complejidad a la hora de calcular los precios de los seguros de no vida, ya que existen numerosos factores de riesgo y casuísticas que influyen en el cálculo de la probabilidad esperada de un suceso aleatorio.

Por último, dentro del marco socio-económico, influye en gran medida la evolución temporal de la siniestralidad especialmente en ciertos ramos y modalidades.
En cierto modo, este ámbito socio-económico tiene gran importancia, ya que es necesario señalar que cualquier actividad que se lleve a cabo, está condicionada por el entorno en el cuál se enmarca, y en el caso particular del ámbito asegurador se encuentra bastante influenciado por los aspectos económicos, así como por los comportamientos y patrones socio-demográficos.

1.3. Generalidades del Seguro de Auto

En la actualidad, la evolución de la sociedad ha provocado que la mayoría de la población vea necesario y de máxima utilidad poseer un vehículo para su vida cotidiana. La gran cantidad de vehículos, ya sean, coches, motos, camiones surge de esta creciente necesidad de desplazarse diariamente, ya sea para fines ociosos, personales o profesionales.

A día de hoy, aproximadamente, 31 millones de vehículos circulan por las carreteras de todo el territorio español. Estas cifras representan la gran evolución y la gran masa dentro del parque automovilístico nacional.

Esta cantidad de vehículos implica que diariamente se produzcan numerosos desplazamientos, lo cual conlleva a un existente y constante riesgo de que se produzcan siniestros, debido a golpes o accidentes que pueden ser provocados por distintos motivos, ya sea por imprudencias, imprevistos o factores ajenos a los conductores. Todos estos riesgos, han de ser cubiertos por los seguros de auto.

En los seguros de automóviles existe una garantía de contratación obligatoria, que cubre la responsabilidad del conductor del vehículo por los daños que cause a las personas o en los bienes con motivo de la circulación.
De esta forma, todo vehículo debe contratar obligatoriamente un **seguro obligatorio de responsabilidad civil en la circulación de vehículos a motor**.
El incumplimiento de esta normativa, prohíbe la circulación de todo vehículo a motor y deriva en una sanción administrativa.
Por otro lado, se pueden contratar otras garantías voluntariamente como la rotura de lunas, Robo, Daños en el vehículo, etc., que son garantías del seguro voluntario del automóvil.
Además, garantías que suelen añadirse a la hora de contratar el seguro de auto son la asistencia en viaje y la defensa jurídica del asegurado.

Recopilando todas las garantías que pueden incluirse o no dentro de los seguros de auto, nos encontramos con diferentes modalidades.
Por un lado, el seguro de terceros básico, que se limita a cubrir la responsabilidad civil obligatoria, y de él derivan diferentes modalidades que van añadiendo coberturas adicionales como son: Terceros Lunas, Terceros Lunas e Incendio y Terceros ampliado (Terceros Lunas Incendio y Robo).
Por otro lado, existe la modalidad de Todo Riesgo que cubre cualquier percance con el vehículo propio y a terceros (siempre que esté recogido y expresado de forma explícita en la póliza de seguro), y la modalidad de Todo Riesgo con Franquicia, donde se presentan las mismas coberturas que en la modalidad de Todo Riesgo con la salvedad de que el asegurado asume una parte del coste de los siniestros, ésta es la franquicia.

Basándonos y centrándonos en el seguro obligatorio de responsabilidad civil de automóviles, es necesario mencionar que se establece un régimen de responsabilidad civil distinto para daños corporales y materiales:

En el caso de **daños a las personas**, el conductor responde siempre salvo que pueda probarse que los daños fueron debidos únicamente a la conducta negligente del perjudicado o a fuerza mayor ajena a la conducción o al funcionamiento del vehículo.
En el caso de **daños materiales**, responde el conductor si viene causado por su culpa o la de las personas que de él dependen.
En el caso de un comportamiento negligente por parte del conductor y del perjudicado se procederá a un reparto de responsabilidad y correspondiente indemnización.

Este seguro destinado a cubrir la responsabilidad del propietario o conductor por las lesiones corporales o daños materiales que pueda ocasionar a terceros, presenta los siguientes límites:

a) En los daños a las personas, 70 millones de euros por siniestro, cualquiera que sea el número de víctimas.
b) En los daños a los bienes, 15 millones de euros por siniestro.

Por otro lado, quedan excluidas las coberturas sobre las lesiones o fallecimiento del conductor causante del siniestro, así como los daños al vehículo asegurado, y material transportado.

Para el cálculo de las indemnizaciones a cargo del seguro obligatorio de responsabilidad civil en caso de daños personales, se establece un anexo en la ley, que contempla un sistema de valoración de los daños acaecidos sobre las personas en los accidentes de circulación, este sistema de valoración es conocido como el "baremo de automóviles".

A través de este sistema, se establecen las tablas de cuantías indemnizatorias, en función de los daños sufridos.

Las compañías aseguradoras, se ven obligadas a satisfacer a los perjudicados el importe de indemnización como consecuencia del siniestro, las cuáles en el plazo de tres meses desde la recepción de la reclamación por parte del perjudicado, deberá presentar una oferta motivada de indemnización si considera que el asegurado es responsable y si se pueden cuantificar los daños, o en caso contrario emitirá una respuesta motivada.

En el caso de daños materiales, debido a la existencia de convenios de indemnización directa entre las entidades aseguradoras, no tiene efecto el régimen de obligaciones impuestas por la oferta y respuesta motivada, ya que en un plazo corto se liquidan los siniestros.

Por lo tanto, es importante tener en cuenta el importante papel que juegan los convenios y las grandes ventajas que los mismos reportan para los asegurados y perjudicados por un siniestro, ya que permiten liquidar los siniestros con carácter amistoso por la vía rápida.

La forma de actuar de dichos convenios consiste en definitiva en que cada aseguradora paga la indemnización de sus asegurados con independencia de a quién corresponde la culpa del siniestro. Con posterioridad y una vez determinada la culpa liquidan posiciones, lo cual agiliza enormemente el pago de las indemnizaciones, pero en ningún caso puede ser opuesto a los perjudicados, ya que se trata de un acuerdo privado entre entidades que en modo alguno afecta al contrato de seguro.

El asegurador, una vez efectuado el pago de la indemnización, podrá exigir el reembolso del pago al conductor y al propietario del vehículo causante del daño, cuando el daño fuera debido por la conducta dolosa o negligente de cualquiera de ellos o por motivos contrarios a la ley, véase por ejemplo conducción bajo la influencia de bebidas alcohólicas o drogas, o bajo la carencia del permiso de conducir.

Por último, comentar levemente el papel del Consorcio de Compensación de Seguros, cuyo cometido es satisfacer a los asegurados las indemnizaciones derivadas de siniestros extraordinarios, tal y como se recoge en el Real Decreto Legislativo 7/2004, de 29 de octubre, por el que se aprueba el texto refundido del Estatuto Legal del Consorcio de Compensación de Seguros (artículo 8).
El Consorcio asume los daños producidos a las personas y en los bienes cuando esté contratada en la póliza, cualquiera de las coberturas de daños, incendios, robo, rotura de lunas o seguro de accidentes. Asimismo, también asumiría los daños cuando fuera producido por un vehículo robado, por un vehículo no asegurado, o asegurado por una entidad aseguradora declarada en concurso. (RD Legislativo 8/2004 por el que se refunde la Ley sobre Responsabilidad Civil y Seguro en la Circulación de Vehículos a Motor, modificado por la Ley 21/2007).

CAPITULO 2. TARIFICACIÓN DEL SEGURO DE AUTOS

2.1. Introducción

En un mercado cada vez más competitivo, la fijación de las primas de seguro se convierte en una tarea primordial para las compañías aseguradoras. Cuanto mayor conocimiento se tenga sobre el riesgo a cubrir, más exacto será el cálculo de la prima de seguro.
La prima de un contrato de seguro viene determinada por el cálculo de la pérdida esperada en el que incurrirá cada una de las pólizas de seguro contratadas.
De esta manera, la prima calculada se le denomina **prima pura**.
Posteriormente, esta prima pura que refleja la siniestralidad esperada se verá incrementada por la inclusión y recargo de gastos de gestión interna, externa y recargos de seguridad, así como de los impuestos correspondientes.
A esta prima final que paga el asegurado se la denominada **prima total o de recibo**.

Presentando la hipótesis de que una aseguradora tiene "n" contratos idénticos con la misma duración en póliza (normalmente 1 año), mientras que cada póliza tiene un coste de siniestro X_1, X_2,..., X_n; siendo N el número de siniestros, observamos:

$$S = \begin{cases} 0 & N = 0 \\ X1, X2, .., Xn & N > 0 \end{cases}$$

En cada una de las pólizas podemos observar si ha acaecido siniestro o no, así como el correspondiente coste para cada una de ellas.
Por lo tanto, tenemos como coste total de una cartera, lo siguiente:

$$S = \sum_{i=1}^{n} Xi$$

Asumiendo, la igualdad de distribución de probabilidad de las cuantías y su independencia, y la independencia entre la propia cuantía por siniestro y el número de siniestros, obtenemos la esperanza del coste total por póliza E [S].

Esta media es lo que se considera **Prima Pura**, y se obtiene como el producto de la esperanza del número de siniestros por la esperanza de la cuantía de un siniestro:

$$E[S] = E[N] * E[X]$$

La prima pura es la base del precio del seguro, con ella la compañía obtiene la cantidad suficiente para hacer frente a los costes esperados generados por los siniestros.

Para hacer frente al cálculo de dicha Prima Pura, es necesario partir de una base de datos bien estructurada y sin datos erróneos que nos puedan distorsionar los cálculos realizados.
En el proceso de elaboración de una tarifa debemos considerar los factores de riesgo más significativos, es decir, aquellas variables que explican más el comportamiento de la siniestralidad, ya que en todo proceso de tarificación, es fundamental que las compañías de seguros ajusten la obtención de primas equitativas para cada riesgo, teniendo en cuenta la solvencia y rentabilidad de la compañía.
Es necesario realizar una buena adecuación práctica de la misma, tanto en términos de solvencia y rentabilidad, como de competitividad dentro del mercado, así como de orientación sobre a qué segmentos de clientes se va a enfocar la tarifa, restringiendo o no las normas de suscripción.

Por todo ello, las primas calculadas por las compañías de seguros han de seguir los siguientes principios.

En primer lugar, el **principio de equidad** que hace referencia a que las primas se ajusten a la siniestralidad esperada por cada póliza, es decir que cada póliza recoja el precio en función del riesgo que le atañe.

En segundo lugar, el **principio de solidaridad**, recoge la repartición del riesgo y de la prima total, por lo que existe un grado de solidaridad de forma que las pólizas que recojan menos riesgo en un momento dado deban pagar más por su seguro de auto, y viceversa, con el objetivo de que siempre se cumplan los términos esperados de siniestralidad en una cartera de seguros.

Por último, el **principio de suficiencia**, hace referencia al objetivo de que las primas establecidas sean suficientes para cubrir los riesgos y garanticen la solvencia y sostenibilidad de una compañía de seguros.

2.2. Sistemas de Tarificación

Los principios técnicos sobre los que se basan la elaboración de los precios de seguro constituyen el sistema de tarificación. Dentro de este ámbito podemos distinguir dos sistemas:

- Sistema de Tarificación a priori (class rating).
- Sistema de Tarificación a posteriori (experience rating).

<table>
<tr><td>TARIFICACIÓN A PRIORI</td><td colspan="4">TARIFICACIÓN A POSTERIORI</td></tr>
<tr><td rowspan="2">Class Reting</td><td colspan="3">Principios de eficacia en la Tarificación</td><td>Principios de eficacia y de estabilidad</td></tr>
<tr><td>Bonus - Malus</td><td>Merit - Rating</td><td>Retrospective - Rating</td><td>Distribución de Dividendos</td></tr>
</table>

Tarificación a priori

El sistema de tarificación a priori, nos permite establecer una tarifa sin tener experiencia sobre la siniestralidad que conllevan los nuevos asegurados que se incorporan a la cartera.
Este sistema aglutina información y masa de un conjunto de datos pasados de una cartera de auto, teniendo en cuenta el conjunto total de los factores de riesgo que son utilizados para establecer el riesgo. Es decir, los niveles de riesgo para cada póliza pueden ser explicados sólo en parte por factores que pueden ser observados, y cuya influencia puede ser estimada a través de modelos predictivos, como por ejemplo los modelos lineales generalizados (GLM).

Un paso fundamental y clave es la adecuada identificación del conjunto de riesgos más significativos, cuya presencia explica una parte importante de la siniestralidad.

Dado el objetivo de equidad y suficiencia en las tarifas, se buscarán grupos de riesgo homogéneos que tendrán una siniestralidad esperada muy similar.

Tarificación a posteriori

La posibilidad de que ocurra un elevado número de siniestros y disparen el coste frente al que la compañía tiene que responder, en relación a lo estimado, hace que, además del recargo de seguridad, sea necesario el planteamiento de tarificación a posteriori que se empieza a aplicar a partir del primer año del asegurado en la compañía. Si todos los factores de riesgo que influyen en la siniestralidad esperada pudieran ser detectados, medidos e introducidos en los sistemas de tarificación, las clases de tarifa deberían ser homogéneas.

En la práctica, esto no sucede, existen importantes factores de riesgo que no pueden ser contemplados en la tarificación a priori ya que son difíciles de cuantificar como por ejemplo la forma de conducción, la responsabilidad de nuestros clientes, etc. De esta forma, existen diferentes factores que se reflejan en la experiencia individual de siniestralidad, por lo que en este proceso de tarificación a posteriori se lleva a cabo un ajuste de la prima de forma individualizada para cada asegurado.
Una manera de incorporar esta información evolutiva de los riesgos es realizar un sistema de bonificaciones y penalizaciones en función de la siniestralidad del asegurado.

2.3. Bonus-Malus System

En el mercado de seguros, es muy común el proceso de tarificación utilizando el sistema bonus-malus, atendiendo al índice de siniestralidad en los últimos años por parte del asegurado.
El sistema Bonus-Malus, que es de aplicación en los ramos de autos, penaliza a los asegurados que realizan reclamaciones mediante aumentos de prima.
Por contra, recompensa a los conductores que no realizan ningún tipo de reclamación con descuentos sobre prima.

El objetivo principal de este sistema es que todos los asegurados paguen, a largo plazo, una prima que se corresponda con su propia experiencia de reclamaciones. La metodología a la hora de tarificar juega aquí un papel importante con la incorporación de la llamada distribución a priori y el cálculo de la distribución a posteriori a partir de la primera anualidad y de la experiencia observada.

En la tarificación a priori, las compañías de seguro implantan este sistema conociendo a través del fichero SINCO (Fichero histórico del seguro del automóvil), en el que se obtiene la información relacionada a la siniestralidad que pertenece a cada asegurado en los últimos años.

Ejemplo de aplicación del sistema de Bonus en tarificación a priori:

AÑOS - SINIESTROS	BONUS
5A - 0S	55%
4A - 0S	50%
3A - 0S	45%
2A - 0S	40%
5A - 1S	30%
4A - 1S	20%
3A - 1S	10%
1A - 0S	0%
0A - 0S	-10%
(4A -2S / 5A -2S / 2A -1S)	-20%
(4A -3S / 5A -3S / 3A -2S / 1A -1S)	-30%
(2A -2S / 5A -4S)	-40%
(5A - 5S / 4A - 4S / 3A - 3S)	-50%
Siniestros > Años	-200%

En cambio, en la tarificación a posteriori o experience rating, a parte contar con la experiencia pasada del asegurado, este bonus-malus y proceso de optimización de la prima, se comienza a aplicar en base a la experiencia observada en cada una de las pólizas de seguro dentro de la cartera de la compañía.

En definitiva, este sistema es un método de tarificación en el que los asegurados se agrupan en clases según el número de reclamaciones que hayan realizado hasta el período actual.
Por lo tanto, se calculan las primas de seguro aplicables para cada póliza individual, a través de un proceso de optimización ajustándose por una cantidad que depende de la experiencia pasada de cada asegurado, penalizando en caso de ocurrencia de siniestros, mediante subidas en prima.

Sin embargo, en muchas ocasiones, las penalizaciones que se producen incrementando la prima de los asegurados "malos" son excesivas, lo que puede llevar a problemas de competitividad y, de suficiencia por parte de la compañía aseguradora.

.

2.4. Factores de Riesgo

En un proceso de tarificación a priori, se dispone de los datos de siniestralidad de una cartera de autos compuesta por todos los factores de riesgo que le atañen a cada una de las observaciones.
Uno de los complementos clave en la elaboración de una tarifa son estos factores de riesgo, ya que sobre ellos se evalúa el riesgo en términos cuantitativos.

Los factores de riesgo serán características definitorias medibles que pueden ser observados y que pueden tener relación con las variables objetivo.

Existen numerosos factores de riesgos que son tenidos en cuenta a la hora de hacer una tarifa. Éstos son:

- Factores de riesgo del vehículo (Valor, Potencia, Tipo de Vehículo, Uso del vehículo, etc.)
- Factores propios del asegurado (Edad, nacionalidad, Estado Civil, Localización Geográfica, etc.)
- Factores sobre las características del seguro (Modalidad, Forma de Pago, Bonus Aplicados, Antigüedad, etc.)
- Factores de comportamiento social y económico (Tasa de Paro, Morosidad, etc.)
- Factores meteorológicos (Precipitaciones, Heladas, Temporal extremo, etc.)
- Factores de competitividad (Comportamiento y Tendencias de Compañías de la Competencia).

Una de las fases previa a todo el proceso de tarificación y de elaboración predictiva es la selección de factores de riesgo que son posibles variables tarificadoras por su riesgo potencial.
Es necesario conocer y procesar el máximo volumen de información en torno a los riesgos que atañen al asegurado, ya que la evolución de la siniestralidad es constante y factores de riesgo que a día de hoy no tienen una correlación suficiente con el riesgo, en un futuro pueden ser factores relevantes.

Como conclusión, el análisis y valoración de los factores de riesgo es un aspecto muy relevante ya que llevaran en el proceso de tarificación el peso del mismo.
Por lo que, también una buena selección de variables explicativas del riesgo llevará a una mejor significatividad y a una mejora de la gestión del riesgo.

2.5. Escenario Actual: Tarifas Unisex

El año 2012 supuso un cambio importante en el ámbito de la tarificación de las compañías aseguradoras.

Hasta ese momento, las compañías aseguradoras discriminaban sus tarifas y sus estudios de siniestralidad por sexo, pero tras la entrada en el año 2012 de la Directiva de Género legalmente dejó de ser así ("Directiva del Consejo 2004/113/CE, de 13 de Diciembre de 2004, por la que se aplica el principio de igualdad de trato entre hombres y mujeres al acceso de bienes y servicios").

La citada sentencia obligó a cambiar el sistema de tarificación de las entidades aseguradoras para aquellos seguros cuyo precio dependa del sexo del asegurado, como, por ejemplo, seguros de vida y seguros de autos.

Esta forma de tarificar supuso una modificación aún más importante en cuanto al principio de equidad citado anteriormente, en el que cada uno debe pagar en función del riesgo que transfiere a la entidad aseguradora.
Esto se daña, en parte, con la aplicación de la Directiva aplicada, en tanto que se trata de igualar precios de los asegurados que tienen riesgos heterogéneos.

Por lo tanto, cierto es que la implantación de esta directiva supuso una situación diferente a la que se venía viviendo en el ámbito de la tarificación en autos. Ante esta situación las compañías tenían por un lado la opción de aceptar la directiva aplicando a sus tarifas el efecto de no contemplar la variable sexo como variable de tarificación, o tenían la posibilidad de buscar técnicas actuariales a través de las cuales mediante una combinación de variables explicativas del riesgo que tengan una relación con la variable sexo, se ajusten con un carácter lo más fiel posible al comportamiento experimentado por la variable a suprimir.

CAPITULO 3. MODELOS PREDICTIVOS

Los modelos predictivos tienen el objetivo principal de llegar a conclusiones fehacientes y predictivas sobre un conjunto de datos observado. A través de estos modelos, obtenemos una función de correlación entre un conjunto de datos de entrada y una variable respuesta a estudiar o predecir.

En la elaboración de cualquier modelo, es óptimo disponer de una base de datos con suficiente masa, tal y como anteriormente referíamos debe estar bien depurada y tratada para poder llegar a obtener datos estadísticamente significativos y análisis correctos.

En el sector asegurador, fundamentalmente aplicable a los ramos de no vida, los modelos predictivos son una herramienta de gran importancia que se utiliza en multitud de aplicaciones, como por ejemplo la elaboración de tarifas, el cálculo de probabilidad de fuga de un cliente, cálculos de probabilidad de fraude, etc.

En el caso específico de los seguros de autos, los modelos predictivos se usan para llevar a cabo una modelización sobre las variables respuesta como son el número de siniestros (frecuencia), el coste de estos siniestros (severidad), o directamente el burning cost (prima pura).
La mayoría de las compañías aseguradoras utilizan modelos lineales generalizados (GLM), para la estimación de frecuencia y severidad.

En los últimos años los modelos predictivos están cobrando más relevancia también en el caso de seguros de vida, ya que el modelo obtenido se ajusta mucho más a la realidad y cada vez tiene más importancia utilizar un modelo de tarificación en el que sobre la tasa de mortalidad influyan otras variables externas como la situación socio-económica, factores meteorológicos que afecten a la salud, el ámbito en que se desarrolla la vida, o el estilo de vida que cada persona conlleva. El "problema" que puede residir a la hora de llevar a cabo una tarifa de seguros de vida con un modelo predictivo, es que muchas de estas variables pueden definirse como declarativas, en las que el asegurado puede no declarar la realidad acerca del estilo de vida que lleva o factores que condicionen su salud, por lo que podría resultar más complicado ajustar con una imagen fiel la prima de riesgo.

A modo de conclusión, en un análisis de regresión de un modelo, se pretende estudiar la asociación entre variables. Tanto en el caso de dos variables (regresión simple) como en el caso de más de dos variables (regresión múltiple), el análisis puede utilizarse para explorar y cuantificar la relación entre una variable dependiente (Y) y una o más variables independientes (X_1, X_2,…, X_n), así como para desarrollar una ecuación lineal con fines predictivos.

Los modelos lineales se basan en los siguientes supuestos:

1- Los errores se distribuyen de forma normal.

2- La varianza es constante.

3- La variable dependiente (Y), se relaciona linealmente con las variables independientes.

De esta forma tenemos en un modelo lineal, la siguiente ecuación:

$$Yi = \beta 0 + \beta 1Xi + \cdots + \beta kXi + \varepsilon i$$

$$\varepsilon\, i \sim N\,(0, \sigma 2)\ ;\quad E\,(\varepsilon_i) = 0;\ i = 1,2,\ldots, n.$$

Un supuesto clave dentro de los **modelos lineales estándar**, es que la variable respuesta o dependiente, es una variable aleatoria que sigue una distribución Normal con media μ y varianza constante σ^2.

$$Y_i \sim N\,(\mu, \sigma^2)$$

$$\mu = \beta_0 + \beta_1 X_i + \ldots + \beta_k X_i$$

En muchas ocasiones observamos que no se cumplen estos supuestos por la naturaleza de los datos e información que manejamos.

Este problema se puede llegar a solucionar mediante una transformación de la variable respuesta tomando logaritmos, sin embargo, no siempre se consigue corregir la falta de normalidad, la heterocedasticidad o la falta de asociación entre nuestros datos.

Como alternativa, se presentan los **modelos lineales generalizados (GLM).**

Por un lado, los modelos lineales generalizados (GLM), son una extensión de los modelos lineales estándar que presentan las siguientes características:

1- La relación entre la variable dependiente y las variables independientes no tiene por qué ser lineal. Los datos siguen una distribución encuadrada dentro de la familia exponencial.

2- La variable dependiente no sigue una distribución normal.

3- Los residuos no tienen por qué ser homocedásticos, es decir, no tienen por qué tener una varianza constante.

4- Existe una relación lineal entre las variables explicativas y una transformación de la media de la variable respuesta g (E (Y))= X'B.

Por otro lado, estos modelos GLM tienen tres componentes básicos:

- **Componente aleatoria**: Identifica la variable respuesta y su distribución de probabilidad.

En ocasiones, las observaciones de Y, pueden ser datos binarios que se identifican como éxitos y fracasos, que se modeliza a través de una distribución binomial.

En otras ocasiones, cada observación es un recuento o conteo de casos, con lo que se puede asignar a Y una distribución de Poisson o una distribución binomial negativa.

Todos estos modelos se incluyen dentro de la **familia exponencial** de distribuciones.

En los modelos lineales generalizados, Y_i sigue una distribución de la familia exponencial, la cual se define como:

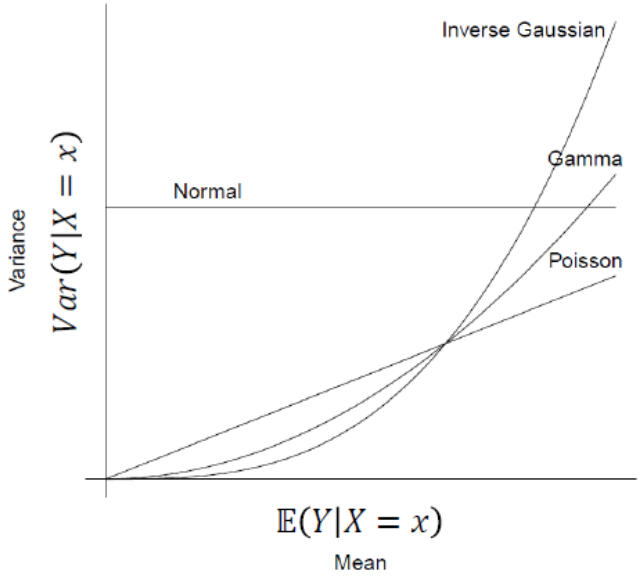
$$fi(yi;\theta,\varphi) = exp\left\{\left[\frac{yi\theta i - b(\theta i)}{\mathrm{ai}(\varphi)}\right] + \mathrm{c}(yi,\varphi)\right\}$$

Esta función depende de tres parámetros: la variable respuesta Y_i, parámetro **θ** y el parámetro **φ**. Además, la función **b (θ_i)** es una función directamente relacionada con la esperanza de la variable (la primera derivada de **b (θi)** es igual a la esperanza de la variable Y) y la función **a_i (φ)** está directamente relacionada con la varianza.

Desde un punto de vista práctico, es útil conocer que las distribuciones que se encuadran dentro de la familia exponencial, quedan completamente especificadas en términos de media y varianza, y la varianza de Y_i es función de su media.

$$Var\ (Yi) = \frac{\varphi\ V\ (\mu i)}{\omega i}$$

La familia exponencial engloba múltiples distribuciones: Normal, Gamma, Binomial, Poisson, Negative Binomial, Gaussian Inverse, etc.



* Cuadro de Relación Media-Varianza.

- **Componente sistemática**: Especifica las variables explicativas utilizadas en la función de predicción lineal, es decir, las variables x_i se relacionan como:

  $$\alpha + \beta_1 X_1 + \ldots + \beta k Xi$$

  Esta combinación lineal de variables explicativas se denomina predictor lineal.

  Alternativamente, se expresa como:

  $$\eta i = \sum_{j}^{N} \beta j Xij$$

Donde x_{ij}, es el valor j-ésimo predictor en el i-ésimo individuo, con i= 1,2,…, N. El término independiente α se obtendría con esta notación haciendo que todos los x_{ij} sean igual a 1 para todos los i.

- **Función link**: Es el elemento fundamental de los modelos GLM.

Los modelos GLM relacionan el valor esperado de la variable dependiente (E [Y]) con el grupo de predictores (β_x), a través de esta función link ($g_{(.)}$).

$$g(\mu) = \eta = \beta 0 + \beta 1X1 + \cdots + \beta kXk = \beta X$$

Cada una de las distribuciones que pertenecen a la familia exponencial da lugar a una función de enlace específica, que ha de ser una función estrictamente monótona y diferenciable.

Distribution	Link Function
Binomial	Logit
Gamma	Inverse (Power (-1))
Geométrica	Log
Inverse Gaussian	Inverse squared (Power (-2))
Negative Binomial	Log
Normal	Identity
Poisson	Log
Tweedie	Log
Zero-Inflated Poisson	Log/Logit
Zero-Inflated Negative Binomial	Log/Logit

Link Function	$\eta_i = g(\mu_i)$	$\mu_i = g^{-1}(\eta_i)$
Identity	μ_i	η_i
Log	$\text{Log}_e\, \mu_i$	$e^{\eta i}$
Inverse	μ_i^{-1}	η_i^{-1}
Inverse Squared	μ_i^{-2}	$\eta_i^{-1/2}$
Logit	$\text{Log}_e\, (\mu_i/1-\mu_i)$	$1/(1+e^{-\eta i})$

$\boldsymbol{\mu_i}$, es el valor esperado de la variable respuesta, y $\boldsymbol{\eta_i}$, es el predictor lineal.

Término OFFSET

El término offset, es una variable que se añade al modelo para equilibrarlo con un (β) parámetro =1.

$$g(\mu) = \beta 0 + \beta 1X1 + \cdots + \beta kXk + \text{offset}$$

Dependiendo de la estructura de los datos, se estiman y si resultan significativos, se introducen en el modelo.

Estimación y Construcción de un modelo GLM

En la construcción de modelos lineales generalizados es importante tener en cuenta que no hay un único modelo válido, ya que en muchos casos, existirán diferentes modelos posibles que se ajusten y tengan la suficiente capacidad predictiva sobre la muestra de datos a analizar. En un proceso de construcción y evaluación de modelos se debe determinar qué modelo es el adecuado y óptimo, que explique una mayor proporción de variabilidad. Es decir, cuál de ellos nos explica más a través de la **Deviance, el AKAIKE (AIC) y el BIC (Bayesian Information Criterium).**

La Deviance proporciona una medida de bondad de ajuste entre los datos observados y los valores ajustados que se obtienen con el modelo (medida de la distancia entre los modelos saturado y ajustado). Denotada como D, se define como dos veces el logaritmo neperiano del cociente de las funciones de verosimilitud del modelo saturado y el modelo calculado, con signo negativo:

$$D = -2 \ln \lambda = -2 [\ln f(y;\beta;\varphi) - \ln f(y;\beta_{sat};\varphi)]$$

El Criterio de información de AKAIKE (AIC), es una medida de la calidad relativa de un modelo estadístico, para un conjunto dado de datos. Como tal, el AIC proporciona un medio para la selección del modelo. Este criterio establece una compensación entre la bondad de ajuste del modelo y la complejidad del mismo.

$$AIC = 2k - 2 \ln (L),$$

Donde k es el número de parámetros del modelo y L, es máximo valor de la función de verosimilitud para el modelo estimado.

El Criterio de información bayesiano (BIC), es un criterio para la selección de modelos entre un conjunto finito de modelos. Se basa, en parte, de la función de probabilidad y que está estrechamente relacionado con el Criterio de Información de Akaike (AIC).

$$BIC = -2 \ln (L) + k \ln (n),$$

Donde k es el número de parámetros del modelo y L, es máximo valor de la función de verosimilitud para el modelo estimado.

En algunos casos queda a nuestro criterio elegir el modelo, ya que pueden existir factores como el propio conocimiento del negocio donde se quiera implantar el modelo que pueda decantar la balanza a favor de un modelo u otro.

Los pasos que hay que seguir en la construcción y evaluación de un GLM son muy similares a los de cualquier otro modelo estadístico.

- **Exploración de los datos**: conocer los datos y buscar posibles relaciones de la variable respuesta con las variables explicativas, considerar la necesidad de aplicar posibles transformaciones de las variables y eliminar las variables explicativas que redunden en la información que aportan.
- **Elección de la función de error y de la función de vínculo**. Una vez analizados los datos, estimar la distribución que sigue la variable dependiente Y, y su correspondiente función de Link.
- **Ajuste del modelo a los datos**: Prestar atención a los test de significación para los estimadores del modelo y la cantidad de varianza explicada (deviance).
- **Análisis de los residuos**. Observar que los residuos se distribuyen de forma normal, con media cero.
  Los residuos son las diferencias entre los valores estimados por el modelo y los valores observados.
- **Simplificación del modelo**. El principio de parsimonia requiere que el modelo sea tan simple como sea posible. Esto significa que no debe contener parámetros o niveles de un factor que sean redundantes e irrelevantes.
- **Validación del modelo**. Una vez elaborado el modelo, se evalúan los resultados de los análisis para garantizar que son independientes entre los datos de entrenamiento y un conjunto de prueba.

Conclusiones sobre los modelos GLM

Los modelos predictivos son de gran utilidad hoy día para las compañías aseguradoras, ya que son capaz de ajustar con gran precisión el riesgo que van a asumir, así como el efecto que tienen distintos factores sobre un suceso aleatorio observado.

En particular, los modelos GLM son modelos avanzados que presentan una forma simple y robusta, y no son modelos muy complicados de desarrollar.
Además, proporcionan la capacidad de realizar y analizar interacciones entre distintos factores de riesgo.

Por el lado negativo cabe destacar que son modelos que requieren gran información para ser precisos, por ejemplo en autos, es habitual realizar un modelo con experiencia previa de hasta tres años. Además, estos modelos tienen una alta sensibilidad frente a los datos observados y estudiados, por lo que las bases de datos tienen que estar bien analizadas y depuradas, con toda la información disponible.

Por último, los modelos lineales generalizados (GLM), ofrecen una plataforma para modelizar, más que una respuesta en sí misma. Una vez elaborado el modelo queda la tarea de interpretar los resultados y el análisis de la tendencia, lo que conlleva una gran cantidad de juicios de valor sobre los resultados aportados.

CAPITULO 4. APLICACIÓN PRÁCTICA

4.1. Hipótesis de Partida

Tras el análisis y el primer estudio teórico realizado sobre los seguros de No vida, y su posterior centralización hacia la tarificación de los seguros de auto, así como el estudio sobre los modelos lineales generalizados, llevaremos a cabo un caso práctico.

En dicho caso práctico, partiremos de una base de datos de auto real, la cual recoge las diferentes variables endógenas y características de una póliza. Posteriormente, extraeremos datos de variables exógenas que consideremos que pueden influir y tener relación con el comportamiento socio-demográfico y que pueden derivar en posibles asociaciones con datos de frecuencia y coste medio.
Estas variables han sido extraídas a nivel de sección censal, que será el hilo de unión frente a cada una de nuestras observaciones de la base de datos de auto.

Uno de los objetivos principales del presente trabajo es ver cómo pueden influir estas variables sobre la frecuencia de siniestros de una cartera de autos y la relevancia que pueden tener a la hora de establecer una tarifa, a través de un modelo GLM.

Toda esta aplicación práctica será desarrollada en el software SAS (Statistical Analysis System), que nos permite efectuar todos los análisis necesarios.

Partiendo de una base de datos completa, haremos los análisis centrándonos en la garantía de daños de automóvil.

La garantía de daños del seguro entra en juego, cuando a parte de la responsabilidad civil obligatoria, el cliente añade las coberturas voluntarias que cubren los daños propios del vehículo.
Por lo tanto, la garantía de daños aplica en las modalidades de Todo Riesgo y Todo Riesgo con Franquicia.

En nuestro caso, vamos a analizar la modalidad de Todo riesgo con Franquicia, ya que dentro de la base de datos es la modalidad sobre la cual disponemos de más información y número de pólizas, tal y como podemos ver a continuación.

Modalidad _Póliza	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
TRCF 120	1100	1.52	1100	1.52
TRCF 125	3348	4.62	4448	6.13
TRCF 180	15355	21.17	19803	27.30
TRCF 200	47758	65.84	67561	93.14
TRCF 300	2985	4.12	70546	97.25
TRCF 450	71	0.10	70617	97.35
TRCF 600	601	0.83	71218	98.18
TRCF 90	227	0.31	71445	98.49
TRCF 99	1093	1.51	72538	100.00

En este primer cuadro se detalla la modalidad de la póliza y el importe de cada una de las franquicias que conforman la base de datos, así como la frecuencia de pólizas en cada una de las franquicias observadas.
Como observamos en el acumulado, la base de datos posee 72538 registros (pólizas), siendo la Franquicia de 200€ la más común, que acumula el 65% de los datos.

SINIESTROS	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0.00	67333	92.82	67333	92.82
1.00	5201	7.17	72534	99.99
2.00	1	0.00	72535	100.00
3.00	1	0.00	72536	100.00
4.00	1	0.00	72537	100.00
5.00	1	0.00	72538	100.00

En este segundo, observamos una relación de los datos a la frecuencia del número de siniestros, sobre la base de datos original, vemos como disponemos de un nivel de frecuencia de 5215 siniestros, que representan un peso del 7.19% sobre el total de pólizas en la base de datos.
Por un lado, la mayor parte de las observaciones (92.82%) no han tenido siniestro en el período de observación, y por otro lado, aquellos que si han tenido siniestros la mayoría se concentran en 1 siniestro declarado (7.17%), el porcentaje restante (0.02%) han declarado 2, 3, 4 ó 5 siniestros.

La base de datos sobre la cual vamos a trabajar recopila información sobre un Rolling Year, o lo que es lo un período de observación de un año.

Dentro de la base de datos cada uno de los registros vienen informados por todas las características endógenas de la póliza, como son las variables de características del vehículo (tipo de vehículo, potencia, valor, etc.), las variables personales del asegurado (edad, antigüedad de carne, localización geográfica, etc.…) y las variables sobre la póliza del seguro (forma de pago, antigüedad en la compañía, modalidad, etc.).

Además de estas variables definitorias, cada observación presenta información sobre la exposición al riesgo transcurrido por la póliza, el número de siniestros en los que ha incurrido y la suma del coste de los siniestros, así como la culpabilidad y la fecha de ocurrencia en el caso de que exista.

Otro de los objetivos de esta parte práctica es ver la asociación de cada uno de los factores de riesgo con las variables dependientes de frecuencia y coste medio, con la finalidad de identificar aquellos factores que son más relevantes y tienen mayor poder explicativo para la posterior aplicación de un modelo GLM.

Los pasos que se han seguido en la aplicación práctica han sido los siguientes:

- Exploración y Depuración de la base de Datos.
- Análisis Exploratorio de variables dependientes e independientes.
- Establecimiento y análisis de asociación de todos los factores de riesgo con las variables dependientes de frecuencia y coste medio.
- Análisis de Correlación entre factores de riesgo.
- Selección de Variables.
- Modelos GLM.
- Validación de los Modelos GLM.
- Análisis de Resultados y Conclusiones.

FACTORES DE RIESGO

En la base de datos, tenemos los siguientes factores de riesgo endógenos:

	VARIABLES CARACTERÍSTICAS DE PÓLIZA	
	Nomenclatura BBDD	DEFINICIÓN
SEGURO	CULPA	Posición asegurado (culpable, dudoso, contrario)
	FOCUR_SINIESTRO	Fecha Ocurrencia del siniestro
	FORMA_PAGO	Forma de Pago (Anual/semestral/Trimestral/Mensual)
	IMPORTE_FRANQUICIA	Franquicias (0/90/99/120/125/180/200/300/450/600)
	CONDUCTOR_ES_TOMADOR	Conductor Tomador Póliza (SI/NO)
	modalidad_pol	Modalidad de la Póliza: TRCF
	PORC_BONUS_APLICADO	Bonus-Malus:clasificación cliente en base a siniestralidad
	TIPO_BONUS_ACREDITADO	Siniestros en los últimos 5 años (ligado al Bonus)
	Antigüedad_pol_efec	Antigüedad de la Poliza
VEHÍCULO	CILINE	Cilindrada(cc)
	GARAJE	Garaje Individual/Garaje Colectivo/Vía Pública
	GRUPOS_MM_DANOS_DIRECTO	Agrupacion cluster de Vehículos en base a su comportamiento en Siniestrlidad de Daños
	MARCAE	Marca
	MODELE	Modelo
	VERSIE	Version
	tipo_vehiculo	Tipo de vehiculo (Monovolumen, Polivalente,Coupe, Cabrio, …)
	MAS_VEHICULOS_UFAMILIAR	Si existen o no más vehiculos en la familia
	MOTORE	Motor
	PESPOE	Peso-Potencia (kg/CV)
	PLAZAS	Plazas del vehículo
	POTENE	Potencia (Cv)
	PUERTE	Puertas del vehículo
	VELMAE	Velocidad (km/h)
	LONMME	Longitud del vehículo (Cm)
	pesoveh	Peso del Vehículo
	sumnufamiliar_autos	Número de vehículos Unidad familiar
	valor_vehiculo	Valor del vehículo (€)
	ant_vehic	Antigüedad del vehículo
	LITERAL_km_anual	Kilometros Anuales
	LITERAL_USO_VEHICULO	Uso dado al vehículo (Particular/Trabajo/Vacaciones)
CONDUCTOR	SEXO_HABITUAL	Sexo del Conductor
	edad_hab	Edad del Conductor
	ant_car_hab	Antigüedad de Carnet
	NACIONALIDAD	Nacionalidad (Española/Extranjera)
	ECIVIL_CONDUCTOR	Estado Civil (Casado/Soltero/Viudo/Divorciado/…)
	LITERAL_PROFESION_CONDUCTOR	Profesión
	LITERAL_MODO_CONTACTO	Modo de Contacto (Tv/Radio/Prensa/Internet…)
	IND_MULTAS	Multas (Si/No)
	ZONA_DEHABITABILIDAD	Zona Rural / Urbano
	CIUDAD_DORMITORIO	Ciudad dormitorio (SI/NO)
	Municipio	Interior / Costa / Islas
	comunidad	Comunidad Autónoma
	provincia	Provincia
	codigo_postal	Código Postal
	CODIGO_SSCC_ANUAL	Sección Censal

Por otro lado, como comentamos anteriormente, a la base de datos de auto con variables características de póliza e internas, le añadiremos variables exógenas que quedarán ligados a cada uno de los registros por la sección censal a la cual pertenezcan.

Los factores de riesgo exógenos que tendremos en cuenta para llevar a cabo los análisis son:

FACTORES DE RIESGO EXÓGENOS	
Nomenclatura BBDD	DEFINICIÓN
SOCIO-DEMOGRÁFICOS	
AUTOS2MANO	Vehículos Segunda mano (%)
AUTOSNEW	Vehículos Nuevos (%)
GASTO_CARBURANTES	Gasto en Carburantes (%)
GASTO_TRANSPORTE	Gasto en Transporte (%)
SEGUROS	Población que tiene como mínimo un seguro (%)
INGRESOSMEDIOS	Nivel medio de ingresos
DENSIDADPOBLACION	Población por Sección Censal
NOTA_MOROSIDAD	Morosidad en la Sección Censal
TASA_PARO	Tasa de paro
COND_ECON	Condición socioeconómica
ACT_20_59_PRC	Tasa actividad 20-59
NO_TERC_PRC	Actividad no terciaria predominante
TER_PESO	Peso del sector terciario
POB_EDAD_MED	Edad media
POB_EDAD_MNA	Mediana de edad
P20MAS_PRC	Población de mas de 20 años (%)
P20MENOS_PRC	Población de menos de 20 años (%)
TAMANO_MED	Tamaño medio (en personas)
SOLT_PRC	Solteros por vivienda famiilar (%)
CASA_PRC	Casados por vivienda famiilar (%)
VIUD_PRC	Viudos por vivienda famiilar (%)
SEPA_PRC	Separados por vivienda famiilar (%)
DIVO_PRC	Divorciados por vivienda famiilar (%)
PAR_HEC_PRC	Parejas de hecho (%)
HIJOS_MED	Media hijos por núcleo familiar
HABITACION_MED	Número medio de habitaciones en vivienda principal
HABITACION_MNA	Número mediano de habitaciones en vivienda principal
METROS_MED	Metros cuadrados promedio de la vivienda
METROS_MNA	Metros cuadrados mediano de la vivienda
PROP_PAG_PRC	Hog. en propiedad por compra, totalmente pagada (%)
PROP_NPAG_PRC	Hog. en propiedad por compra, con pagos pendientes (hipotecas...)(%)
PROP_HER_PRC	Hog. en propiedad por herencia o donacion (%)
ALQUILER_PRC	En alquiler (%)
VIV_SEC_PRC	Viviendas secundarias (%)
SEG_VIV_PRC	Disponibilidad de segunda vivienda (%)
VIV_VAC_PRC	Viviendas vacías (%)
VIV_ANTIG_MED	Antigüedad media en la vivienda
EDIF_ALTURA_MED	Altura media (plantas)
HOG_DELIN_PRC	Hogares con delincuencia o vandalismo en la zona (%)
POB_ESP_PRC	Pob. de España (%)
POB_EXT_PRC	Pob. Extranjera (%)
EST_PRE_OBL_PRC	Estudios pre-obligatorios (%)
EST_POS_OBL_PRC	Estudios post-obligatorios (%)
NUM_VEHIC_MED	Número medio de vehículos
JUV_ACT_IND	Indicador de juventud de la población potencialmente activa
LOC_ACTIV_PRC	Locales activos (%)
LOC_INAC_PRC	Locales inactivos (%)
LOC_AGRA_PRC	Locales agrarios (%)
OFICINAS_PRC	Oficinas (%)

FACTORES DE RIESGO EXÓGENOS	
Nomenclatura BBDD	DEFINICIÓN
METEOROLÓGICOS	
ALTITUD_MED	Altitud sobre el nivel del mar en metros
TEM_MED	Temperatura media anual
PREC_TOTAL	Precipitación (mm - l/m2) total anual promedio
PREC_APR_DIAS	Número de días de precipitación apreciable promedio anual
LLUVIA_DIAS	Número de días de lluvia promedio anual
NIEVE_DIAS	Número de días de nieve promedio anual
GRANIZO_DIAS	Número de días de granizo promedio anual
HELADA_DIAS	Número de días de helada en promedio anual
RACH_VEL	Velocidad máxima de la racha de viento en Km/h promedio anual
VIENTO_VEL_MED	Velocidad media del viento en Km/h promedio anual
INSOLA_PRC	Insolación anual (%)
PRES_MED	Presion atmosférica (hPa) media anual
OTROS FACTORES	
BOMB_DISTA	Distancia parque de bomberos más cercano
HOG_EXP_RSK_FOR	Hogares Expuestos a Riesgo Incendio Forestal
CUENCA_NIVAL	Cuenca nival
ZONA_INUNDABLE	Zona inundable

4.2. Primeros Pasos

Los pasos iniciales dados, fue en primer lugar analizar la base de datos, comenzando con una depuración de la misma.

Esta depuración se basó en eliminar variables que no eran necesarias para los análisis que se van a llevar a cabo y eliminar observaciones con datos incoherentes. Por ejemplo, variables y registros con alta cantidad de valores informados (missings), observaciones con datos erróneos como datos de exposición iguales a 0.001, primas iguales a 0, etc.

Una vez eliminados datos y variables innecesarias que pueden condicionar el estudio, hemos pasado a calcular ciertos ratios por cada uno de los registros, como son la Frecuencia, el Coste Medio, el Coste Medio incorporando la Franquicia que asume el asegurado, el Burning Cost o Prima Pura y el Ratio esperado de pérdida por cada asegurado. Todas estas medidas las trataremos y definiremos a continuación en el inicio del análisis exploratorio univariable.

Por otro lado, otro de los tratamientos realizados sobre la base de datos fue prescindir de aquellos datos con valores de frecuencia muy elevados, por lo que se llevó a cabo un proceso de eliminación de datos atípicos donde se depuró el 2% de los datos de la cola de la variable frecuencia, que representaban datos de siniestralidad en el primer mes de suscripción de la póliza. Esto supuso la eliminación de 1478 observaciones con siniestros en dicha situación, y también aquellas que presentaban más de un siniestro.

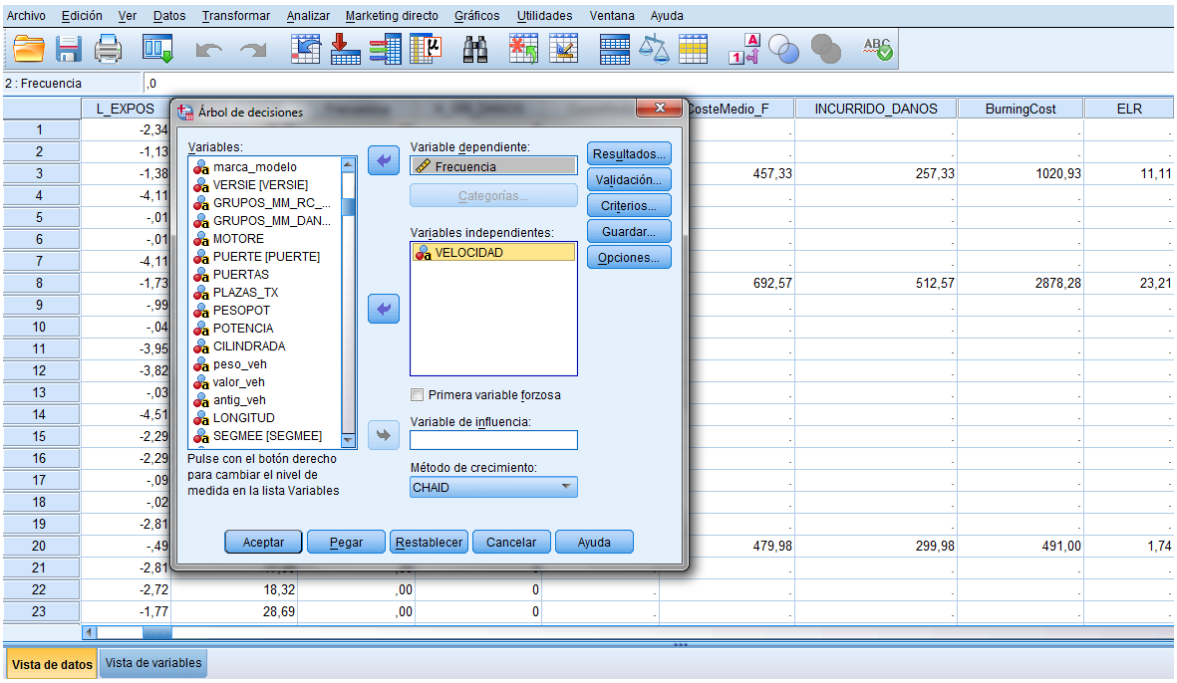
Por lo tanto, la base de datos con la que plantearemos los estudios se posicionó de la siguiente forma:

Modalidad_Póliza	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
TRCF 120	1070	1.51	1070	1.51
TRCF 125	3286	4.62	4356	6.13
TRCF 180	15005	21.12	19361	27.25
TRCF 200	46804	65.87	66165	93.11
TRCF 300	2934	4.13	69099	97.24
TRCF 450	70	0.10	69169	97.34
TRCF 600	595	0.84	69764	98.18
TRCF 90	221	0.31	69985	98.49
TRCF 99	1075	1.51	71060	100.00

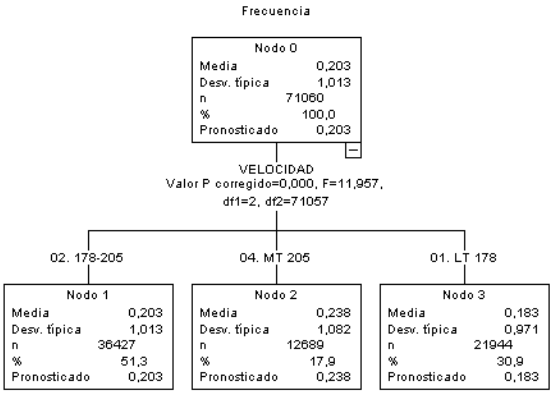
Como se puede observar en el gráfico, la base de datos queda compuesta finalmente de 71060 pólizas, repartidas en cada una de las franquicias observadas en la muestra.

Por último, otro de los pasos dados fue la segmentación de las varibles explicativas.

El proceso seguido a la hora de segementar las variables fue en primer lugar incorporar la base de datos al software SPSS 21, donde a través de dicho software podemos segmentar los factores de riesgo con la técnica de árboles de decisión, a través del método de crecimiento CHAID.



En el árbol de decisión contraponemos la variable dependiente de la Frecuencia, frente a cada una de las variables independientes, obteniendo así los resultados como se expone a continuación en el ejemplo seguido con la variable explicativa VELOCIDAD.



4.3. Análisis Exploratorio

4.3.1. Análisis Univariable

En un primer momento analizaremos y llevaremos a cabo un estudio univariable sobre las variables dependientes o variables a estudiar, en este caso, Frecuencia y Coste Medio, y posteriormente mostraremos como se distribuyen ciertos factores de riesgo relevantes.

La variable dependiente de Frecuencia hace referencia al número de siniestros que se han contabilizado en nuestra cartera de autos, la cual está condicionada por una variable muy relevante, la exposición temporal de cada uno de los asegurados.

$$Frecuencia = \frac{Número\ Siniestros}{Exposición\ al\ Riesgo}$$

La exposición al riesgo es la fracción temporal que el asegurado ya ha vencido en su año natural de contrato en póliza. Por ejemplo, una póliza que lleve 3 meses como asegurada tiene una exposición de 0.25

Por otro lado, el Coste Medio hace referencia al incurrido total de la cartera entre el número de siniestros acaecidos.

$$Coste\ Medio = \frac{Incurrido\ Total}{Número\ de\ Siniestros}$$

Con estas dos variables, podemos llegar a la definición de la Prima Pura, que es el coste esperado de la siniestralidad.

$$Prima\ Pura\ o\ Burning\ Cost = Frecuencia * CosteMedio = \frac{Incurrido\ Total}{Exposición\ al\ Riesgo}$$

Dentro de nuestra base de datos, obtenemos los siguientes datos:

CALCULOS_SQL	INCURRIDO TOTAL	NÚMERO SINIESTROS	FRECUENCIA	COSTE MEDIO	BURNING COST	LOSS RATIO
VALORES	3.275.683,00 €	3727	11,67%	878,90 €	102,58 €	69,76%

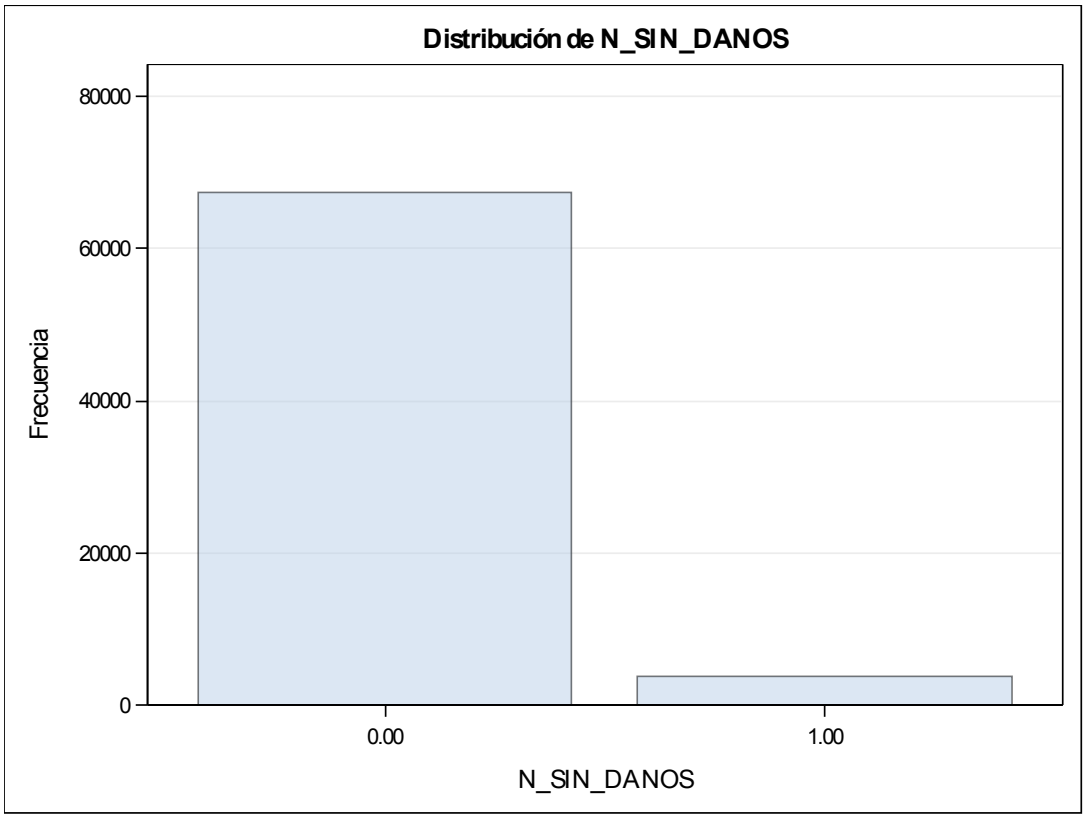
La variable Loss Ratio (ELR), hace referencia a la prima pagada sobre el incurrido total. Si este ratio es superior al 100%, la compañía presenta pérdidas.

VARIABLES DEPENDIENTES

1) FRECUENCIA

- **Distribución del número de siniestros en la garantía de daños.**

SINIESTROS	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
0.00	67333	94.76	67333	94.76
1.00	3727	5.24	71060	100.00



	Número de Siniestros
Pólizas	71060
Missings	0
Suma	3727
Mínimo	0.00
Máximo	1.00
Kurtosis	14.12
Asimetría	4.02
Starndard Desviation	0.22
Media	0.05
Moda	0.00
Mediana	0.00

Tras la eliminación de datos de Frecuencia atípicos como comentamos anteriormente en uno de los primeros pasos realizados en el estudio, la variable de Frecuencia refleja los datos mostrados en los gráficos previos.

En relación a estos datos de frecuencia que obtenemos, vemos como disponemos de un nivel de frecuencia de 3727 siniestros, que representan un peso del 5.24% sobre el total de pólizas de la base de datos. Por un lado, la mayor parte de las observaciones (94.76%) no han tenido siniestro en el período de observación, y por otro lado, aquellos que si han tenido siniestros, el volumen máximo declarado es de un siniestro.

- **Ajuste de la distribución a la variable frecuencia del número de siniestros.**

La variable número de siniestros es una variable discreta que sólo puede tomar algunos valores dentro de un mínimo conjunto numerable, es decir, no acepta cualquier valor, sólo aquellos que pertenecen al conjunto. Por lo tanto, la variable número de siniestros es una variable de conteo que puede tomar valores 0, 1, 2, 3, etc.

A continuación analizamos qué distribución teórica sigue la frecuencia del número de siniestros. Este estudio es necesario para conocer con qué distribución vamos a estimar en los modelos de frecuencia.

Al tratarse de una variable de conteo, centraremos los análisis realizados en las distribuciones de Poisson y Negative Binomial, y los respectivos modelos inflados de cero, ya que como pudimos observar en la tabla de frecuencias del número de siniestros, el 94.76% de los datos no tienen asignados datos de siniestralidad.

Gracias a los procedimientos de SAS podemos ajustar cada una de las distribuciones teóricas anteriores a la distribución empírica representada por nuestros datos. Además, éste procedimiento nos permite comparar el ajuste de cada distribución gracias a que se basa en criterios de información como el AIC o el BIC en lugar de basarse en los tradicionales test de ajuste como Kolmogorov-Smirnov o Chi-Cuadrado. De tal forma que podemos comparar el ajuste y seleccionar la distribución teórica que mejor explica nuestra distribución empírica.

Por lo tanto, para comparar el ajuste entre diferentes distribuciones nos basaremos en el criterio de información Akaike, que es una medida de la calidad relativa de un modelo estadístico, para un conjunto de datos.
Como tal, el AIC proporciona un medio para la selección del modelo.

Ajuste Frecuencia Número de Siniestros			
GARANTÍA	ERROR FUNCTION	LINK FUNCTION	AIC Statistic
DAÑOS	POISSON	LOG	32275
	NEGATIVE BINOMIAL	LOG	32277
	ZI POISSON	LOG	32277
	ZI NEGATIVE BINOMIAL	LOG	32279

Como podemos observar en la tabla superior, existen muy pocas diferencias en el ajuste de las distribuciones, a través del estadístico de bondad del ajuste AIC, por lo que por simplicidad utilizaremos la distribución de Poisson a la hora de llevar a cabo el proceso de modelización, además el exceso de ceros no impide que se ajuste a dicha distribución.

La distribución de Poisson es una distribución de probabilidad discreta que expresa, a partir de una frecuencia de ocurrencia media, la probabilidad de que ocurra un determinado número de eventos durante cierto período de tiempo. Concretamente, se caracteriza por ser una distribución especializada en la probabilidad de ocurrencia de sucesos de conteo.

La función de probabilidad de Poisson se define de la siguiente forma:

$$P\ (\lambda) = \frac{e^{-\lambda}}{K!} * \lambda\text{^}k$$

Salida SAS de ajuste a la Frecuencia del Número de Siniestros sobre la distribución de Poisson.

Sistema SAS

Procedimiento GENMOD

Información del modelo	
Conjunto de datos	TFM.BBDDJUAN
Distribución	Poisson
Función de vínculo	Log
Variable dependiente	N_SIN_DANOS
Variable Offset	L_EXPOS

Número de observaciones leídas	71060
Número de observaciones usadas	71060

Criterios para valorar la bondad de ajuste			
Criterio	DF	Valor	Valor/DF
Desviación	71E3	24819.7937	0.3493
Desviación escalada	71E3	24819.7937	0.3493
Chi-cuadrado de Pearson	71E3	119976.2113	1.6884
Pearson X2 escalado	71E3	119976.2113	1.6884
Verosimilitud log		-16136.8969	
Verosimilitud log completa		-16136.8969	
AIC (mejor más pequeño)		32275.7937	
AICC (mejor más pequeño)		32275.7938	
BIC (mejor más pequeño)		32284.9650	

Algoritmo convergido.

Análisis de estimadores de parámetros de máxima verosimilitud							
Parámetro	DF	Estimador	Error estándar	% de límites de confianza 95de Wald		Chi-cuadrado de Wald	Pr > Chi Sq
Intercept	1	-2.1480	0.0164	-2.1801	-2.1159	17196.1	<.0001
Escala	0	1.0000	0.0000	1.0000	1.0000		

Note: The scale parameter was held fixed.

2) COSTE MEDIO

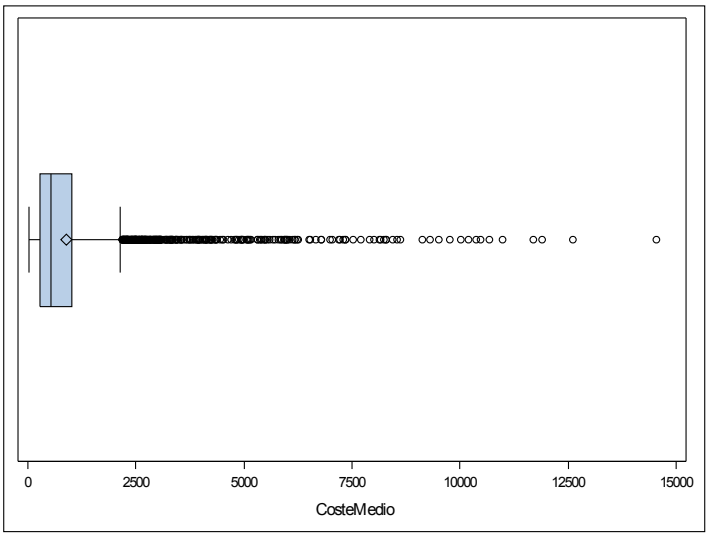
- **Distribución del Coste Medio en la garantía de daños.**

Al igual que representamos con la variable Número de Siniestros, la variable Coste Medio nos aporta los siguientes datos:

	Coste Medio
Pólizas	3727
Missings	67333
Suma	3275683
Mínimo	16.00
Máximo	14535.20
Kurtosis	26.24
Asimetría	4.32
Dev std	1210.17
Media	878.91
Moda	70.57
Mediana	528.49

El total de pólizas que recogen cuantías positivas de incurrido son aquellas que representan el número de siniestros (3727), y suman un incurrido total a asumir por parte de la compañía de 3.275.683 €, lo cual equivale a una media por póliza que ha declarado siniestro de 878.91€.

El diagrama de caja nos muestra claramente la asimetría existente en los datos del coste medio, representando la media con el punto en el interior de la caja y la mediana con la barra vertical. Presenta una cola de siniestralidad prolongada, pero que no representa valores de siniestros extremos, semi-punta (> 15.000€) o punta (> 90.000€). (También fueron eliminados como datos atípicos los siniestros con un coste superior a 15.000€)



- **Ajuste de la distribución a la variable Coste Medio.**

A continuación para la variable del Coste Medio, seguiremos los mismos pasos de bondad de ajuste realizados en el ajuste de la variable Frecuencia del número de siniestros.

En primer lugar, comentar que la variable coste medio, es una variable continua que toma valores a lo largo de un continuo, esto es, en todo un intervalo de valores.
Por lo tanto, la variable Coste Medio se tiene que ajustar a una distribución de probabilidad continua.

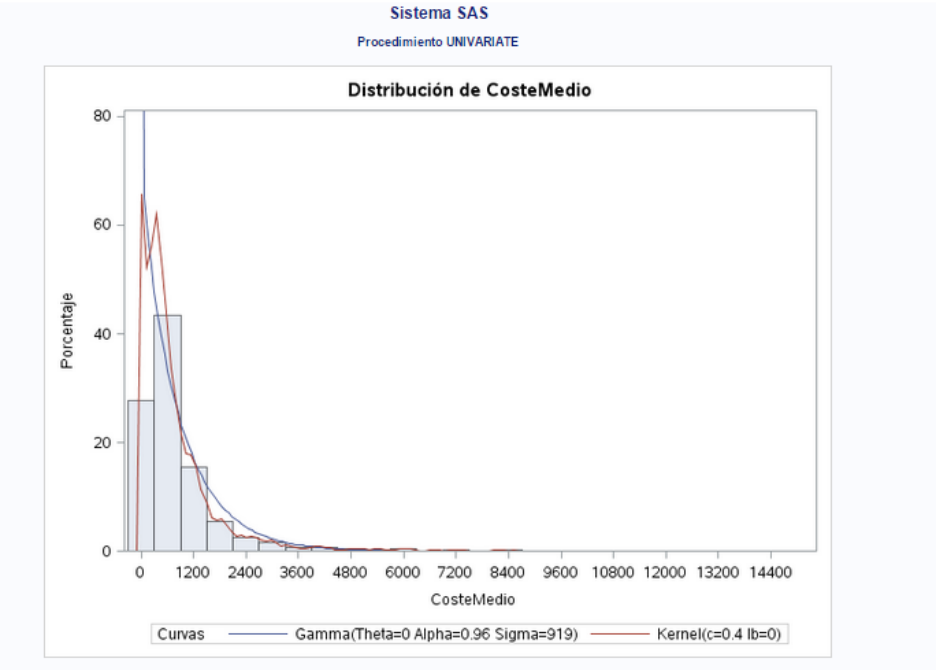
En segundo lugar, observaremos a través de los criterios de información de bondad de ajuste, qué distribución se ajusta mejor al coste medio.

A través de las salidas programadas en SAS, obtenemos los siguientes resultados:

Ajuste Coste Medio			
GARANTÍA	ERROR FUNCTION	LINK FUNCTION	AIC Statistic
DAÑOS	LOG NORMAL	LOG	67844
	GAMMA	INVERSE	57981
	INVERSE GAUSSIAN	INVERSE SQUARED	58551

Como observamos en los valores del ajuste AIC, la variable Coste Medio sigue una distribución Gamma.
La distribución Gamma es una distribución continúa con dos parámetros k y b, cuya función de densidad para valores X > 0 es:

$$f(x) = (1/(\tau(k)*b^{k}))* X^{k-1} * e^{-x/b}$$



Salida SAS de ajuste al Coste Medio sobre la distribución Gamma.

Sistema SAS

Procedimiento GENMOD

Información del modelo	
Conjunto de datos	TFM.BBDDJUAN
Distribución	Gamma
Función de vínculo	Log
Variable dependiente	CosteMedio

Número de observaciones leídas	71060
Número de observaciones usadas	3727
Valores ausentes	67333

Criterios para valorar la bondad de ajuste			
Criterio	DF	Valor	Valor/DF
Desviación	3726	4525.1984	1.2145
Desviación escalada	3726	4325.6801	1.1609
Chi-cuadrado de Pearson	3726	7063.9658	1.8959
Pearson X2 escalado	3726	6752.5119	1.8123
Verosimilitud log		-28988.6398	
Verosimilitud log completa		-28988.6398	
AIC (mejor más pequeño)		57981.2795	
AICC (mejor más pequeño)		57981.2827	
BIC (mejor más pequeño)		57993.7262	

Algoritmo convergido.

Análisis de estimadores de parámetros de máxima verosimilitud							
Parámetro	DF	Estimador	Error estándar	% de límites de confianza 95de Wald		Chi-cuadrado de Wald	Pr > Chi Sq
Intercept	1	6.7787	0.0168	6.7458	6.8115	163707	<.0001
Escala	1	0.9559	0.0194	0.9186	0.9947		

Note: The scale parameter was estimated by maximum likelihood.

3) BURNING COST

El Ratio del **Burning Cost** o **Prima Pura**, representa el ratio del exceso de pérdidas totales sobre el total de los expuestos de una cartera, siendo esto lo mismo que el producto de la Frecuencia y el Coste Medio.
Coste esperado de siniestralidad y porcentaje estimado de frecuencia.

El Burning Cost, por lo tanto es una convolución o mixtura de ambas, del Coste Medio y Frecuencia.
La distribución que se utiliza para modelizar la prima pura, es una distribución **Tweedie**.

La distribución Tweedie tiene un soporte no negativo y un punto de masa discreto en el valor cero. Siendo así útil para modelizar eventos en los que exista una mixtura de observaciones cero y positivas.

La media y la varianza de la distribución Tweedie es:

$$E(Y) = \mu. \qquad VAR(Y) = \varphi\mu^{P};$$

Donde **φ** es el parámetro de dispersión y ***p*** es un parámetro extra que controla la varianza de la distribución.

La familia de distribuciones Tweedie incluye varias distribuciones importantes para los GLM:

–Cuando $p = 0$ la distribución Tweedie degenera en una normal.

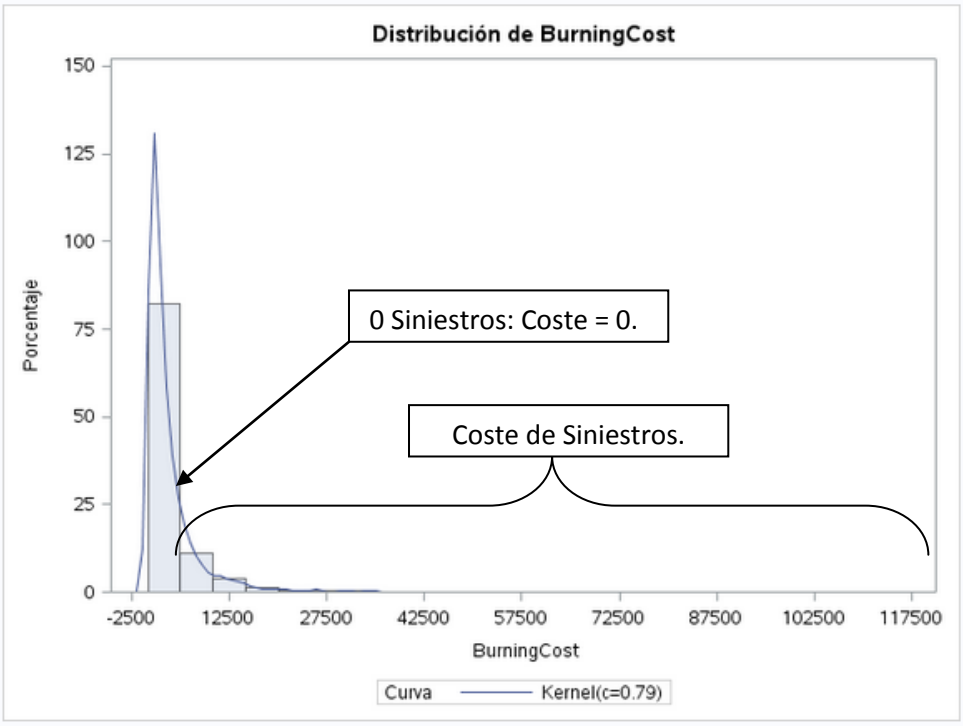
–Cuando $p = 1$ la distribución Tweedie se convierte en una Poisson.

–Cuando $p = 2$ se convierte en una Gamma.

En la práctica el rango de valores más interesante se encuentra entre 1 y 2. Cuando pasamos de 1 a 2 en el parámetro *p* la distribución Tweedie progresivamente va perdiendo su punto de masa discreto en cero para ir derivando hacia la Gamma. En éste caso decimos que la distribución Tweedie se ha generado como una distribución de Poisson compuesta.

Sistema SAS

Procedimiento UNIVARIATE



Sistema SAS

Procedimiento GENMOD

Información del modelo	
Conjunto de datos	TFM.BBDDJUAN
Distribución	Tweedie
Función de vínculo	Log
Variable dependiente	BurningCost
Número de subprocesos	2

Número de observaciones leídas	71060
Número de observaciones usadas	3727
Valores ausentes	67333

Criterios para valorar la bondad de ajuste			
Criterio	DF	Valor	Valor/DF
Chi-cuadrado de Pearson	3726	364.1693	0.0977
Pearson X2 escalado	3726	5736.8388	1.5397
Verosimilitud log		-33734.2719	
Verosimilitud log completa		-33734.2719	
AIC (mejor más pequeño)		67474.5439	
AICC (mejor más pequeño)		67474.5503	
BIC (mejor más pequeño)		67493.2139	

Algoritmo convergido.

Análisis de estimadores de parámetros de máxima verosimilitud							
Parámetro	DF	Estimador	Error estándar	% de límites de confianza 95de Wald		Chi-cuadrado de Wald	Pr > ChiSq
Intercept	1	8.1358	0.0228	8.0912	8.1804	127834	<.0001
Dispersion	1	0.0635	0.0091	0.0456	0.0814		
Power	1	2.4196	0.0203	2.3797	2.4595		

Note: The Tweedie dispersion parameter was estimated by maximum likelihood.

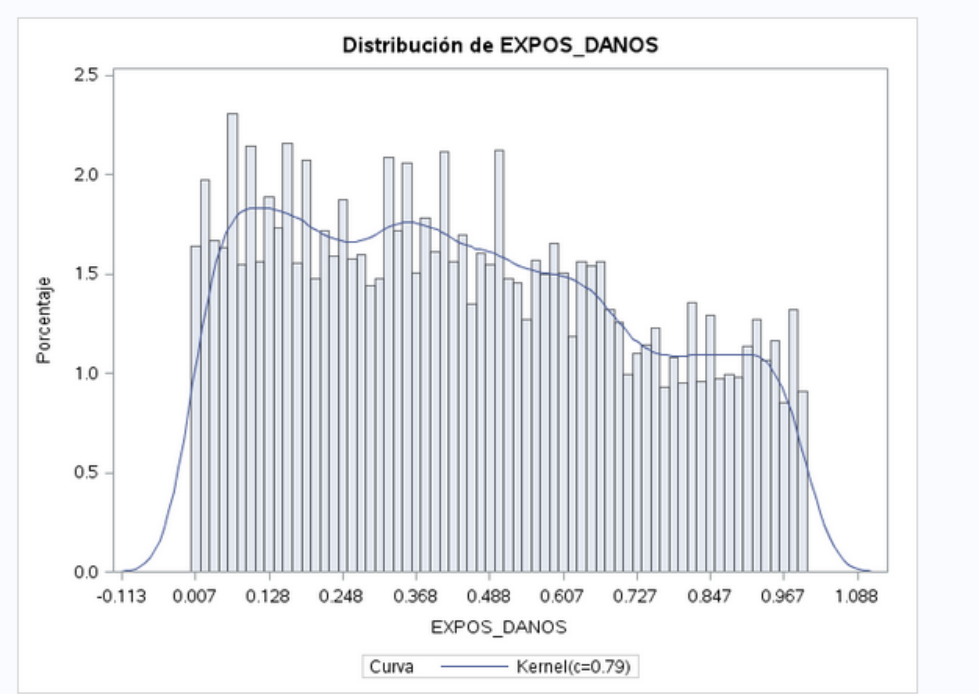
4) EXPOSICIÓN

La exposición juega un papel muy importante, ya que sobretodo influye directamente en la variable Frecuencia. En los datos expuestos llama la atención una media tan baja de la exposición, esto es debido a que en la muestra de la base de datos existe un elevado porcentaje de casos que presentan una baja exposición al riesgo como podemos observar en el histograma.

Por ejemplo, el 30% de la base de datos presenta una exposición inferior a 0.25 años (3 meses).

Por lo tanto, tenemos una base de datos con observaciones que han recorrido poca exposición al riesgo en media.

	Exposición
Pólizas	71060
Missings	0
Suma	31931.95
Mínimo	0.00
Máximo	1.00
Kurtosis	-1.06
Asimetría	0.23
Dev std	0.28
Media	0.45
Moda	0.02
Mediana	0.42

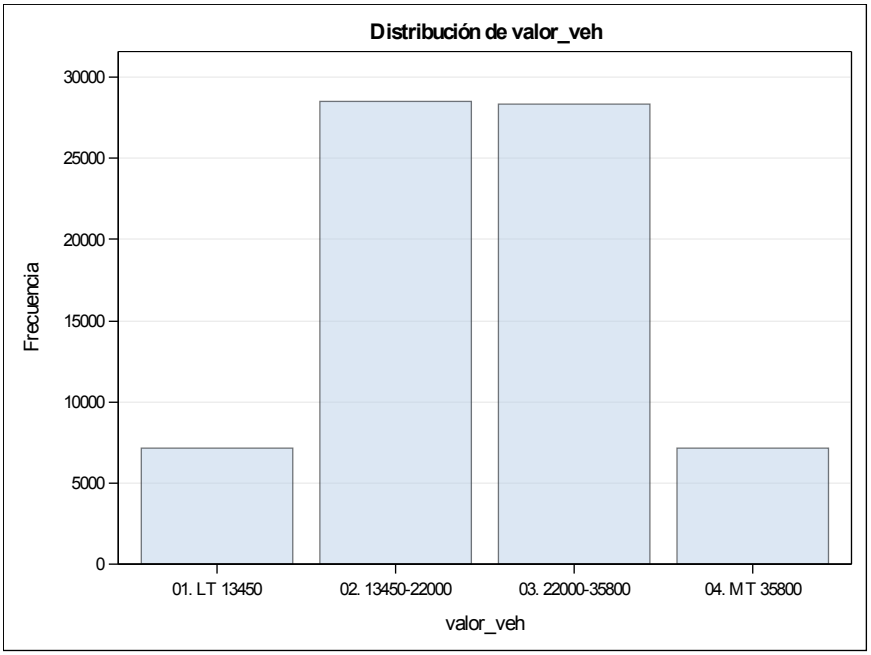


VARIABLES INDEPENDIENTES

A continuación representaremos ciertas variables o factores de riesgo relevantes, en dónde podamos observar cómo se distribuyen dichas variables y los datos que se recogen en la base de datos.

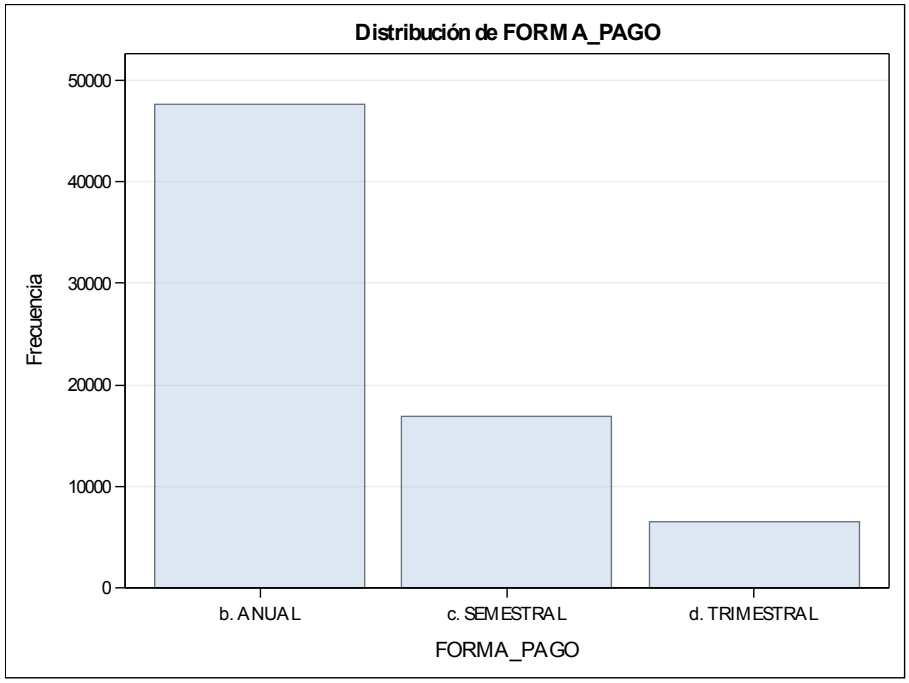
Valor del Vehículo:

Valor del Vehículo	Frecuencia	Porcentaje	Frecuencia Acumulada	Porcentaje Acumulado
01. LT 13450 €	7147	10.06	7147	10.06
02. 13450-22000 €	28486	40.09	35633	50.14
03. 22000-35800 €	28299	39.82	63932	89.97
04. MT 35800 €	7128	10.03	71060	100.00



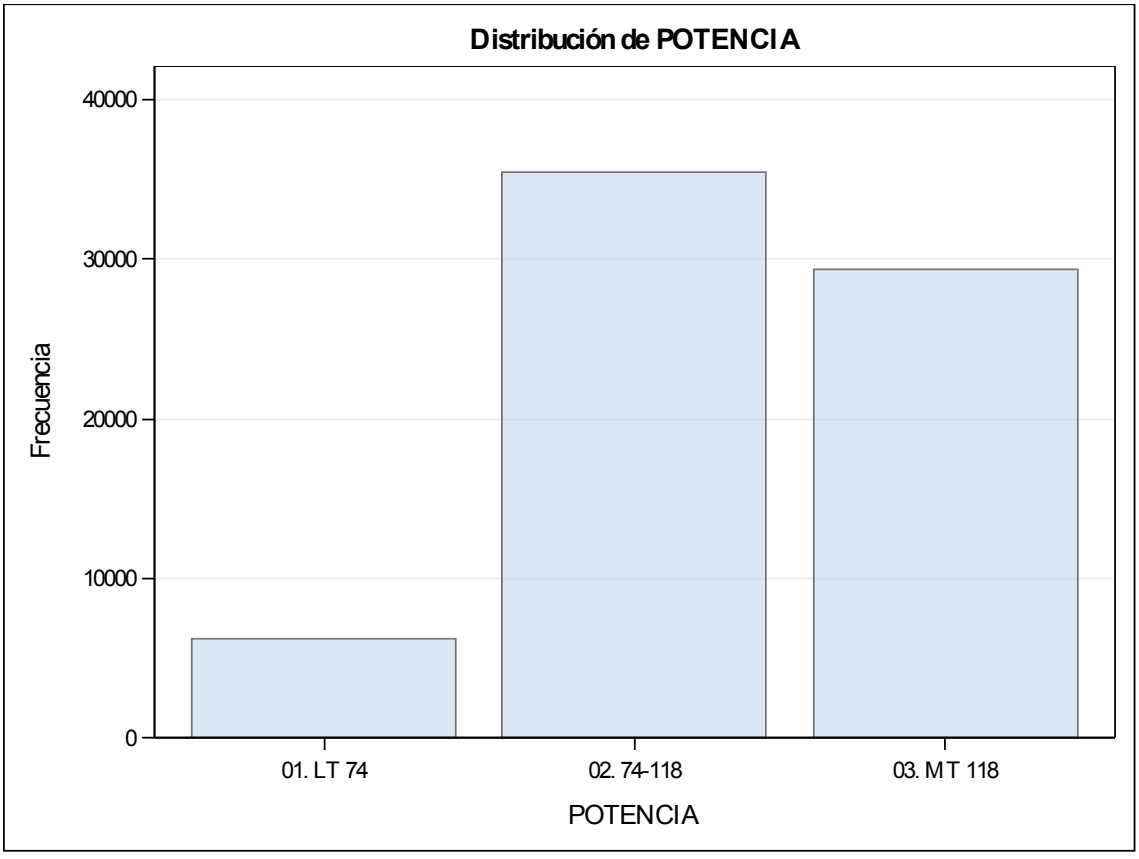
Forma de Pago:

FORMA_PAGO	Frecuencia	Porcentaje	Frecuencia Acumulada	Porcentaje Acumulado
b. ANUAL	47620	67.01	47620	67.01
c. SEMESTRAL	16916	23.81	64536	90.82
d. TRIMESTRAL	6524	9.18	71060	100.00



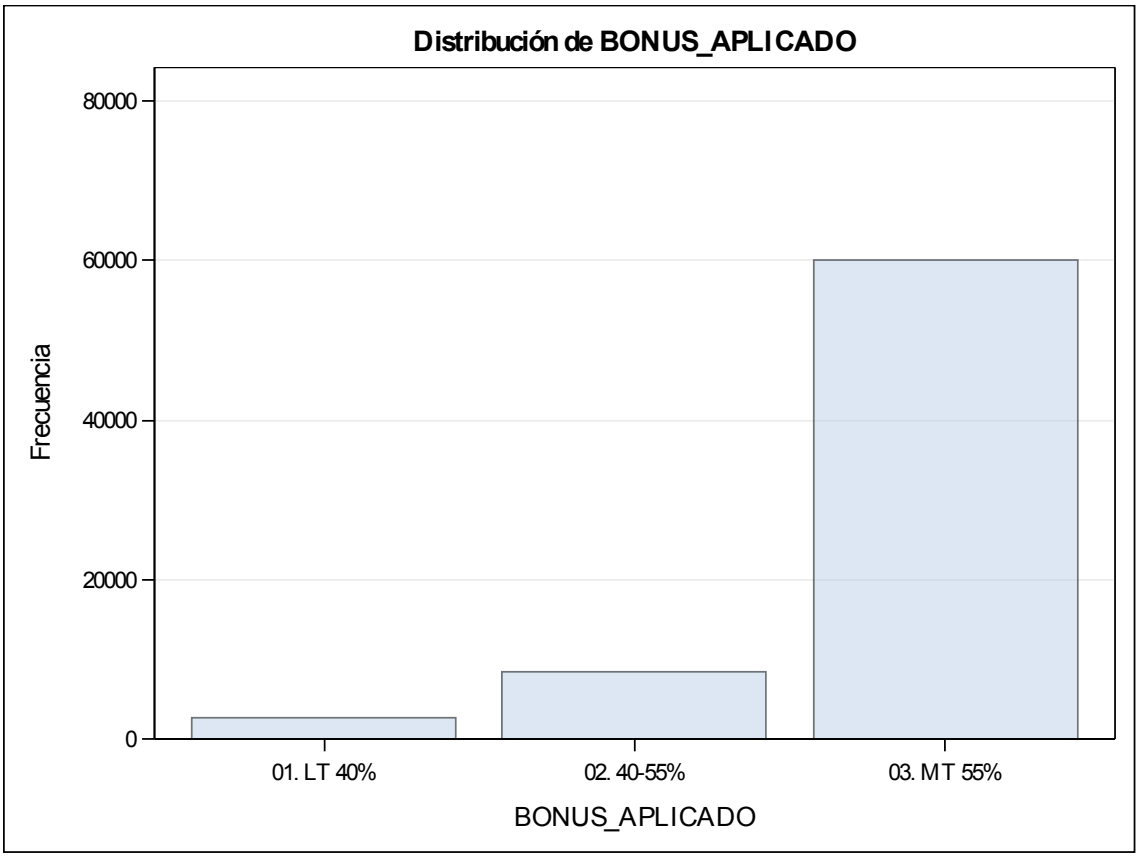
Potencia del Vehículo:

POTENCIA	Frecuencia	Porcentaje	Frecuencia Acumulada	Porcentaje Acumulado
01. LT 74 Cv	6230	8.77	6230	8.77
02. 74-118 Cv	35503	49.96	41733	58.73
03. MT 118 Cv	29327	41.27	71060	100.00



Bonus Aplicado a la Póliza:

BONUS_APLICADO	Frecuencia	Porcentaje	Frecuencia Acumulada	Porcentaje Acumulado
01. LT 40%	2656	3.74	2656	3.74
02. 40-55%	8340	11.74	10996	15.47
03. MT 55%	60064	84.53	71060	100.00



4.3.2. Análisis Bivariable

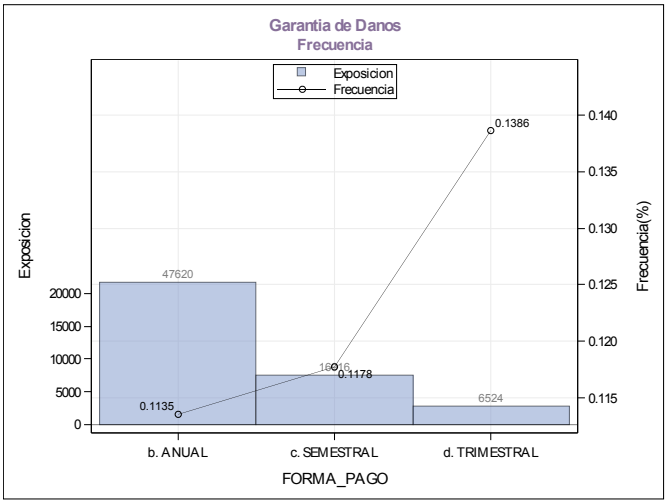
Una vez analizadas las variables de forma individual y observar su comportamiento, continuaremos el estudio a través de un análisis bivariable en el cual se puede apreciar en primer lugar la relación existente de los factores de riesgo con Frecuencia y Severidad, y en segundo lugar, analizar la asociación de estos factores de riesgo entre sí, con la finalidad de visualizar qué variables tienen una alta correlación entre sí y poder así evitar el efecto de la multicolinealidad.
Por lo tanto, la finalidad de estos análisis es ver que variables tienen un gran potencial para ser introducidas como variables explicativas dentro de los posteriores análisis de modelización de Frecuencia y Coste Medio.

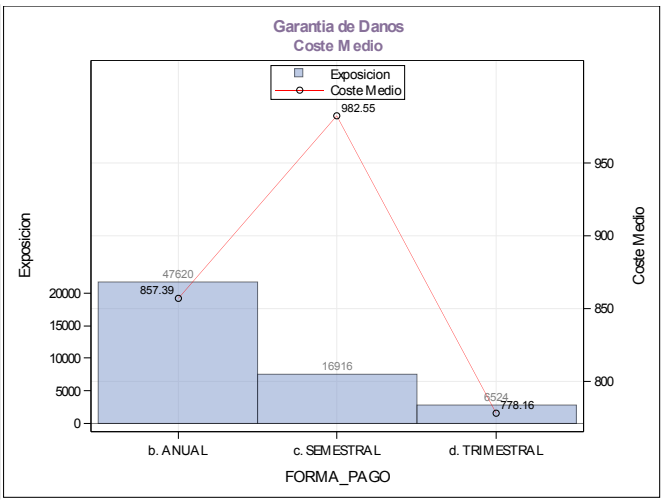
Al igual que en el análisis univariable, representaremos la asociación sobre variables significativas, esto es, sobre aquellas que posean una clara asociación con Frecuencia y Coste.*

Variables Endógenas

La variable Forma de Pago, es una variable que está asociada a la Frecuencia y Coste de una póliza de seguros. Como podemos observar en la situación de pago fraccionado de la póliza, existe mayor frecuencia de siniestralidad.

Además, esta variable presenta el hándicap de estar sujeta a posibles fraudes por parte de los asegurados, ya que si el asegurado elige pagar de forma fraccionada la póliza, existe el riesgo permanente de que pueda declarar un siniestro y no termine de pagar la anualidad completa.

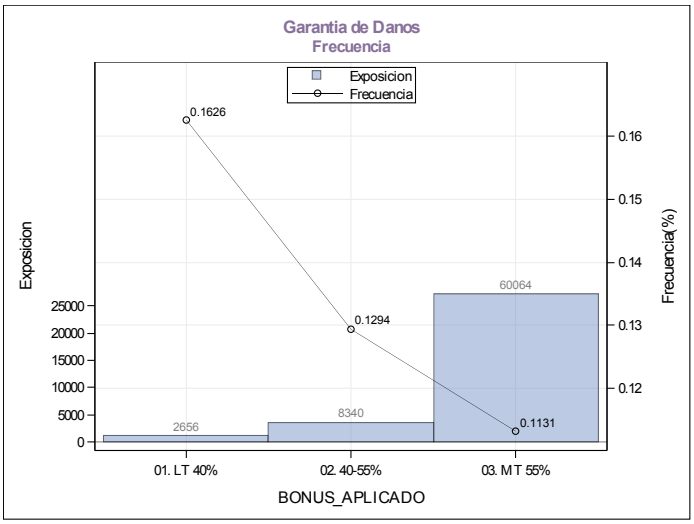


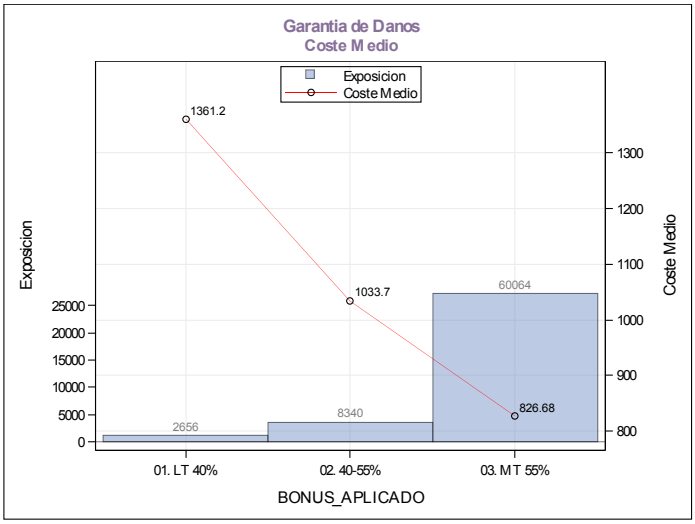


*Estos análisis de asociación han sido realizados con todos y cada uno de los factores de riesgo, con la finalidad de determinar su grado de asociación con Frecuencia y Severidad.

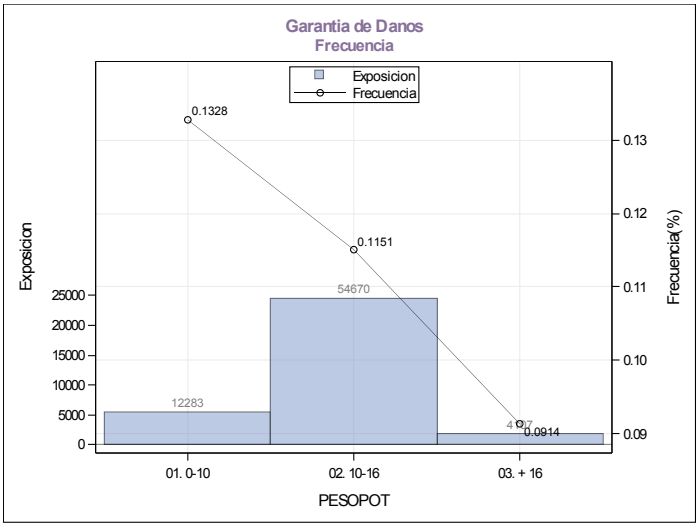
La variable Porcentaje de Bonus que se le aplica al asegurado, se basa en la calidad de clientes que estas incorporando a tu cartera de autos, ya que como podemos observar a mayor porcentaje de descuento menor siniestralidad y menor coste medio.

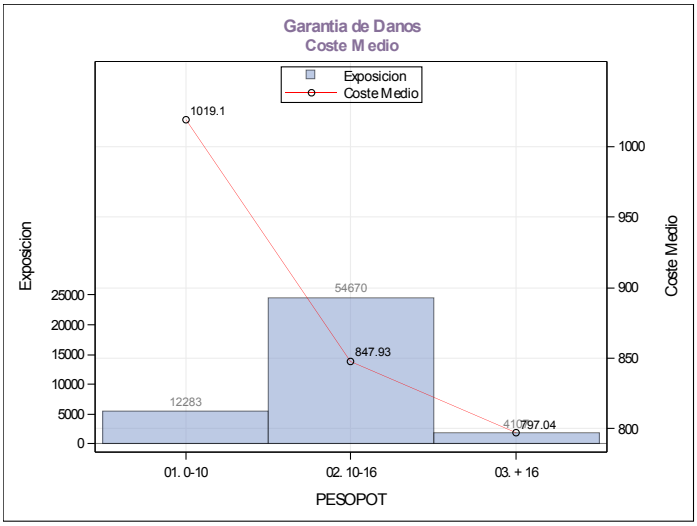
Se otorga mayor porcentaje de Bonus a aquellos clientes que llevan más años sin declarar un siniestro. Por ejemplo en este caso, un 55% de Bonus representa a clientes que llevan 5 años o más sin declarar un siniestro.



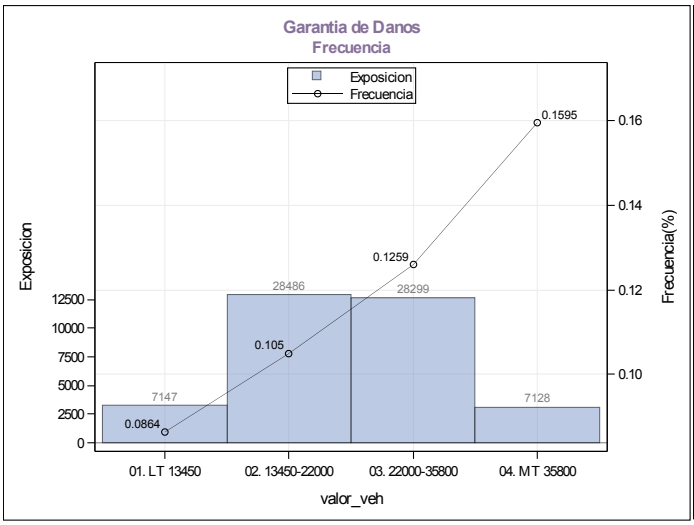


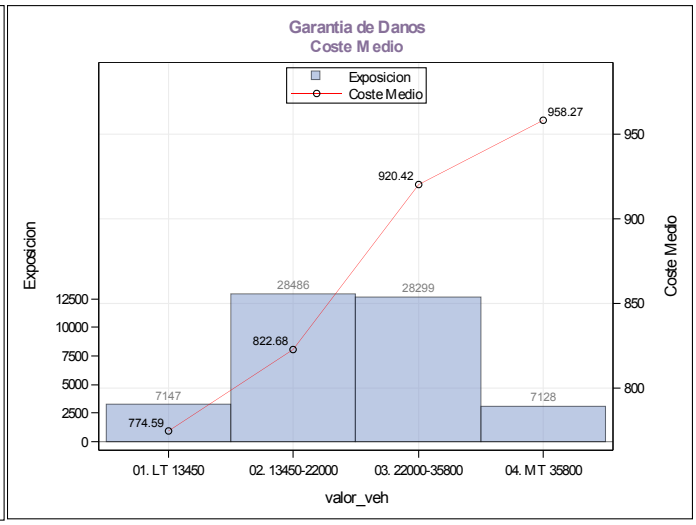
La variable Peso Potencia (Kg/Cv) del Vehículo, es otra de las variables que presentan una alta asociación con la Frecuencia y el Coste Medio. Aquellos vehículos que menor Peso Potencia tienen, presentan mayor siniestralidad y coste.



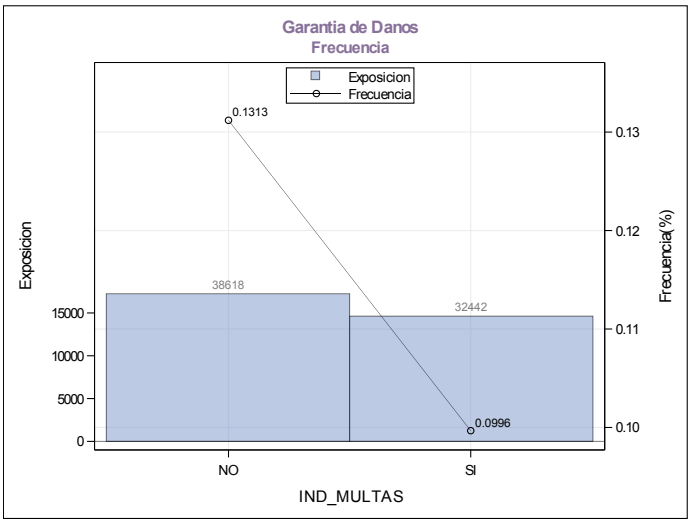


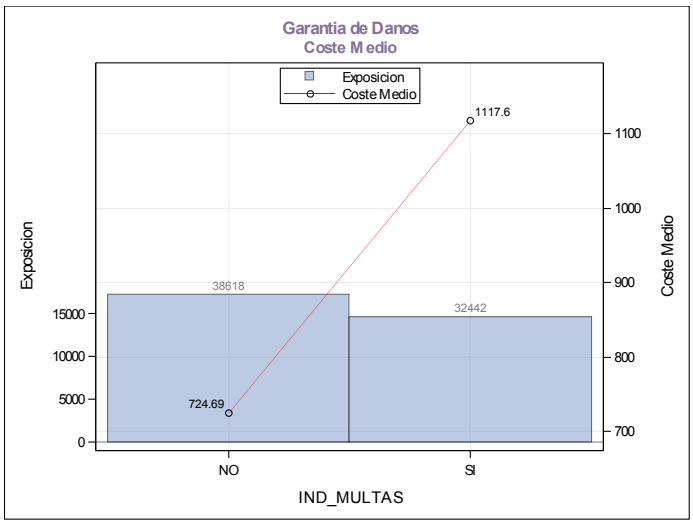
La variable Valor del Vehículo, es otro de los que presenta una clara asociación tanto con frecuencia como con el coste medio, tal y como podemos visualizar en los gráficos.





Contrario a lo que puede parecer el sentido común, la variable Indicador de Multas, representa una alta frecuencia en aquellos que no han tenido ninguna multa y una baja frecuencia en los que sí han tenido multas. En cambio, en relación al coste medio, aquellos que presentan multas, tienen un coste medio mucho mayor que los que no tienen multas.



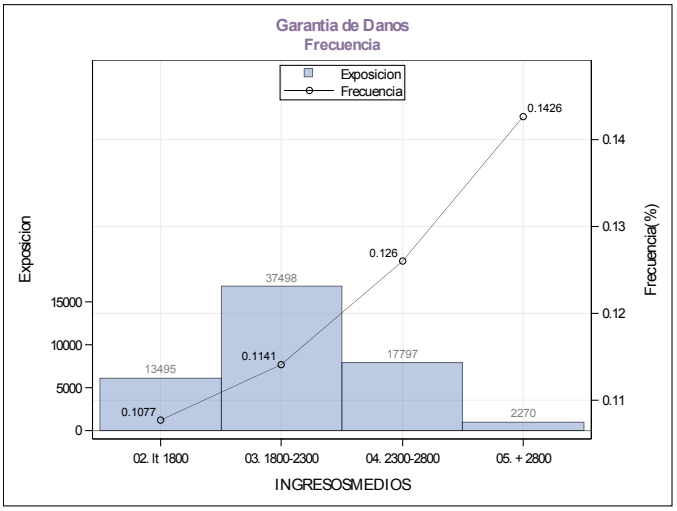


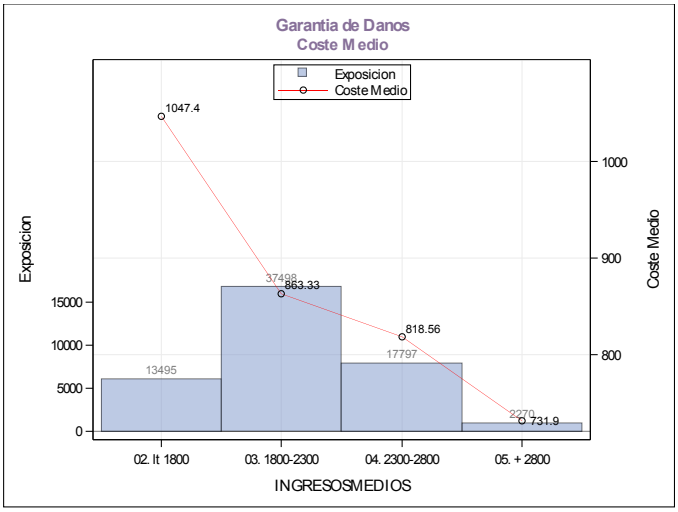
Variables Exógenas

Tras visualizar ciertas variables endógenas, mostraremos los mismos análisis sobre algunos factores de riesgo exógenos relevantes.

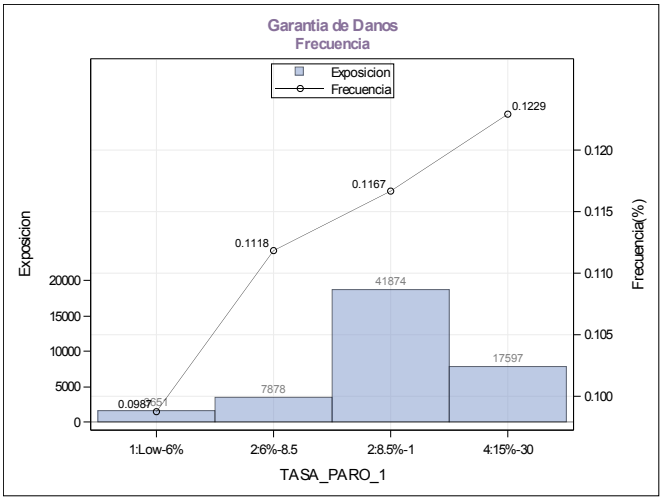
El factor de riesgo de ingresos medios, es un factor bastante discriminante ya que como podemos observar a mayores ingresos mayor frecuencia y menores costes medios.

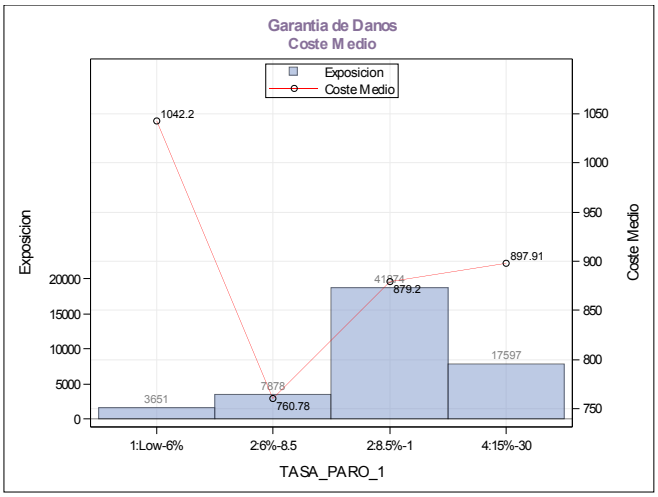
Esto puede estar asociado claramente a que al tener mayores ingresos en media, existe menor reparo a declarar un siniestro y pagar la franquicia correspondiente.





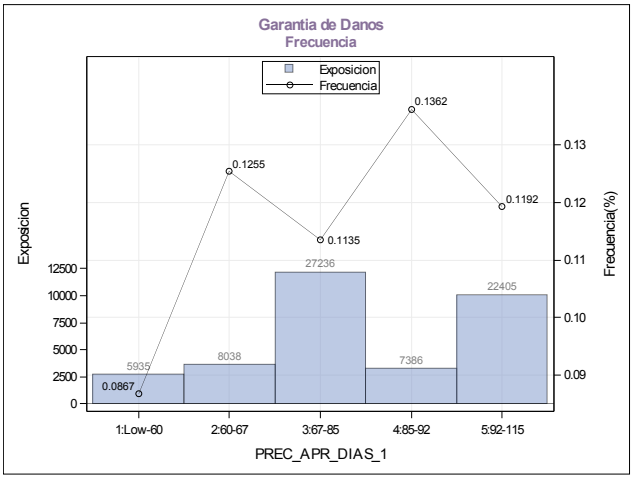
En cuanto a la variable Tasa de Paro, se muestra una progresiva asociación con la frecuencia según aumenta el porcentaje de paro en una sección censal.

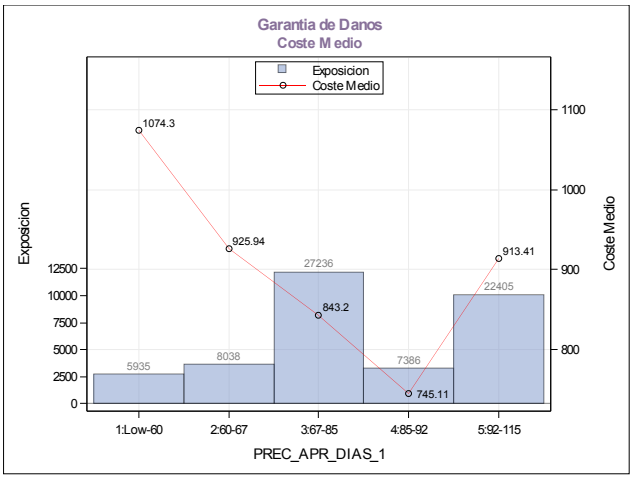




El número de días de precipitación presenta una clara tendencia tanto en frecuencia como en Severidad.

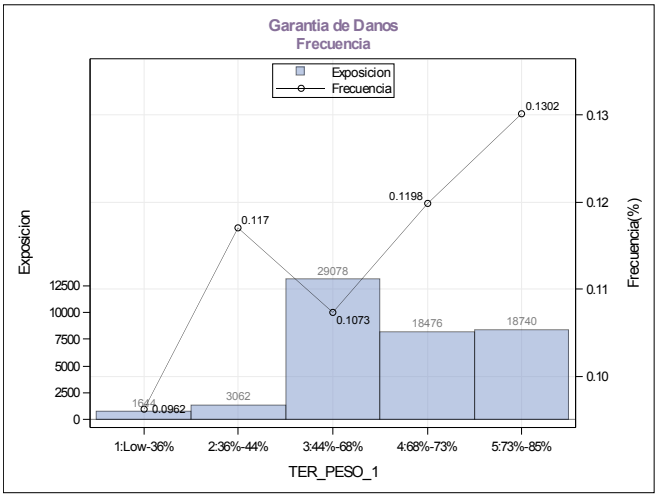
Tal y como hace pensar la lógica, en aquellos lugares donde más precipitaciones hay existe mayor frecuencia.

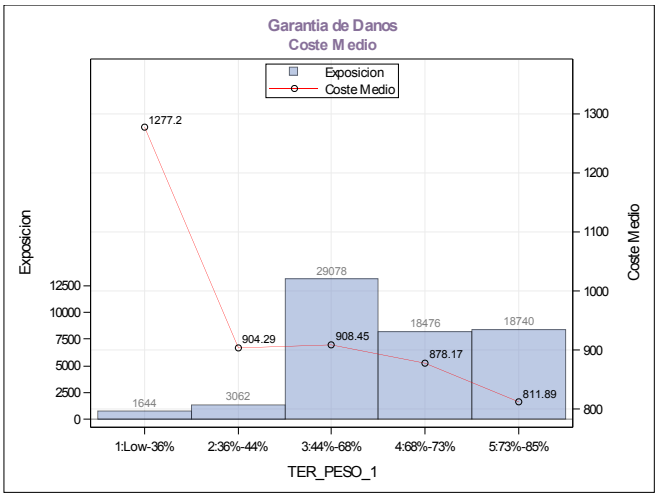




Otra variable que se comporta de la misma forma que la anterior es el peso del sector servicios en cada una de las secciones censales.

En zonas donde existe mayor peso del sector servicios, existen mayores desplazamientos y mayor volumen de vehículos que inciden en un mayor riesgo de siniestralidad.





Matriz de Correlaciones

Una vez analizados todos los factores de riesgo frente a las variables dependientes, es necesario detenerse a realizar un análisis bivariable de los factores de riesgo entre sí.

El análisis estadístico de la asociación (correlación) entre variables, representa una parte básica del análisis de datos en cuanto a que muchas de las preguntas e hipótesis que se plantean en los estudios que se llevan a cabo en la práctica implican analizar la existencia de relación entre variables.

La posible existencia de algún tipo de asociación entre dos o más variables representa la presencia de algún tipo de tendencia o patrón de emparejamiento entre los distintos valores de esas variables.

En nuestro caso práctico, llevaremos a cabo esta medida de asociación a través del coeficiente de contingencia V de Cramer, ya que es una medida simétrica para la intensidad de la relación entre dos o más variables categóricas.

El coeficiente V de Cramer se obtiene a través de la siguiente fórmula:

$$V = \sqrt{\chi 2}/\sqrt{n(\min[r, c] - 1)}$$

Dicho coeficiente V de Cramer oscila entre 0 y 1, de modo que cuanto más próximos a 1 sean los valores mayor intensidad en la asociación de las variables.

Si el valor de la V de Cramer es 0 ó próximo a 0, no hay relación entre las variables. En cambio si la V de Cramer es igual a 1 o muy próxima a 1 existe una relación perfecta entre las variables, por lo que podemos afirmar que ambas variables explican lo mismo. Por otro lado, si la V de Cramer está en torno a 0.6, hay una correlación relativamente intensa entre las variables.

Dado que la V de Cramer es siempre un coeficiente positivo, no se pueden hacer afirmaciones acerca de la dirección de la relación.

A continuación, sobre el estudio de variables realizado, representaremos en una tabla todos los factores de riesgo correlacionados entre sí, con un valor de la V de Cramer por encima de 0.5.

Variable 1	Variable 2	Cramer_V
IND_MULTAS	LITERAL_USO_VEHICULO	0,99991663
POB_EXT_PRC	POB_ESP_PRC	0,98748909
IND_MULTAS	LITERAL_MODO_CONTACTO	0,8751552
PREC_APR_DIAS	provincia	0,82036396
ant_carne	edad_hab_char	0,81840226
PREC_TOTAL	provincia	0,79046837
INSOLA_PRC	provincia	0,77842841
GRANIZO_DIAS	provincia	0,76287334
provincia	Municipio	0,75589257
LLUVIA_DIAS	provincia	0,74646417
VIENTO_VEL_MED	provincia	0,7426321
Municipio	ZONA_DEHABITABILIDAD	0,71597737
tipo_vehiculo	PUERTE	0,70344005
RACH_VEL	provincia	0,6857376
ALTITUD_MED	provincia	0,68170216
NIEVE_DIAS	provincia	0,66058479
HELADA_DIAS	provincia	0,65170151
valor_veh	peso_veh	0,63197367
valor_veh	POTENCIA	0,61082866
tipo_vehiculo	PLAZAS_TX	0,60335013
LLUVIA_DIAS	PREC_APR_DIAS	0,59060327
INSOLA_PRC	GRANIZO_DIAS	0,5884578
INSOLA_PRC	PREC_APR_DIAS	0,58488779
INSOLA_PRC	PREC_TOTAL	0,58450379
INSOLA_PRC	LLUVIA_DIAS	0,57848521
PREC_APR_DIAS	PREC_TOTAL	0,56786732
IND_MULTAS	LITERAL_PROFESION_CONDUCTOR	0,56770875
GRANIZO_DIAS	PREC_APR_DIAS	0,56158076
GRANIZO_DIAS	PREC_TOTAL	0,55649508
valor_veh	CILINDRADA	0,54257028
INSOLA_PRC	VIENTO_VEL_MED	0,52668383
peso_veh	POTENCIA	0,52261471
VIENTO_VEL_MED	GRANIZO_DIAS	0,51887681
VELOCIDAD	PESOPOT	0,50986006
VIENTO_VEL_MED	PREC_TOTAL	0,50952995
GRANIZO_DIAS	ALTITUD_MED	0,50867353
GRANIZO_DIAS	LLUVIA_DIAS	0,50807361
VELOCIDAD	POTENCIA	0,50751299
HOG_EXP_RSK_FOR	PREC_TOTAL	0,5008675

El gráfico nos muestra la definición de correlación de una variable frente a otra, con el coeficiente V de Cramer que muestra el grado de asociación entre ellas.

Hemos señalado en color aquellas relaciones que parecen interesantes analizar, y que posteriormente van a influir en nuestro modelo a la hora de descartar variables redundantes y que nos aportan la misma información.

En color verde quedan señalados los factores de riesgo endógenos relevantes. Como podemos observar, por ejemplo la variable Indicador de Multas, presenta una correlación prácticamente exacta frente a las Variables Uso de Vehículo y el Modo de Contacto. Además, también presenta una alta relación con la profesión del asegurado.

Por otro lado, observamos que la mayoría de las variables que son características del vehículo tienen una alta relación entre sí, como pueden ser el valor del vehículo, el peso del vehículo, la potencia, la cilindrada y la velocidad.

Llama la atención, aquellas variables coloreadas en azul. Éstas son todas aquellas variables exógenas que tienen una alta correlación con la variable Provincia.

Las variables exógenas que presentan esta alta correlación con la Provincia, son variables todas ellas meteorológicas, como son el número de días de precipitaciones, el volumen de precipitaciones totales, los días de granizo, de lluvias, de nieve, de heladas, las rachas y velocidad de viento, y la altitud.

Por ello, podemos interpretar y deducir que la variable territorial Provincia recoge perfectamente las implicaciones meteorológicas a nivel nacional que puedan afectar a un cliente asegurado, ya que presenta altas correlaciones con todas las variables significativas.

Por último, señalar que estas variables meteorológicas también presentan una alta correlación entre ellas.

4.4. Selección de Variables

Finalizados los estudios sobre qué variables son relevantes e irrelevantes, y cuáles son redundantes dentro de nuestro conjunto de datos, el próximo paso a dar es ver que variables seleccionar para posteriormente introducirlas en los modelos de frecuencia y coste medio.

En la práctica ocurre en numerosas ocasiones, que se dispone de un elevado conjunto de variables explicativas que presentan una relación relevante frente a las variables dependientes. Una primera pregunta a plantearse es qué variables introducir en el modelo.

Para ello existen diferentes procedimientos automatizados que permiten elegir el subconjunto de variables que deben estar en el modelo.

Los procedimientos más usuales para la selección de variables son:

- **Método backward**: Este procedimiento comienza por considerar incluidas en el modelo teórico a todas las variables disponibles y se van eliminando del modelo de una en una según su capacidad explicativa. En concreto, la primera variable que se elimina es aquella que presenta menor poder explicativo con la variable dependiente, es decir, aquella que menos reduce la Deviance, y así sucesivamente hasta llegar a una situación óptima.
- **Método forward**: El procedimiento forward es el procedimiento inverso al Método backward, ya que se comienza por un modelo que no contiene ninguna variable explicativa y se añade como primera de ellas aquella que tenga mayor poder explicativo -en valor absoluto- con la variable dependiente.
- **Método stepwise**: El método Stepwise, es uno de los más empleados y consiste en una combinación de los dos anteriores. En el primer paso se procede como en el método forward pero a diferencia de éste en el que cuando una variable entra en el modelo ya no vuelve a salir, en el procedimiento stepwise es posible que la inclusión de una nueva variable haga que otra que ya estaba en el modelo resulte redundante y sea expulsada de él.

El modelo de ajuste al que se llega partiendo del mismo conjunto de variables explicativas es distinto según cuál sea el método de selección de variables.
La consecuencia de este hecho resultante es que ninguno de los algoritmos garantiza encontrar el modelo óptimo.

Por otro lado, también existen medidas de la bondad de ajuste de un modelo que permiten elegir entre diferentes subconjuntos de variables explicativas el "mejor" subconjunto para construir el modelo.

La utilización combinada de los algoritmos de selección de las variables y los criterios de bondad de ajuste permiten seleccionar adecuadamente el modelo óptimo que se debe utilizar.

En nuestro caso práctico, llevaremos a cabo el proceso de selección de variables a través del método de evaluación de la bondad de ajuste de cada variable explicativa frente a cada una de las variables a estudiar.

A través de los procedimientos de SAS, hemos realizado de forma individual un modelo GLM de cada una de las variables explicativas frente a cada una de las variables dependientes, con la finalidad de ver qué factores de riesgo tienen mayor poder explicativo, observando cuales reducen más la Deviance y los criterios de información de ajuste AIC y BIC.

Esta reducción de la Deviance y de los criterios de información se compara entre los factores de riesgo para observar las diferencias, y se calcula el porcentaje de reducción frente a la Deviance, AIC y BIC obtenidos en el modelo realizado sin factores de riesgo (desarrollado en el apartado de ajuste de la distribución de las variables dependientes).

Mostraremos a modo de ejemplo, la metodología seguida con la variable dependiente Frecuencia frente a cada factor de riesgo*

Ajuste de la Distribución	Deviance	AIC	BIC
POISSON	24819	32275	32284

* Con variable coste medio se siguen los mismos pasos.

VARIABLES ENDÓGENAS	ColumnDeviance	ColumnAIC	ColumnBIC	%Deviance	%AIC	%BIC
GRUPOS_MM_DANOS_DIRECTO	24722	32186	32232	-0,39%	-0,27%	-0,16%
valor_veh	24723	32185	32222	-0,39%	-0,28%	-0,19%
Provincia	24724	32298	32848	-0,38%	0,07%	1,75%
CILINDRADA	24739	32199	32226	-0,32%	-0,24%	-0,18%
LITERAL_USO_VEHICULO	24749	32213	32259	-0,28%	-0,19%	-0,08%
IND_MULTAS	24751	32209	32227	-0,27%	-0,20%	-0,18%
N_autofamilia	24754	32212	32231	-0,26%	-0,19%	-0,16%
POTENCIA	24755	32215	32242	-0,26%	-0,19%	-0,13%
peso_veh	24757	32217	32245	-0,25%	-0,18%	-0,12%
VELOCIDAD	24775	32235	32262	-0,18%	-0,12%	-0,07%
LONGITUD	24785	32245	32272	-0,14%	-0,09%	-0,04%
Antiguedad	24792	32254	32291	-0,11%	-0,07%	0,02%
BONUS_APLICADO	24794	32254	32282	-0,10%	-0,06%	-0,01%
LITERAL_km_anual	24794	32254	32282	-0,10%	-0,06%	-0,01%
modalidad_pol	24795	32267	32350	-0,10%	-0,02%	0,20%
MAS_VEHICULOS_UFAMILIAR	24796	32254	32272	-0,09%	-0,06%	-0,04%
PESOPOT	24797	32257	32284	-0,09%	-0,06%	0,00%
edad_hab_char	24801	32263	32300	-0,07%	-0,04%	0,05%
MOTORE	24801	32259	32278	-0,07%	-0,05%	-0,02%
ZONA_DEHABITABILIDAD	24805	32265	32293	-0,06%	-0,03%	0,03%
FORMA_PAGO	24807	32267	32295	-0,05%	-0,02%	0,03%
antig_veh	24808	32266	32284	-0,04%	-0,03%	0,00%
PLAZAS_TX	24810	32268	32286	-0,04%	-0,02%	0,01%
ant_carne	24815	32273	32291	-0,02%	-0,01%	0,02%
PUERTAS	24816	32274	32292	-0,01%	0,00%	0,03%
CONDUCTOR_ES_TOMADOR	24817	32275	32293	-0,01%	0,00%	0,03%
MUNICIPIO	24819	32281	32317	0,00%	0,02%	0,10%
CIUDAD_DORMITORIO	24819	32277	32295	0,00%	0,01%	0,03%
ECIVIL_CONDUCTOR	24819	32277	32296	0,00%	0,01%	0,04%
NACIONALIDAD	24819	32277	32296	0,00%	0,01%	0,04%
GARAJE	24820	32280	32307	0,00%	0,01%	0,07%

En estos gráficos, podemos observar el poder de reducción y de explicación de cada una de las variables independientes frente a la variable de estudio.

Para la selección de variables se tendrá en cuenta aquellas variables que sean relevantes frente a las variables dependientes, y aquellas que no sean redundantes.

En el caso de que dos variables de riesgo presenten alto poder explicativo y alta asociación frente a las variables dependientes, (como por ejemplo, el valor del vehículo y la potencia, o el Indicador de multas y el uso del vehículo), habrá que escoger entre una de ellas en el momento de seleccionar las variables para incorporarlas en el modelo (normalmente aquella que tenga mayor asociación y poder explicativo), ya que no sería óptimo introducir variables redundantes en el modelo.

VARIABLES EXÓGENAS	ColumnDeviance	ColumnAIC	ColumnBIC	%Deviance	%AIC	%BIC
PREC_APR_DIAS_1	24761	32219	32265	-0,23%	-0,17%	-0,06%
PREC_TOTAL_1	24762	32216	32243	-0,23%	-0,18%	-0,13%
GRANIZO_DIAS_1	24768	32224	32261	-0,20%	-0,16%	-0,07%
TER_PESO_1	24770	32230	32286	-0,20%	-0,14%	0,00%
VIENTO_VEL_MED_1	24773	32227	32255	-0,19%	-0,15%	-0,09%
SOLT_PRC_1	24774	32234	32289	-0,18%	-0,13%	0,02%
LOC_ACTIV_PRC_1	24775	32231	32268	-0,18%	-0,14%	-0,05%
LOC_AGRA_PRC_1	24775	32243	32335	-0,18%	-0,10%	0,16%
ALTITUD_MED_1	24775	32229	32257	-0,18%	-0,14%	-0,08%
OFICINAS_PRC_1	24777	32233	32269	-0,17%	-0,13%	-0,05%
SEG_VIV_PRC_1	24778	32232	32260	-0,17%	-0,13%	-0,08%
HELADA_DIAS_1	24779	32235	32271	-0,16%	-0,12%	-0,04%
TAMANO_MED_1	24779	32237	32283	-0,16%	-0,12%	0,00%
PRES_MED_1	24780	32234	32261	-0,16%	-0,13%	-0,07%
RACH_VEL_1	24780	32248	32340	-0,16%	-0,08%	0,17%
NIEVE_DIAS_1	24780	32238	32284	-0,16%	-0,11%	0,00%
RACHA_DIR_1	24781	32237	32274	-0,15%	-0,12%	-0,03%
EST_POS_OBL_PRC_1	24781	32235	32263	-0,15%	-0,12%	-0,07%
NUM_VEHIC_MED_1	24782	32250	32341	-0,15%	-0,08%	0,18%
ALQUILER_PRC_1	24783	32251	32343	-0,14%	-0,07%	0,18%
HOG_EXP_RSK_FOR_1	24783	32239	32276	-0,14%	-0,11%	-0,02%
CASA_PRC_1	24785	32241	32278	-0,14%	-0,10%	-0,02%
HIJOS_MED_1	24786	32242	32278	-0,13%	-0,10%	-0,02%
INSOLA_PRC_1	24786	32242	32279	-0,13%	-0,10%	-0,02%
VIVIEN_TOT_1	24786	32246	32301	-0,13%	-0,09%	0,05%
BOMB_DISTA_1	24786	32254	32346	-0,13%	-0,06%	0,19%
JUV_ACT_IND_1	24788	32256	32348	-0,13%	-0,06%	0,20%
POB_EDAD_MNA_1	24788	32248	32303	-0,12%	-0,08%	0,06%
COND_ECON_1	24789	32245	32281	-0,12%	-0,09%	-0,01%
ACT_20_59_PRC_1	24789	32243	32270	-0,12%	-0,10%	-0,04%
REMPLAZA_IND_1	24790	32244	32271	-0,12%	-0,10%	-0,04%
LLUVIA_DIAS_1	24790	32248	32294	-0,12%	-0,08%	0,03%
TASA_PARO_1	24790	32246	32283	-0,12%	-0,09%	0,00%
EDI_TOT_1	24791	32245	32272	-0,11%	-0,09%	-0,04%
VIUD_PRC_1	24794	32252	32298	-0,10%	-0,07%	0,04%
VIV_ANTIG_MED_1	24794	32262	32354	-0,10%	-0,04%	0,22%
PROP_NPAG_PRC_1	24795	32249	32276	-0,10%	-0,08%	-0,02%
PROP_HER_PRC_1	24795	32249	32276	-0,10%	-0,08%	-0,02%
HOG_DELIN_PRC_1	24796	32252	32288	-0,09%	-0,07%	0,01%
POB_ESP_PRC_1	24796	32252	32289	-0,09%	-0,07%	0,02%
POB_EXT_PRC_1	24796	32252	32289	-0,09%	-0,07%	0,02%
INGRESOSMEDIOS	24803	32265	32302	-0,06%	-0,03%	0,06%
AUTOSNEW	24813	32275	32312	-0,02%	0,00%	0,09%
GASTO_CARBURANTES	24813	32275	32312	-0,02%	0,00%	0,09%
GASTO_TRANSPORTE	24818	32280	32317	0,00%	0,02%	0,10%
MOTOSYCICLOS	24819	32279	32306	0,00%	0,01%	0,07%
DENSIDADPOBLACION	24819	32281	32318	0,00%	0,02%	0,10%
SEGUROS	24820	32282	32318	0,00%	0,02%	0,11%
AUTOS2MANO	24820	32280	32307	0,00%	0,01%	0,07%

4.5. Modelos Lineales Generalizados (GLM)

Tras los análisis exploratorios y el proceso de selección de variables, nos introduciremos en el proceso de elaboración de los modelos lineales generalizados.

Este apartado va a constar de la elaboración de dos modelos.

El primero de ellos lo elaboraremos tan sólo con la selección de variables endógenas, y el segundo modelo incorporará variables exógenas junto con las variables endógenas del primer modelo. En cada uno de ellos, mostraremos las variables seleccionadas, la ejecución del mismo y la validación de los modelos.

MODELO 1: VARIABLES ENDÓGENAS

En primer lugar, la selección de variables ha sido llevada a cabo a través del método explicado en el apartado anterior.

En segundo lugar, tras la realización de diferentes modelos con las variables explicativas, hemos seguido un proceso de simplificación de los modelos descartando variables e introduciendo tan solo aquellas más significativas.

Por lo tanto, las variables finalmente seleccionadas para cada uno de los modelos han sido:

	MODELO DE FRECUENCIA
VARIABLES SELECCIONADAS	Modalidad de la póliza
	Indicador de Multas
	Num.Autos en la familia
	Antigüedad Póliza
	Bonus Aplicado
	Grupo de Vehiculos Daños
	Valor Vehiculo
	Zona Habitabilidad

	MODELO DE COSTE MEDIO
VARIABLES SELECCIONADAS	Indicador de Multas
	Num.Autos en la familia
	Bonus Aplicado
	Velocidad
	Valor Vehiculo
	Zona Habitabilidad

Al tratarse de una modalidad con la peculiaridad de la franquicia, dentro de los modelos hemos seguido la siguiente casuística.
Para el modelo de frecuencia, utilizamos la variable modalidad de la póliza como una variable explicativa más del modelo, donde en cada uno de los niveles de franquicia estima los βetas correspondientes para cada nivel de franquicia.

En cambio, para la variable dependiente del Coste Medio, hemos llevado a cabo la modelización del coste medio que asume la compañía más el coste medio que asume el cliente, modelizando así el incurrido total de los siniestros.

A través de los procedimientos de SAS, podemos llevar a cabo la ejecución de cada uno de los modelos:

- Salidas del Modelo de Frecuencia.

```
proc genmod data=TFM.basefinal_2 order=data plots=all;
CLASS modalidad_pol valor_veh IND_MULTAS N_autofamilia Antiguedad
GRUPOS_MM_DANOS_DIRECTOZONA_DEHABITABILIDAD
BONUS_APLICADO/order=freq ref=first;
model N_SIN_DANOS = modalidad_pol valor_veh IND_MULTAS N_autofamilia
Antiguedad GRUPOS_MM_DANOS_DIRECTO ZONA_DEHABITABILIDAD
BONUS_APLICADO/ dist=poisson link=log type3 offset=L_EXPOS;
output out=DATOS_F p= Freq_Estimada RESDEV=yresid cooksd=Cook
leverage=leverage Betas= Xbeta;
run;
```

Model Information	
Data Set	TFM.BASEFINAL_2
Distribution	Poisson
Link Function	Log
Dependent Variable	N_SIN_DANOS
Offset Variable	L_EXPOS

Number of Observations Read	71059
Number of Observations Used	71059

Class Level Information		
Class	Levels	Values
modalidad_pol	9	p. TRCF 180 p. TRCF 125 p. TRCF 300 p. TRCF 99 p. TRCF 120 p. TRCF 600 p. TRCF 90 p. TRCF 450 p. TRCF 200
valor_veh	4	03. 22000-35800 01. LT 13450 04. MT 35800 02. 13450-22000
IND_MULTAS	2	SI NO
N_autofamilia	2	01. 0 auto 02. + DE 1 auto
Antiguedad	2	NB Cartera
GRUPOS_MM_DANOS_DIRE	5	GRUPO DANOS 3 GRUPO DANOS 4 GRUPO DANOS 1 GRUPO DANOS 5 GRUPO DANOS 2
ZONA_DEHABITABILIDAD	2	b.RURAL a.URBANO
BONUS_APLICADO	3	02. 40-55% 01. LT 40% 03. MT 55%

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	71E3	24431.6578	0.3439
Scaled Deviance	71E3	24431.6578	0.3439
Pearson Chi-Square	71E3	119627.6498	1.6840
Scaled Pearson X2	71E3	119627.6498	1.6840
Log Likelihood		-15942.8289	
Full Log Likelihood		-15942.8289	
AIC (smaller is better)		31929.6578	
AICC (smaller is better)		31929.6720	
BIC (smaller is better)		32131.4256	

Algorithm converged.

Este cuadro nos muestra la información del modelo y las variables que han sido introducidas en él, así como los niveles que presenta cada una de las variables.

En el último cuadro, y quizá el más importante, podemos observar los diferentes criterios de ajuste al modelo como son la Deviance, el AIC y el BIC.

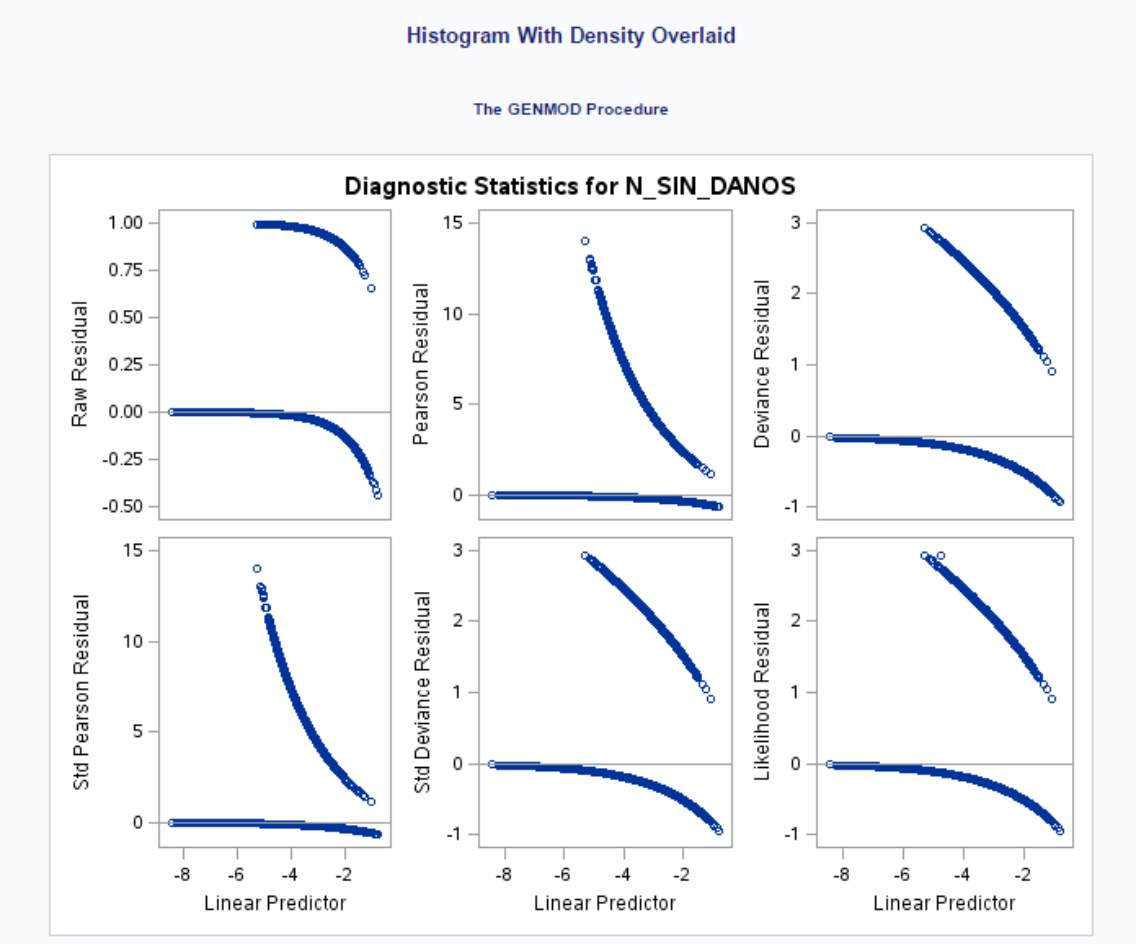
Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > ChiSq
Intercept		1	-2.2274	0.0431	-2.3118	-2.1429	2673.67	<.0001
modalidad_pol	p. TRCF 180	1	0.0228	0.0449	-0.0653	0.1108	0.26	0.6122
modalidad_pol	p. TRCF 125	1	0.1325	0.0838	-0.0318	0.2968	2.50	0.1141
modalidad_pol	p. TRCF 300	1	-0.1057	0.0849	-0.2722	0.0608	1.55	0.2133
modalidad_pol	p. TRCF 99	1	0.2928	0.1415	0.0154	0.5702	4.28	0.0386
modalidad_pol	p. TRCF 120	1	0.1756	0.1242	-0.0678	0.4190	2.00	0.1573
modalidad_pol	p. TRCF 600	1	-0.4250	0.2058	-0.8284	-0.0216	4.26	0.0389
modalidad_pol	p. TRCF 90	1	0.2533	0.2693	-0.2745	0.7811	0.88	0.3469
modalidad_pol	p. TRCF 450	1	-1.2801	1.0003	-3.2405	0.6804	1.64	0.2006
modalidad_pol	p. TRCF 200	0	0.0000	0.0000	0.0000	0.0000	.	.
valor_veh	03. 22000-35800	1	0.1515	0.0400	0.0732	0.2298	14.37	0.0001
valor_veh	01. LT 13450	1	-0.1975	0.0661	-0.3270	-0.0680	8.94	0.0028
valor_veh	04. MT 35800	1	0.3839	0.0583	0.2696	0.4982	43.34	<.0001
valor_veh	02. 13450-22000	0	0.0000	0.0000	0.0000	0.0000	.	.
IND_MULTAS	SI	1	-0.3400	0.0437	-0.4257	-0.2544	60.57	<.0001
IND_MULTAS	NO	0	0.0000	0.0000	0.0000	0.0000	.	.
N_autofamilia	01. 0 auto	1	0.4586	0.0484	0.3638	0.5533	89.95	<.0001
N_autofamilia	02. + DE 1 auto	0	0.0000	0.0000	0.0000	0.0000	.	.
Antiguedad	NB	1	-0.2459	0.0634	-0.3703	-0.1216	15.03	0.0001
Antiguedad	Cartera	0	0.0000	0.0000	0.0000	0.0000	.	.
GRUPOS_MM_DANOS_DIRE	GRUPO DANOS 3	1	0.1274	0.0431	0.0430	0.2118	8.75	0.0031
GRUPOS_MM_DANOS_DIRE	GRUPO DANOS 4	1	0.2057	0.0466	0.1144	0.2970	19.51	<.0001
GRUPOS_MM_DANOS_DIRE	GRUPO DANOS 1	1	-0.0716	0.0664	-0.2018	0.0587	1.16	0.2814
GRUPOS_MM_DANOS_DIRE	GRUPO DANOS 5	1	0.2360	0.0798	0.0795	0.3925	8.74	0.0031
GRUPOS_MM_DANOS_DIRE	GRUPO DANOS 2	0	0.0000	0.0000	0.0000	0.0000	.	.
ZONA_DEHABITABILIDAD	b.RURAL	1	-0.0966	0.0339	-0.1631	-0.0301	8.11	0.0044
ZONA_DEHABITABILIDAD	a.URBANO	0	0.0000	0.0000	0.0000	0.0000	.	.
BONUS_APLICADO	02. 40-55%	1	0.2389	0.0509	0.1392	0.3386	22.05	<.0001
BONUS_APLICADO	01. LT 40%	1	0.5109	0.0789	0.3563	0.6655	41.96	<.0001
BONUS_APLICADO	03. MT 55%	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

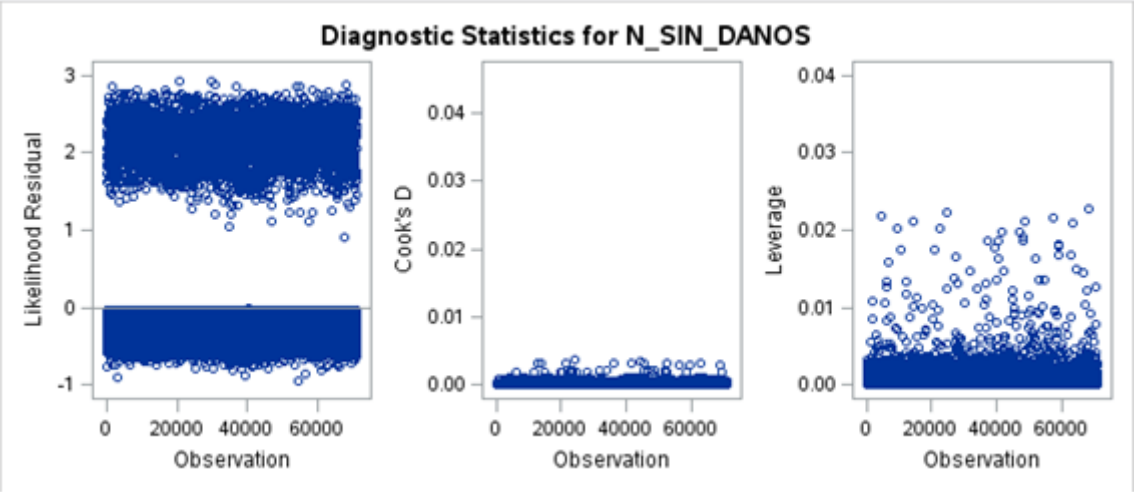
Note: The scale parameter was held fixed.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
modalidad_pol	8	18.79	0.0160
valor_veh	3	63.21	<.0001
IND_MULTAS	1	61.04	<.0001
N_autofamilia	1	82.36	<.0001
Antiguedad	1	15.45	<.0001
GRUPOS_MM_DANOS_DIRE	4	29.85	<.0001
ZONA_DEHABITABILIDAD	1	8.17	0.0043
BONUS_APLICADO	2	51.98	<.0001

A través de los gráficos superiores, sobre las salidas del modelo, podemos observar los estimadores para cada uno de los niveles. La columna "Estimate" nos aporta los βetas que salen del modelo.
Utilizando la fórmula $\mathbf{e^{\beta}}$, obtenemos los factores relativos de recargo o bonificación que se aplican a una tarifa.

Por último, el cuadro de Type 3 Analysis, nos muestra que todos los βetas (Estimates) de nuestro modelo son significativos.





En los gráficos presentes vemos por un lado la convergencia de los predictores lineales del modelo, y por otro lado los gráficos de Cook y Leverage.

La distancia de Cook es una medida que permite detectar las observaciones atípicas en combinación del Leverage y la concordancia del proceso generador de dichas

observaciones con el proceso generador del resto de observaciones, y por otro lado, una observación presenta apalancamiento si está muy alejada del resto de observaciones.

- Salidas del Modelo de Coste Medio.

```
proc genmod data=TFM.basefinal_2 order=data plots=all;
CLASS valor_veh IND_MULTAS N_autofamilia ZONA_DEHABITABILIDAD BONUS_APLICADO VELOCIDAD/order=freq ref=first;
model CosteMedio_F = valor_veh IND_MULTAS N_autofamilia ZONA_DEHABITABILIDAD BONUS_APLICADO VELOCIDAD/ dist=gamma link=log type3 offset=L_EXPOS;
output out=DATOS_CMe p= CMe_Estimado RESDEV=yresid cooksd=Cook leverage=leverage Betas= Xbeta;
run;
```

Model Information	
Data Set	TFM.BASEFINAL_2
Distribution	Gamma
Link Function	Log
Dependent Variable	CosteMedio_F

Number of Observations Read	71059
Number of Observations Used	3727
Missing Values	67332

Class Level Information		
Class	Levels	Values
valor_veh	4	03. 22000-35800 01. LT 13450 04. MT 35800 02. 13450-22000
IND_MULTAS	2	SI NO
N_autofamilia	2	01. 0 auto 02. + DE 1 auto
VELOCIDAD	3	01. LT 178 04. MT 205 02. 178-205
ZONA_DEHABITABILIDAD	2	b.RURAL a.URBANO
BONUS_APLICADO	3	02. 40-55% 01. LT 40% 03. MT 55%

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	3716	2303.3040	0.6198
Scaled Deviance	3716	4068.7079	1.0949
Pearson Chi-Square	3716	3941.6573	1.0607
Scaled Pearson X2	3716	6962.8031	1.8737
Log Likelihood		-29349.6254	
Full Log Likelihood		-29349.6254	
AIC (smaller is better)		58723.2507	
AICC (smaller is better)		58723.3347	
BIC (smaller is better)		58797.9310	

Algorithm converged.

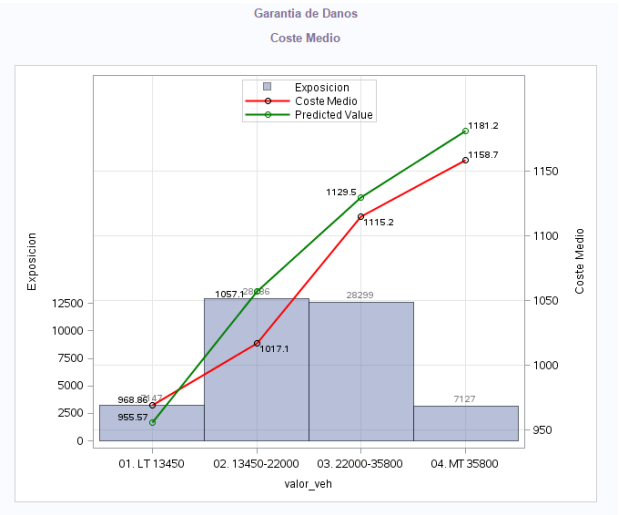
Analysis Of Maximum Likelihood Parameter Estimates							
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits	Wald Chi-Square	Pr > ChiSq
Intercept		1	6.6946	0.0267	6.6422 6.7469	62747.1	<.0001
valor_veh	03. 22000-35800	1	0.0409	0.0296	-0.0171 0.0988	1.91	0.1669
valor_veh	01. LT 13450	1	-0.0518	0.0529	-0.1555 0.0519	0.96	0.3273
valor_veh	04. MT 35800	1	0.0312	0.0449	-0.0569 0.1193	0.48	0.4874
valor_veh	02. 13450-22000	0	0.0000	0.0000	0.0000 0.0000	.	.
IND_MULTAS	SI	1	0.2895	0.0261	0.2384 0.3405	123.43	<.0001
IND_MULTAS	NO	0	0.0000	0.0000	0.0000 0.0000	.	.
N_autofamilia	01. 0 auto	1	0.0912	0.0345	0.0236 0.1589	6.99	0.0082
N_autofamilia	02. + DE 1 auto	0	0.0000	0.0000	0.0000 0.0000	.	.
VELOCIDAD	01. LT 178	1	0.0272	0.0321	-0.0357 0.0902	0.72	0.3967
VELOCIDAD	04. MT 205	1	0.1420	0.0357	0.0721 0.2120	15.84	<.0001
VELOCIDAD	02. 178-205	0	0.0000	0.0000	0.0000 0.0000	.	.
ZONA_DEHABITABILIDAD	b.RURAL	1	0.1291	0.0255	0.0791 0.1791	25.58	<.0001
ZONA_DEHABITABILIDAD	a.URBANO	0	0.0000	0.0000	0.0000 0.0000	.	.
BONUS_APLICADO	02. 40-55%	1	0.1173	0.0377	0.0434 0.1912	9.68	0.0019
BONUS_APLICADO	01. LT 40%	1	0.3271	0.0587	0.2120 0.4421	31.06	<.0001
BONUS_APLICADO	03. MT 55%	0	0.0000	0.0000	0.0000 0.0000	.	.
Scale		1	1.7665	0.0377	1.6941 1.8419		

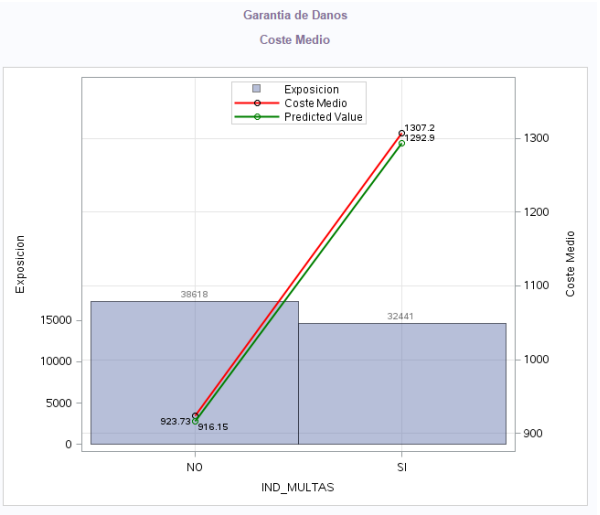
Note: The scale parameter was estimated by maximum likelihood.

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
valor_veh	3	3.64	0.3025
IND_MULTAS	1	123.78	<.0001
N_autofamilia	1	7.12	0.0076
VELOCIDAD	2	15.99	0.0003
ZONA_DEHABITABILIDAD	1	25.76	<.0001
BONUS_APLICADO	2	39.72	<.0001

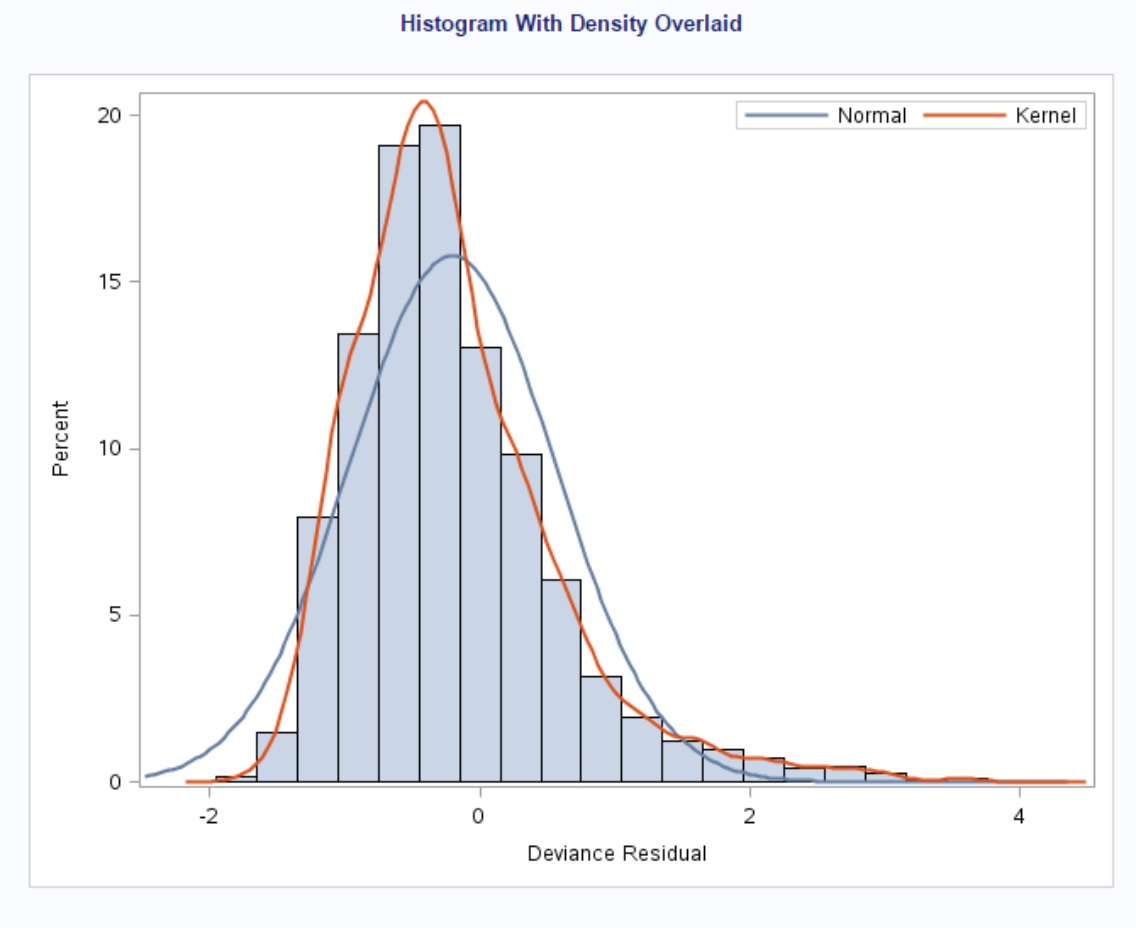
Al igual que en el modelo de Frecuencia, en el modelo de coste medio obtenemos las mismas salidas y los diferentes estimadores para cada una de las variables introducidas en el modelo.

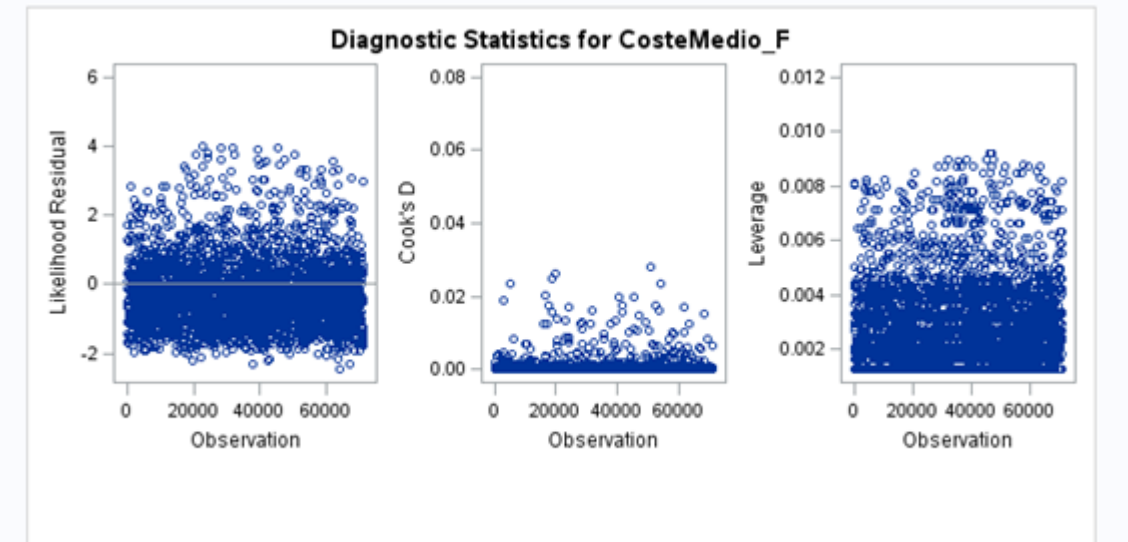
Además, podemos reflejar como ajusta el modelo frente a cada factor de riesgo, teniendo en cuenta el Coste Medio observado frente al Coste Medio estimado.
A modo de ejemplo, vamos a observar el comportamiento de las variables Valor de vehículo e Indicador de Multas.





En el presente gráfico, podemos observar como los residuos del modelo se distribuyen con forma normal.





- Validación de los Modelos

El método empleado y seguido para la validación de los modelos ha sido la validación cruzada.

La validación cruzada (cross validation) es una técnica utilizada para evaluar los resultados de un análisis estadístico y garantizar que son independientes de la partición entre datos de entrenamiento y prueba. Consiste en repetir y calcular la media aritmética obtenida de las medidas de evaluación sobre diferentes particiones.
Esta validación cruzada tiene como objetivo principal la predicción y cómo de preciso es el modelo.

En nuestro caso hemos realizado una validación cruzada en K-iteraciones, esto es, los datos se subdividen en K conjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. El proceso de validación cruzada es repetido durante K iteraciones, con cada uno de los posibles subconjuntos de datos de prueba. Finalmente, se realiza la media aritmética de los resultados de cada iteración para obtener un único resultado. Este método es muy preciso puesto que evaluamos a partir de K combinaciones de datos de entrenamiento y de prueba.

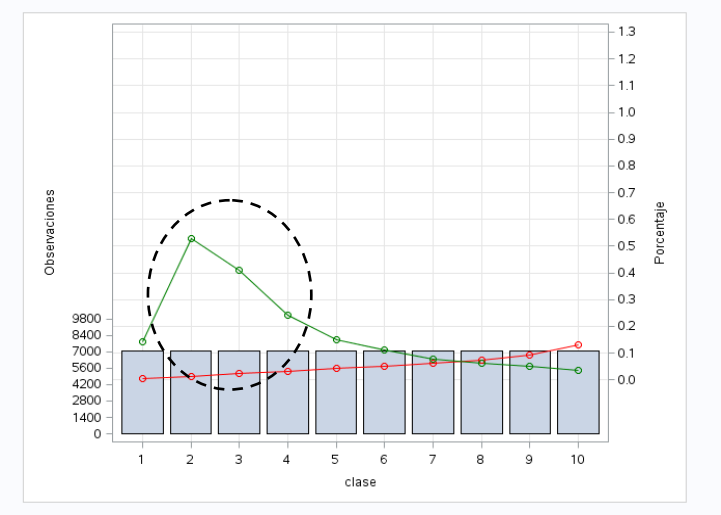
En la práctica realizada, hemos dividido la base de datos en 5 subconjuntos para emplear este método de validación cruzada.
En primer lugar creamos un conjunto aleatorio de 5 niveles, que van a conformar estos 5 subconjuntos que serán utilizados como datos de prueba y entrenamiento, para posteriormente puntuar las observaciones de entrenamiento y validación.
A través de estos pasos, se valida el modelo sobre los datos de la muestra.

GRÁFICOS DE VALIDACIÓN.

Tras la ejecución del método de validación cruzada extraemos los siguientes gráficos de validación.
En el proceso de los gráficos validación se ordenan de menor a mayor los valores estimados y observados y se divide el número total de observaciones en percentiles (en grupos de 10), donde observamos la media de cada valor observado y estimado en cada uno de los percentiles.

Gráfico de Validación de Frecuencia.

En el gráfico de validación representamos la frecuencia observada (línea verde) y la frecuencia estimada (línea roja).



Tras la elaboración del modelo de frecuencia, dentro del gráfico de validación nos encontramos con el problema señalado en el gráfico (el modelo no ajusta bien en los percentiles señalados).

Esto puede deberse a que tal y como se comentó en el análisis exploratorio, la variable Exposición que contempla la base de datos, presenta una distribución en la que la mayor parte de la exposición de las observaciones se acumula en los primeros meses de suscripción de la póliza. Esto puede disparar la frecuencia observada en caso de siniestro, en aquellas observaciones con una baja exposición al riesgo.

Por lo tanto, en los datos analizados se observa una media relativamente baja de exposición (0.44 años), cuando lo normal en una base de datos es que esta media esté en torno 0.65 – 0.70 años.

Gráfico de Validación de Coste Medio.

En el presente gráfico, observamos el Coste medio observado (línea verde) frente al Coste medio estimado (línea roja).
Tal y como muestra el gráfico, el modelo de coste medio se ajusta perfectamente a los datos observados.
En este caso, como el Coste Medio no está influenciado por la variable Exposición como la Frecuencia, no nos encontramos con el problema anterior.

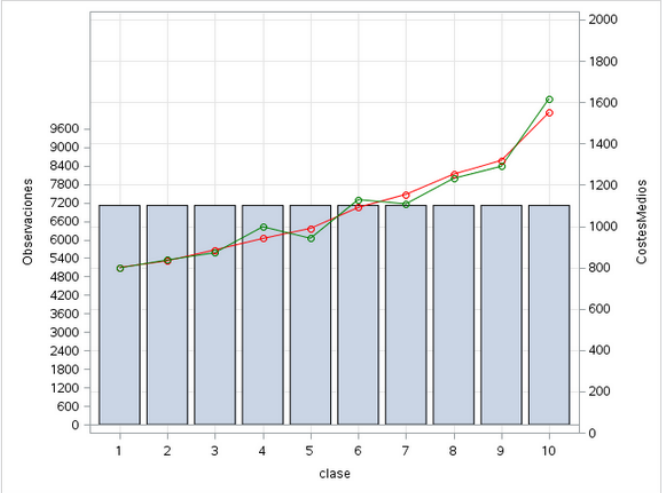
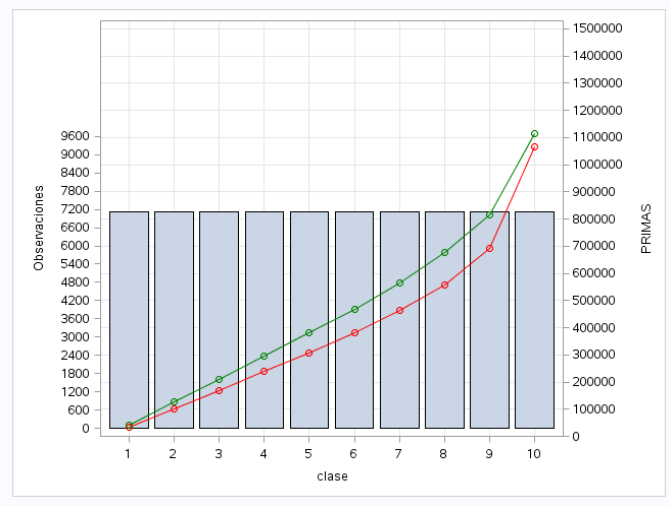


Gráfico de Validación de Prima Pura.

En el gráfico que muestra la combinación de ambos modelos (Frecuencia y Coste Medio), observamos que los modelos establecen una Prima pura estimada (línea roja) inferior a la Prima Pura observada (línea verde). En el cuadro vemos como la Prima Pura estimada es inferior a la observada, por lo que podemos deducir que puede existir un caso de sobre tarificación en nuestra base de datos.



MODELO 2: VARIABLES EXÓGENAS

Una vez analizado el primer modelo sobre la explicación de las variables endógenas, en este segundo modelo seguiremos los mismos pasos que los realizados en el previo.
Tal y como comentamos al inicio de este apartado 4.5., una vez realizado el modelo con variables internas, le incorporamos al modelo las variables exógenas seleccionadas.

Las variables exógenas seleccionadas han sido:

	MODELO DE FRECUENCIA
VARIABLES SELECCIONADAS	Peso del sector terciario
	Tasa de paro
	Días de helada
	Días de precipitación
	Velocidad media del viento
	Locales activos

	MODELO DE COSTE MEDIO
VARIABLES SELECCIONADAS	Días de lluvia
	Peso del sector terciario
	Velocidad máxima de la racha de viento
	Hogares con delincuencia
	Número medio de vehículos
	Segundas viviendas

A continuación expondremos las salidas que nos proporciona SAS, para cada uno de los modelos, siendo éstas las mismas que para el modelo con variables endógenas.

- Salidas del Modelo de Frecuencia.

```
proc genmod data=TFM.basefinal_2 order=data plots=all;
CLASS modalidad_pol valor_veh IND_MULTAS N_autofamilia Antiguedad GRUPOS_MM_DANOS_DIRECTOZONA_DEHABITABILIDAD
BONUS_APLICADO TER_PESO_1 TASA_PARO_1 PREC_APR_DIAS_1 LOC_ACTIV_PRC_1 VIENTO_VEL_MED_1 HELADA_DIAS_1 /order=freq ref=first;
model N_SIN_DANOS = modalidad_pol valor_veh IND_MULTAS N_autofamilia Antiguedad GRUPOS_MM_DANOS_DIRECTO ZONA_DEHABITABILIDAD BONUS_APLICADO TER_PESO_1 TASA_PARO_1 PREC_APR_DIAS_1 LOC_ACTIV_PRC_1 VIENTO_VEL_MED_1 HELADA_DIAS_1/ dist=poisson link=log type3 offset=L_EXPOS;
output out=DATOS_F p= Freq_Estimada RESDEV=yresid cooksd=Cook leverage=leverage Betas= Xbeta;
run;
```

Model Information	
Data Set	TFM.BASEFINAL_2
Distribution	Poisson
Link Function	Log
Dependent Variable	N_SIN_DANOS
Offset Variable	L_EXPOS

Number of Observations Read	71059
Number of Observations Used	70999
Missing Values	60

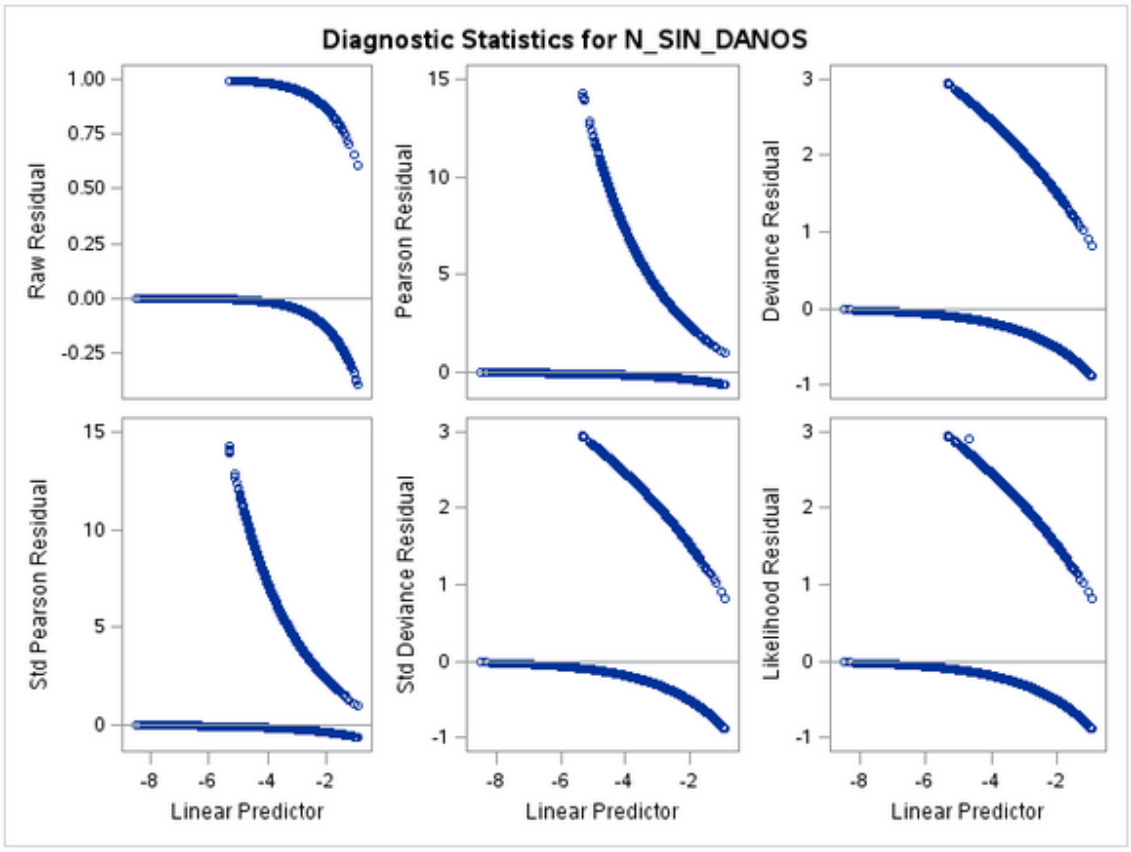
Class Level Information		
Class	Levels	Values
modalidad_pol	9	p. TRCF 180 p. TRCF 125 p. TRCF 300 p. TRCF 99 p. TRCF 120 p. TRCF 600 p. TRCF 90 p. TRCF 450 p. TRCF 200
valor_veh	4	03. 22000-35800 01. LT 13450 04. MT 35800 02. 13450-22000
IND_MULTAS	2	SI NO
N_autofamilia	2	01. 0 auto 02. + DE 1 auto
Antiguedad	2	NB Cartera
GRUPOS_MM_DANOS_DIRE	5	GRUPO DANOS 3 GRUPO DANOS 4 GRUPO DANOS 1 GRUPO DANOS 5 GRUPO DANOS 2
ZONA_DEHABITABILIDAD	2	b.RURAL a.URBANO
BONUS_APLICADO	3	02. 40-55% 01. LT 40% 03. MT 55%
TER_PESO_1	6	5 4 3 2 1 6
TASA_PARO_1	4	4 2 1 3
PREC_APR_DIAS_1	5	5 2 4 1 3
LOC_ACTIV_PRC_1	4	3 2 4 1
VIENTO_VEL_MED_1	3	1 2 3
HELADA_DIAS_1	4	3 1 4 2

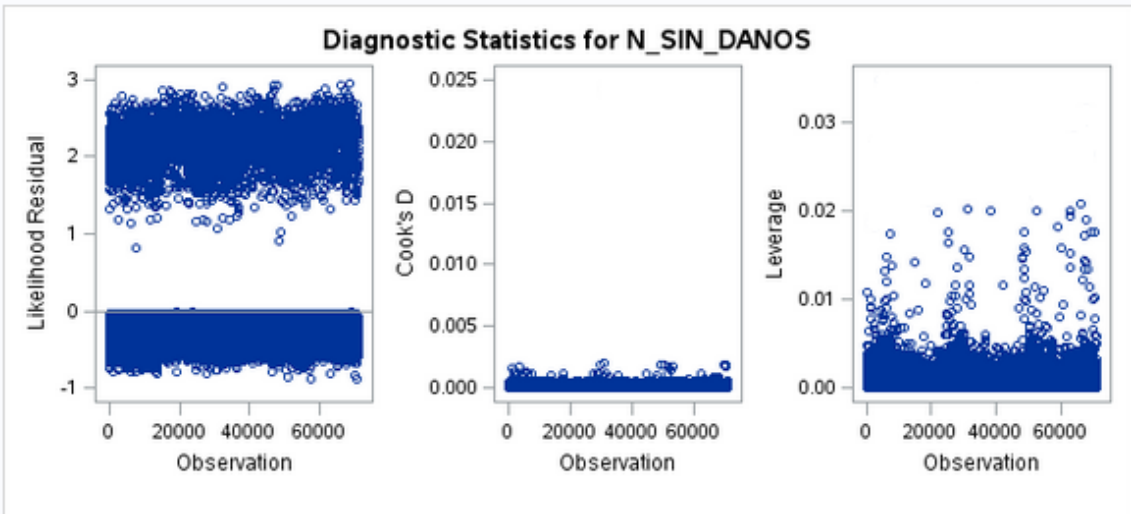
Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	71E3	24340.8170	0.3430
Scaled Deviance	71E3	24340.8170	0.3430
Pearson Chi-Square	71E3	119460.7497	1.6836
Scaled Pearson X2	71E3	119460.7497	1.6836
Log Likelihood		-15894.4085	
Full Log Likelihood		-15894.4085	
AIC (smaller is better)		31872.8170	
AICC (smaller is better)		31872.8679	
BIC (smaller is better)		32257.9746	

Algorithm converged.

Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > Chi Sq
Intercept		1	-2.3554	0.0663	-2.4853	-2.2256	1263.84	<.0001
modalidad_pol	p. TRCF 180	1	0.0149	0.0450	-0.0734	0.1032	0.11	0.7406
modalidad_pol	p. TRCF 125	1	0.1247	0.0839	-0.0398	0.2892	2.21	0.1372
modalidad_pol	p. TRCF 300	1	-0.1249	0.0853	-0.2921	0.0422	2.15	0.1430
modalidad_pol	p. TRCF 99	1	0.2907	0.1417	0.0129	0.5684	4.21	0.0402
modalidad_pol	p. TRCF 120	1	0.1790	0.1243	-0.0646	0.4227	2.08	0.1497
modalidad_pol	p. TRCF 600	1	-0.4326	0.2061	-0.8364	-0.0287	4.41	0.0358
modalidad_pol	p. TRCF 90	1	0.2511	0.2694	-0.2769	0.7792	0.87	0.3513
modalidad_pol	p. TRCF 450	1	-1.2966	1.0003	-3.2572	0.6640	1.68	0.1949
modalidad_pol	p. TRCF 200	0	0.0000	0.0000	0.0000	0.0000	.	.
valor_veh	03. 22000-35800	1	0.1516	0.0401	0.0730	0.2301	14.31	0.0002
valor_veh	01. LT 13450	1	-0.2007	0.0661	-0.3303	-0.0711	9.22	0.0024
valor_veh	04. MT 35800	1	0.3672	0.0589	0.2519	0.4826	38.93	<.0001
valor_veh	02. 13450-22000	0	0.0000	0.0000	0.0000	0.0000	.	.
IND_MULTAS	SI	1	-0.3355	0.0438	-0.4213	-0.2498	58.81	<.0001
IND_MULTAS	NO	0	0.0000	0.0000	0.0000	0.0000	.	.
N_autofamilia	01. 0 auto	1	0.4638	0.0484	0.3689	0.5588	91.66	<.0001
N_autofamilia	02. + DE 1 auto	0	0.0000	0.0000	0.0000	0.0000	.	.
Antiguedad	NB	1	-0.2383	0.0635	-0.3628	-0.1138	14.08	0.0002
Antiguedad	Cartera	0	0.0000	0.0000	0.0000	0.0000	.	.
GRUPOS_MM_DANOS_DIRE	GRUPO DANOS 3	1	0.1244	0.0431	0.0399	0.2089	8.32	0.0039
GRUPOS_MM_DANOS_DIRE	GRUPO DANOS 4	1	0.2030	0.0467	0.1115	0.2945	18.92	<.0001
GRUPOS_MM_DANOS_DIRE	GRUPO DANOS 1	1	-0.0664	0.0665	-0.1967	0.0639	1.00	0.3180
GRUPOS_MM_DANOS_DIRE	GRUPO DANOS 5	1	0.2336	0.0800	0.0768	0.3903	8.53	0.0035
GRUPOS_MM_DANOS_DIRE	GRUPO DANOS 2	0	0.0000	0.0000	0.0000	0.0000	.	.
ZONA_DEHABITABILIDAD	b.RURAL	1	-0.0467	0.0397	-0.1245	0.0311	1.38	0.2397
ZONA_DEHABITABILIDAD	a.URBANO	0	0.0000	0.0000	0.0000	0.0000	.	.
BONUS_APLICADO	02. 40-55%	1	0.2369	0.0510	0.1370	0.3368	21.61	<.0001
BONUS_APLICADO	01. LT 40%	1	0.4960	0.0792	0.3409	0.6512	39.27	<.0001
BONUS_APLICADO	03. MT 55%	0	0.0000	0.0000	0.0000	0.0000	.	.
TER_PESO_1	5	1	-0.0022	0.0460	-0.0924	0.0879	0.00	0.9613
TER_PESO_1	4	1	-0.0410	0.0502	-0.1393	0.0574	0.67	0.4144
TER_PESO_1	3	1	-0.0855	0.0587	-0.2006	0.0295	2.12	0.1450
TER_PESO_1	2	1	0.0614	0.0905	-0.1160	0.2388	0.46	0.4975
TER_PESO_1	1	1	-0.1003	0.1279	-0.3509	0.1503	0.62	0.4328
TER_PESO_1	6	0	0.0000	0.0000	0.0000	0.0000	.	.
TASA_PARO_1	4	1	0.0784	0.0415	-0.0030	0.1597	3.57	0.0590
TASA_PARO_1	2	1	-0.0207	0.0555	-0.1296	0.0881	0.14	0.7089
TASA_PARO_1	1	1	-0.1288	0.0825	-0.2904	0.0329	2.44	0.1185
TASA_PARO_1	3	0	0.0000	0.0000	0.0000	0.0000	.	.
PREC_APR_DIAS_1	5	1	0.0929	0.0418	0.0110	0.1748	4.94	0.0262
PREC_APR_DIAS_1	2	1	0.0929	0.0592	-0.0230	0.2088	2.47	0.1162
PREC_APR_DIAS_1	4	1	0.0929	0.0632	-0.0311	0.2168	2.16	0.1419
PREC_APR_DIAS_1	1	1	-0.1978	0.0736	-0.3421	-0.0535	7.22	0.0072
PREC_APR_DIAS_1	3	0	0.0000	0.0000	0.0000	0.0000	.	.
LOC_ACTIV_PRC_1	3	1	0.1489	0.0426	0.0655	0.2323	12.24	0.0005
LOC_ACTIV_PRC_1	2	1	0.0407	0.0437	-0.0449	0.1263	0.87	0.3511
LOC_ACTIV_PRC_1	4	1	-0.0230	0.0638	-0.1480	0.1020	0.13	0.7182
LOC_ACTIV_PRC_1	1	0	0.0000	0.0000	0.0000	0.0000	.	.
VIENTO_VEL_MED_1	1	1	0.0991	0.0486	0.0039	0.1943	4.16	0.0413
VIENTO_VEL_MED_1	2	1	0.0576	0.0680	-0.0757	0.1909	0.72	0.3972
VIENTO_VEL_MED_1	3	0	0.0000	0.0000	0.0000	0.0000	.	.
HELADA_DIAS_1	3	1	0.0452	0.0489	-0.0506	0.1410	0.86	0.3549
HELADA_DIAS_1	1	1	0.0352	0.0441	-0.0512	0.1216	0.64	0.4244
HELADA_DIAS_1	4	1	-0.0747	0.0974	-0.2657	0.1163	0.59	0.4433
HELADA_DIAS_1	2	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		0	1.0000	0.0000	1.0000	1.0000		

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > Chi Sq
modalidad_pol	8	19.25	0.0136
valor_veh	3	59.58	<.0001
IND_MULTAS	1	59.27	<.0001
N_autofamilia	1	83.86	<.0001
Antiguedad	1	14.47	0.0001
GRUPOS_MM_DANOS_DIRE	4	28.51	<.0001
ZONA_DEHABITABILIDAD	1	1.39	0.2391
BONUS_APLICADO	2	49.47	<.0001
TER_PESO_1	5	4.56	0.4719
TASA_PARO_1	3	7.10	0.0686
PREC_APR_DIAS_1	4	20.65	0.0004
LOC_ACTIV_PRC_1	3	13.60	0.0035
VIENTO_VEL_MED_1	2	4.18	0.1239
HELADA_DIAS_1	3	2.16	0.5399


Diagnostic Statistics for N_SIN_DANOS
Raw Residual
Pearson Residual
Deviance Residual
Std Pearson Residual
Std Deviance Residual
Likelihood Residual
Linear Predictor
Linear Predictor
Linear Predictor


Diagnostic Statistics for N_SIN_DANOS
Likelihood Residual
Cook's D
Leverage
Observation
Observation
Observation

- Salidas del Modelo de Coste Medio.

```
proc genmod data=TFM.basefinal_2 order=data plots=all;
CLASS valor_veh IND_MULTAS N_autofamilia ZONA_DEHABITABILIDAD BONUS_APLICADO VELOCIDAD LLUVIA_DIAS_1 TER_PESO_1 HOG_DELIN_PRC_1 SEG_VIV_PRC_1 RACH_VEL_1 NUM_VEHIC_MED_1/order=freq ref=first;
model CosteMedio_F = valor_veh IND_MULTAS N_autofamilia ZONA_DEHABITABILIDAD BONUS_APLICADO VELOCIDAD LLUVIA_DIAS_1 TER_PESO_1 HOG_DELIN_PRC_1 SEG_VIV_PRC_1 RACH_VEL_1 NUM_VEHIC_MED_1/ dist=gamma link=log type3 offset=L_EXPOS;
output out=DATOS_CMe p= CMe_Estimado RESDEV=yresid cooksd=Cook leverage=leverage Betas= Xbeta;
run;
```

Model Information	
Data Set	TFM.BASEFINAL_2
Distribution	Gamma
Link Function	Log
Dependent Variable	CosteMedio_F

Number of Observations Read	71059
Number of Observations Used	3724
Missing Values	67335

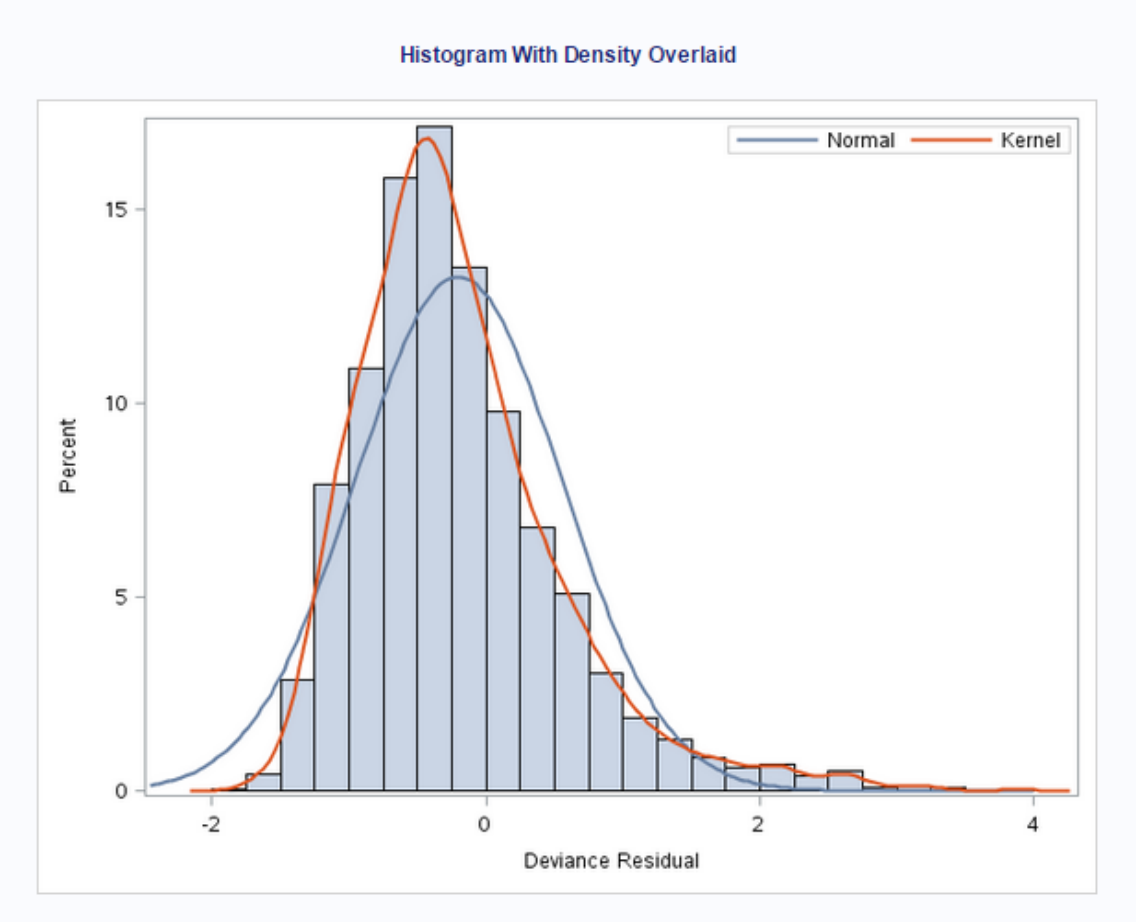
Class Level Information		
Class	Levels	Values
BONUS_APLICADO	3	02. 40-55% 01. LT 40% 03. MT 55%
IND_MULTAS	2	SI NO
N_autofamilia	2	01. 0 auto 02. + DE 1 auto
ZONA_DEHABITABILIDAD	2	b.RURAL a.URBANO
PESOPOT	3	01. 0-10 03. + 16 02. 10-16
peso_veh	3	02. 1140-1331 01. LT 1140 03. MT 1331
LLUVIA_DIAS_1	5	4 1 3 2 5
TER_PESO_1	6	5 4 3 2 1 6
HOG_DELIN_PRC_1	4	3 2 1 4
SEG_VIV_PRC_1	3	3 1 2
RACH_VEL_1	10	7 6 4 8 3 2 9 1 5 10
NUM_VEHIC_MED_1	10	9 8 7 6 5 4 3 2 1 10

Criteria For Assessing Goodness Of Fit			
Criterion	DF	Value	Value/DF
Deviance	3682	2249.7986	0.6110
Scaled Deviance	3682	4058.6840	1.1023
Pearson Chi-Square	3682	3705.5138	1.0064
Scaled Pearson X2	3682	6684.8246	1.8155
Log Likelihood		-29278.7470	
Full Log Likelihood		-29278.7470	
AIC (smaller is better)		58643.4940	
AICC (smaller is better)		58644.5222	
BIC (smaller is better)		58911.0638	

Algorithm converged.

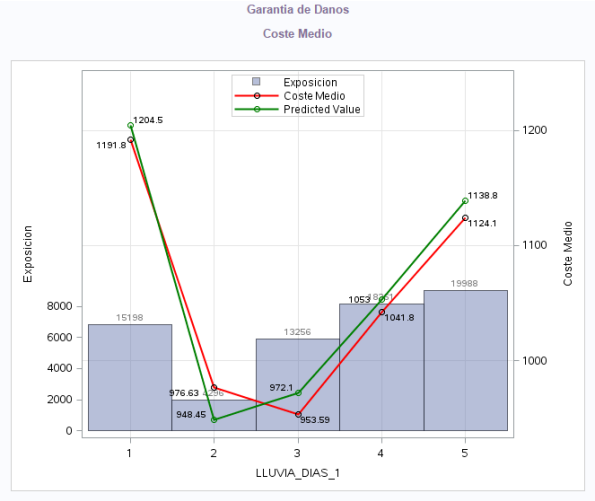
Analysis Of Maximum Likelihood Parameter Estimates								
Parameter		DF	Estimate	Standard Error	Wald 95% Confidence Limits		Wald Chi-Square	Pr > Chi Sq
Intercept		1	6.7992	0.0594	6.6828	6.9156	13114.1	<.0001
BONUS_APLICADO	02. 40-55%	1	0.1252	0.0375	0.0516	0.1988	11.12	0.0009
BONUS_APLICADO	01. LT 40%	1	0.3486	0.0588	0.2333	0.4639	35.11	<.0001
BONUS_APLICADO	03. MT 55%	0	0.0000	0.0000	0.0000	0.0000	.	.
IND_MULTAS	SI	1	0.2933	0.0263	0.2419	0.3448	124.73	<.0001
IND_MULTAS	NO	0	0.0000	0.0000	0.0000	0.0000	.	.
N_autofamilia	01. 0 auto	1	0.0917	0.0344	0.0243	0.1592	7.10	0.0077
N_autofamilia	02. + DE 1 auto	0	0.0000	0.0000	0.0000	0.0000	.	.
ZONA_DEHABITABILIDAD	b.RURAL	1	0.0372	0.0333	-0.0281	0.1024	1.25	0.2641
ZONA_DEHABITABILIDAD	a.URBANO	0	0.0000	0.0000	0.0000	0.0000	.	.
PESOPOT	01. 0-10	1	0.1372	0.0322	0.0741	0.2003	18.16	<.0001
PESOPOT	03. + 16	1	-0.0352	0.0598	-0.1524	0.0820	0.35	0.5566
PESOPOT	02. 10-16	0	0.0000	0.0000	0.0000	0.0000	.	.
peso_veh	02. 1140-1331	1	-0.0296	0.0289	-0.0862	0.0271	1.05	0.3063
peso_veh	01. LT 1140	1	-0.0953	0.0355	-0.1650	-0.0256	7.19	0.0073
peso_veh	03. MT 1331	0	0.0000	0.0000	0.0000	0.0000	.	.
LLUVIA_DIAS_1	4	1	-0.0917	0.0419	-0.1738	-0.0095	4.79	0.0287
LLUVIA_DIAS_1	1	1	0.0585	0.0419	-0.0235	0.1406	1.95	0.1621
LLUVIA_DIAS_1	3	1	-0.1640	0.0458	-0.2536	-0.0743	12.84	0.0003
LLUVIA_DIAS_1	2	1	-0.1808	0.0599	-0.2982	-0.0634	9.11	0.0025
LLUVIA_DIAS_1	5	0	0.0000	0.0000	0.0000	0.0000	.	.
TER_PESO_1	5	1	0.0490	0.0349	-0.0194	0.1174	1.97	0.1603
TER_PESO_1	4	1	0.0394	0.0392	-0.0375	0.1163	1.01	0.3151
TER_PESO_1	3	1	0.0657	0.0466	-0.0253	0.1568	2.00	0.1572
TER_PESO_1	2	1	0.0126	0.0716	-0.1278	0.1529	0.03	0.8606
TER_PESO_1	1	1	0.2532	0.0997	0.0578	0.4485	6.45	0.0111
TER_PESO_1	6	0	0.0000	0.0000	0.0000	0.0000	.	.
HOG_DELIN_PRC_1	3	1	0.0525	0.0300	-0.0064	0.1114	3.06	0.0804
HOG_DELIN_PRC_1	2	1	0.0561	0.0477	-0.0374	0.1495	1.38	0.2395
HOG_DELIN_PRC_1	1	1	0.0633	0.1197	-0.1713	0.2979	0.28	0.5968
HOG_DELIN_PRC_1	4	0	0.0000	0.0000	0.0000	0.0000	.	.
SEG_VIV_PRC_1	3	1	-0.0674	0.0294	-0.1250	-0.0097	5.24	0.0221
SEG_VIV_PRC_1	1	1	0.0294	0.0393	-0.0476	0.1064	0.56	0.4545
SEG_VIV_PRC_1	2	0	0.0000	0.0000	0.0000	0.0000	.	.
RACH_VEL_1	7	1	-0.0057	0.0577	-0.1188	0.1075	0.01	0.9216
RACH_VEL_1	6	1	-0.0439	0.0535	-0.1488	0.0609	0.67	0.4115
RACH_VEL_1	4	1	0.0027	0.0591	-0.1132	0.1186	0.00	0.9636
RACH_VEL_1	8	1	-0.0460	0.0594	-0.1625	0.0705	0.60	0.4391
RACH_VEL_1	3	1	0.0156	0.0585	-0.0990	0.1301	0.07	0.7898
RACH_VEL_1	2	1	-0.1018	0.0587	-0.2168	0.0132	3.01	0.0828
RACH_VEL_1	9	1	-0.1544	0.0563	-0.2646	-0.0441	7.52	0.0061
RACH_VEL_1	1	1	-0.0248	0.0559	-0.1343	0.0847	0.20	0.6567
RACH_VEL_1	5	1	0.0104	0.0646	-0.1162	0.1371	0.03	0.8719
RACH_VEL_1	10	0	0.0000	0.0000	0.0000	0.0000	.	.
NUM_VEHIC_MED_1	9	1	0.0552	0.0439	-0.0308	0.1412	1.58	0.2086
NUM_VEHIC_MED_1	8	1	0.0532	0.0482	-0.0413	0.1476	1.22	0.2697
NUM_VEHIC_MED_1	7	1	0.0435	0.0481	-0.0509	0.1379	0.82	0.3662
NUM_VEHIC_MED_1	6	1	0.0721	0.0513	-0.0285	0.1728	1.97	0.1601
NUM_VEHIC_MED_1	5	1	-0.0340	0.0530	-0.1379	0.0700	0.41	0.5219
NUM_VEHIC_MED_1	4	1	-0.0213	0.0540	-0.1271	0.0845	0.16	0.6931
NUM_VEHIC_MED_1	3	1	0.0657	0.0595	-0.0509	0.1823	1.22	0.2693
NUM_VEHIC_MED_1	2	1	-0.0884	0.0598	-0.2055	0.0287	2.19	0.1391
NUM_VEHIC_MED_1	1	1	-0.0302	0.0666	-0.1607	0.1004	0.21	0.6505
NUM_VEHIC_MED_1	10	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale		1	1.8040	0.0386	1.7300	1.8812		

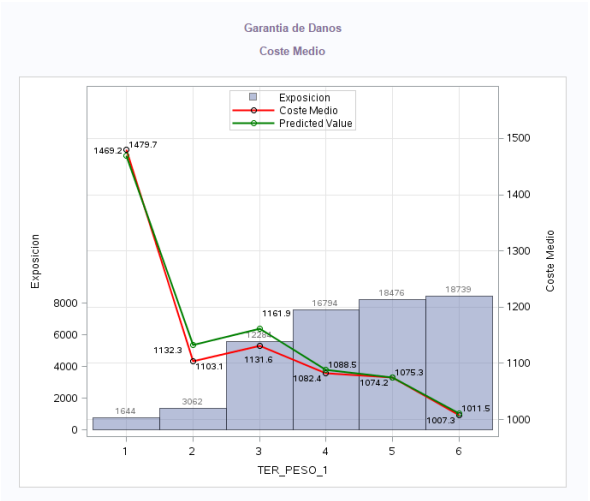
LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > Chi Sq
BONUS_APLICADO	2	45.11	<.0001
IND_MULTAS	1	125.03	<.0001
N_autofamilia	1	7.23	0.0072
ZONA_DEHABITABILIDAD	1	1.25	0.2638
PESOPOT	2	19.46	<.0001
peso_veh	2	7.14	0.0282
LLUVIA_DIAS_1	4	33.43	<.0001
TER_PESO_1	5	8.27	0.1420
HOG_DELIN_PRC_1	3	3.27	0.3519
SEG_VIV_PRC_1	2	6.92	0.0315
RACH_VEL_1	9	15.03	0.0900
NUM_VEHIC_MED_1	9	13.21	0.1532



El hecho de incorporar variables exógenas al modelo, no daña la normalidad de los residuos del modelo.

A modo de ejemplo presentamos la visualización de cómo se ajustan en el modelo las variables externas peso del sector terciario y número de días de lluvia.





GRÁFICOS DE VALIDACIÓN

Al igual que en el primer modelo de variables endógenas, presentamos los gráficos de validación del segundo modelo con la inclusión de variables exógenas.
En el modelo de frecuencia, nos encontramos con el mismo error que en el modelo anterior.
En el modelo de coste medio, observamos un buen ajuste, pero con ciertas diferencias en relación al modelo anterior, ya que este estima un coste medio superior.
En el gráfico de prima pura se ve que en este modelo se recarga ligeramente frente al modelo de Variables endógenas, esto es debido al sobreajuste que realiza sobre el Coste Medio.

Gráfico de Validación de Frecuencia.

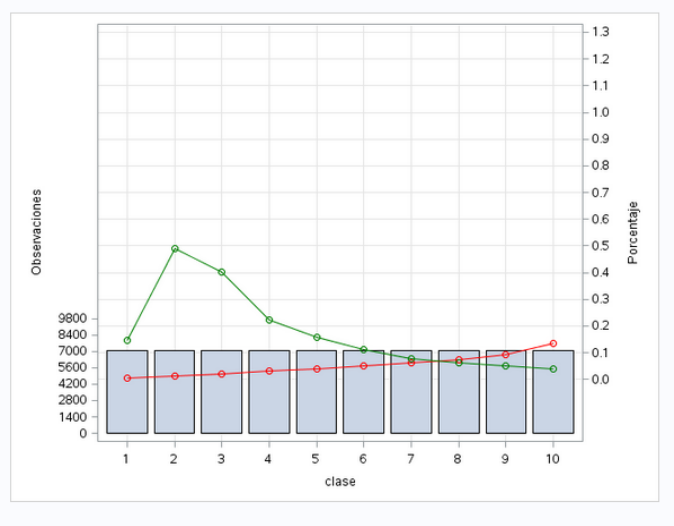


Gráfico de Validación de Coste Medio.

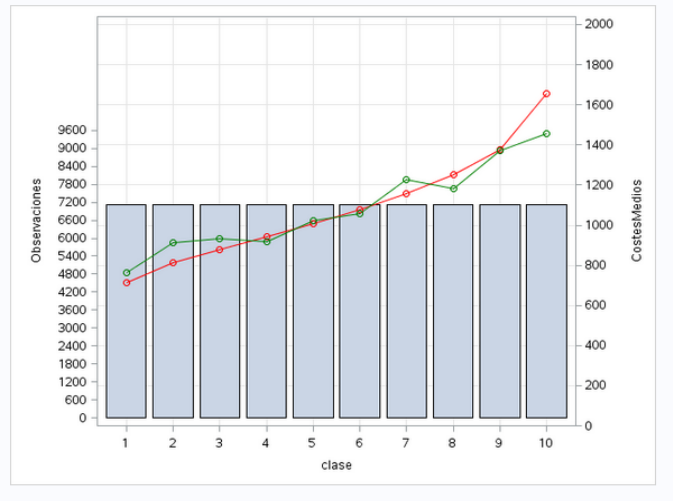
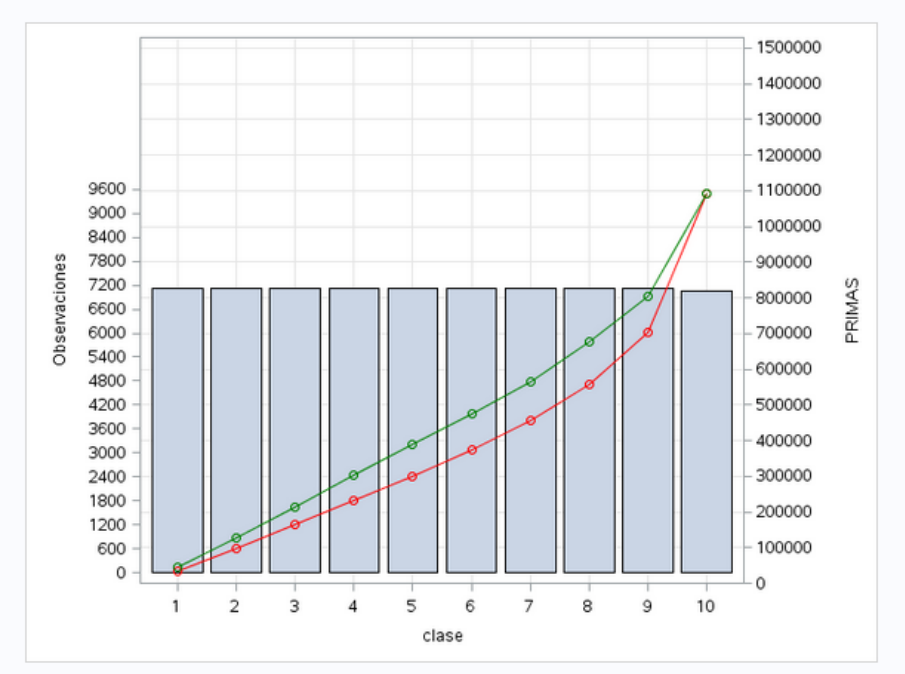


Gráfico de Validación de Prima Pura.



CAPITULO 5. ANÁLISIS DE RESULTADOS Y CONCLUSIONES

El presente trabajo ha plasmado la idea principal sobre la correlación de factores de riesgo con variables dependientes, y la elaboración práctica de un modelo GLM, en primer lugar con variables endógenas y posteriormente con la inclusión de variables exógenas.

En este apartado se trata de reflexionar sobre los resultados obtenidos y ver qué conclusiones se extraen del proceso elaborado.

En primer lugar, se hará hincapié en los resultados desprendidos de los modelos sobre los datos de ajuste a cada uno de ellos.

FRECUENCIA	DEVIANCE	AIC	BIC
Modelo de Ajuste	24819	32275	32284
Modelo Variables Endógenas	24431	31929	32131
Modelo Variables Exógenas	24340	31872	32257

REDUCCIÓN DE VARIABILIDAD	DEVIANCE	AIC	BIC
Modelo Variables Endógenas	-1,6%	-1,1%	-0,5%
Modelo Variables Exógenas	-1,9%	-1,2%	-0,1%

En el modelo de frecuencia, podemos ver cómo la inclusión de variables exógenas reduce la Deviance y el AIC, cierto es que como se ve reflejado, lo hace en un porcentaje pequeño (-0.3%) en relación al modelo que solo alberga variables endógenas.

COSTE MEDIO	DEVIANCE	AIC	BIC
Modelo de Ajuste	3774	60753	60766
Modelo Variables Endógenas	2303	58723	58797
Modelo Variables Exógenas	2248	58644	58911

REDUCCIÓN DE VARIABILIDAD	DEVIANCE	AIC	BIC
Modelo Variables Endógenas	-39,0%	-3,3%	-3,2%
Modelo Variables Exógenas	-40,4%	-3,5%	-3,1%

Al igual que en el modelo de frecuencia, en el modelo de Coste Medio la inclusión de variables exógenas reduce más la Deviance y el AIC, pero también en un pequeño porcentaje (-1.4%).

Por otro lado, podemos ver como la prima especificada en la base de datos es muy superior al incurrido asumido por la compañía aseguradora, por lo que podemos decir que hay un sobrecargo en las tarifas. Además, las primas estimadas se ajustan en un 85% a la prima observada.

Los modelos ajustan perfectamente (ligeramente por encima), el coste total ocurrido por los siniestros declarados.

	AJUSTE DEL MODELO
Σ Coste Medio	3.275.683,71 €
Σ Coste Medio Con Franquicia	4.003.807,71 €
Σ Prima de Daños	4.695.383,69 €
Σ Prima Estimada	4.007.811,80 €
Σ Prima Estimada Var.Externas	4.016.137,00 €

Por lo tanto, llevados a cabo y finalizados todos los análisis, llegamos a las siguientes conclusiones:

1. Tal y como ha sido posible observar en los gráficos de análisis bivariable y en el proceso de Stepwise, las variables endógenas o características de póliza tienen mayor relevancia que las variables exógenas ya que tienen mayor poder de correlación con las variables frecuencia y coste medio, recogen mayor grado de asociación y explican más.

2. Como es posible observar, el factor geográfico juega un papel muy importante dentro del ámbito de los seguros.
   A través del análisis V de Cramer, se puede ver cómo la variable Provincia explica prácticamente la misma información que las variables meteorológicas, por lo que recoge el comportamiento de los aspectos meteorológicos a nivel nacional.
   Además, es una variable que explica claramente la asociación con la Frecuencia y el Coste Medio. Dicha variable no ha sido incluida en el modelo por la cantidad de niveles que presenta y que no favorece la simplificación del modelo. Además es una variable sobre la que no pueden realizarse técnicas de segmentación o agrupación de niveles, ya que si se realiza este proceso, se daña el componente geográfico y territorial.

3. Al comparar los resultados de los modelos GLM realizados, se aprecia la importancia y relevancia de las variables endógenas a la hora de establecer una tarifa de autos.
   Las variables exógenas ayudan y tienen un carácter complementario sobre las variables endógenas, pero no sustitutivo.
   Tienen un carácter complementario y no sustitutivo, ya que la mayoría de las variables endógenas tienen mayor correlación y mayor poder de explicación sobre las variables dependientes llevadas a estudio.
   Bien es cierto, que la inclusión de las variables exógenas aporta aún más a la explicación y predicción del modelo, tal y como reflejan los criterios de ajuste observados en los modelos.

4. Las variables endógenas introducidas en el modelo se ajustan prácticamente a la perfección frente a los siniestros observados declarados ante la compañía, por lo que estas variables tienen un poder de ajuste tan elevado y un nivel de predicción tan exhaustivo que la inclusión de variables exógenas tienen poco recorrido.
   Por lo tanto, las variables exógenas ayudan a explicar levemente las variables dependientes, ya que las variables endógenas tienen un poder predictivo y explicativo muy alto.

5. Una vez terminado todo el proceso y realizadas las consideraciones oportunas, es necesario reflexionar en próximos pasos que pueden llevarse a cabo:

   a. Extender estos análisis realizados a otras modalidades del seguro de auto, con la finalidad de ver si las variables exógenas tienen mayor efecto que en la modalidad estudiada.

   b. Analizar este estudio realizado sobre una base de datos más amplia que recoja mayor información y nos permita llegar a conclusiones más robustas.

CAPITULO 6. REFERENCIAS BIBLIOGRÁFICAS

- SAS Institute Trainig Courses:

  www.sas.com

  "Course SAS Programming 1: Essentials"

  https://support.sas.com/edu/schedules.html?ctry=gb&id=277

  Ron Cody., 2007. "Learning Sas By Example A Programmers Guide"

  "Course Statistics 1: Introduction to ANOVA, Regression, and Logistic Regression"

  https://support.sas.com/edu/schedules.html?id=1979&ctry=GB&view=nearby

  Ron Cody., 2011. "SAS Statistics by Example"

- Books & Papers:

  Samprit Chattefuee, Ali S. Hadi., 2006. "Regression Analysis By Example" (4^{th} edition)

  Yiu Kuen Tse., 2009. "Non-Life Actuarial Models Theory, Methods and Evaluation"

  Klugman, Stuart. A. 2004. "Loss Models: From Data to Decisions" (3^{rd} Edition)

  Annette J. Dobson.2002. "An Introduction to Generalized Linear Models" (2^{nd} Edition)

  Michel Denuit, Xavier Maréchal, Sandra Pitrebois, Jean-François Walhin., 2007. "Actuarial Modelling of Claim Counts: Risk Classification, Credibility and Bonus-Malus Systems".

  Esbjörn Ohlsson, Björn Johansson., 2010., "Non-life Insurance Pricing with Generalized Linear Models"

  Jong and Heller., 2008. "Generalized Linear Models for Insurance Data"

Edward W. Frees, Richard A. Derrig, Glenn Meyers., 2014. "Predictive modeling applications in actuarial science: Volume I: Predictive Modeling Techniques"

M.A. Keyser., IAE., 2015. "Tarificación GLM en Autos y Hogar con SAS: Introducción a los modelos lineales generalizados (GLM) en SAS"

Towers Watson. 2010. "Modelos GLM: Aplicación orientada a la Tarificación"

Milliman., 2009. "Motor Pure Premium Modeling with Deductible"

Eva Boj del Val, Mª Mercè Claramunt y Josep Fortiana. (Fundación Mapfre). 2004. "Análisis multivariante aplicado a la selección de factores de riesgo en la tarificación"

Eva Boj del Val y Mª Mercè Claramunt. 2005. "Bases de datos y estadísticas del seguro de automóviles en España: influencia en el cálculo de primas"

Carlos Bousoño. (Fundación Mapfre)., 2008. "Factores de riesgo y Cálculo de primas mediante técnicas de aprendizaje"

Evelien Brisard., 2014. "Pricing of Car Insurance with Generalized Linear Models"

James Smith., 2004. "Generalized Linear Models and Their Applications to Actuarial Modeling"

David Cummings., 2011. "Practical GLM Modeling of Deductibles"

Universidad de Valencia., 2010. "Estadística descriptiva. Asociación entre variables"

A. I. McLeod, C. Xu (University of Western Ontario)., 2010. "Bestglm: Best Subset GLM"

Caro Carretero, Raquel. (Universidad Pontifica Comillas). "Segmentación y Predicción en los modelos de tarificación"

- Legislación & Webs:

Ley 50/1980, de 8 de octubre, de Contrato de Seguro.

Real Decreto Legislativo 6/2004, de 29 de octubre,: "Texto refundido de la Ley de ordenación y supervisión de los seguros privados".

Real Decreto Legislativo 8/2004, de 29 de octubre: "Texto refundido de la Ley sobre responsabilidad civil y seguro en la circulación de vehículos a motor".

www.ine.es

http://www.dgsfp.mineco.es/

CAPITULO 7. ANEXO

ANEXO 1. Código SAS

```
/*PRIMEROS PASOS*/

OPTION COMPRESS=YES;
LIBNAME TFM '/folders/myfolders/TARIFA ACTUARIAL';
/*Establecer y Determinar Base de Datos. Depuración Base de Datos*/
        /*Colocar, crear y eliminar variables, corregir datos erróneos en variables, Identificadores, etc...*/

Data TFM.bbddjuan;
  set TFM.bbddjuan (keep=MATRICULA COD_CIA COD_PRODUCTO NUM_POLIZA ESTADO
    FOCUR_SINIESTRO CULPA EXPOS_DANOS PRIMA_DANOS N_SIN_DANOS
    INCURRIDO_DANOS maduracion IND_MULTAS LITERAL_USO_VEHICULO
    LITERAL_km_anual LITERAL_PROFESION_CONDUCTOR codigo_postal provincia
    comunidad ZONA_DEHABITABILIDAD MUNICIPIO CIUDAD_DORMITORIO FORMA_PAGO
    LITERAL_MODO_CONTACTO modalidad_pol IMPORTE_FRANQUICIA tipo_vehiculo
    MARCAE MODELE marca_modelo VERSIE MOTORE POTENE CILINE VELMAE PESPOE
    PUERTE PLAZAS LONMME SEGMEE pesoveh valor_vehiculo ant_vehic
    NACIONALIDAD edad_hab ant_car_hab PORC_BONUS_APLICADO SEXO_HABITUAL
    CONDUCTOR_ES_TOMADOR ECIVIL_CONDUCTOR GARAJE CODIGO_SSCC_ANUAL
    TIPO_MOSAIC NOTA_MOROSIDAD DENSIDAD_POBLACION INGRESOS_MEDIOS
    PORC_INGRESOS_LIBRE_DE_GASTOS PORC_GASTO_TRANSPORTE N_MEDIO_VEHICULOS
    PORC_AUTOSNEW_SGTOG7 PORC_AUTOS2MANO_SGTOG7 PORC_MOTOSYCICLOS_SGTOG7
    PORC_MANTYREPARVEH_SGTOG7 PORC_CARBURANTES_SGTOG7 PORC_SEGUROS_SGTOG12
    GRUPOS_MM_RC_DIRECTO GRUPOS_MM_DANOS_DIRECTO sumnufamiliar_autos
    grupo_vehic_dir MAS_VEHICULOS_UFAMILIAR antiguedad_pol_efec
    perm_modalidad_vehiculo_ent TIPO_BONUS_ACREDITADO);
run;
Data TFM.bbddjuan;
  retain Poliza MATRICULA COD_CIA COD_PRODUCTO NUM_POLIZA ESTADO
    modalidad_pol IMPORTE_FRANQUICIA maduracion FORMA_PAGO
    Antiguedad_pol_efec perm_modalidad_vehiculo_ent PORC_BONUS_APLICADO
    TIPO_BONUS_ACREDITADO FOCUR_SINIESTRO CULPA EXPOS_DANOS L_EXPOS PRIMA_DANOS
    Frecuencia N_SIN_DANOS CosteMedio CosteMedio_F INCURRIDO_DANOS BurningCost ELR
    grupo_vehic_dir tipo_vehiculo MARCAE MODELE marca_modelo VERSIE MOTORE
    POTENE CILINE VELMAE PESPOE PUERTE PLAZAS LONMME SEGMEE pesoveh
    valor_vehiculo ant_vehic LITERAL_USO_VEHICULO LITERAL_km_anual GARAJE
    sumnufamiliar_autos MAS_VEHICULOS_UFAMILIAR GRUPOS_MM_RC_DIRECTO
    GRUPOS_MM_DANOS_DIRECTO NACIONALIDAD edad_hab ant_car_hab SEXO_HABITUAL
    CONDUCTOR_ES_TOMADOR ECIVIL_CONDUCTOR LITERAL_PROFESION_CONDUCTOR
    LITERAL_MODO_CONTACTO IND_MULTAS provincia comunidad
    ZONA_DEHABITABILIDAD MUNICIPIO CIUDAD_DORMITORIO codigo_postal
    CODIGO_SSCC_ANUAL TIPO_MOSAIC NOTA_MOROSIDAD N_MEDIO_VEHICULOS
    DENSIDAD_POBLACION INGRESOS_MEDIOS PORC_INGRESOS_LIBRE_DE_GASTOS
    PORC_GASTO_TRANSPORTE PORC_AUTOSNEW_SGTOG7 PORC_AUTOS2MANO_SGTOG7
    PORC_MOTOSYCICLOS_SGTOG7 PORC_MANTYREPARVEH_SGTOG7
    PORC_CARBURANTES_SGTOG7 PORC_SEGUROS_SGTOG12;
  set TFM.bbddjuan;
  if CODIGO_SSCC_ANUAL=' ' then delete;
  if modalidad_pol='a. TERCEROS' then  delete;
  if modalidad_pol='b. TERC + LUNAS' then delete;
  if modalidad_pol='d. TERC + ROBO' then delete;
  if modalidad_pol='g. TERC + LUNAS + ROBO' then delete;
  if modalidad_pol='m. TERC + LUNAS + ROBO+ INCENDIO' then delete;
  if modalidad_pol='p. TRCF' then delete;
  if modalidad_pol='q. TRSF' then delete;
  if PRIMA_DANOS=0 then delete;
  if INCURRIDO_DANOS=0 then INCURRIDO_DANOS='.';
  if PLAZAS='505' or PLAZAS='305' then PLAZAS='.';
```

```
  if LITERAL_km_anual='menos de 5.000 km' then
    LITERAL_km_anual='hasta 5.000 Km';
  if LITERAL_km_anual='de 5.001 km a 10.000 km' then
    LITERAL_km_anual='De 5.001 a 10.000 Km';
  if LITERAL_km_anual='de 10.001 km a 15.000 km' then
    LITERAL_km_anual='De 10.001 a 15.000 Km';
  if LITERAL_km_anual='de 15.001 km a 20.000 km' then
    LITERAL_km_anual='De 15.001 a 20.000 Km';
  if LITERAL_km_anual='de 20.001 km a 25.000 km' then
    LITERAL_km_anual='De 20.001 a 25.000 Km';
  if LITERAL_km_anual='de 25.001 km a 30.000 km' then
    LITERAL_km_anual='De 25.001 a 30.000 Km';
   if LITERAL_km_anual='de 30.001 km a 40.000 km' then
    LITERAL_km_anual='De 30.001 a 40.000 Km';
  if LITERAL_km_anual='más de 50.000 km' then
    LITERAL_km_anual='Desde 40.001';
  if LITERAL_km_anual='de 40.001 km a 50.000 km' then
    LITERAL_km_anual='Desde 40.001';

  CosteMedio=INCURRIDO_DANOS / N_SIN_DANOS;
  BurningCost=INCURRIDO_DANOS / EXPOS_DANOS;
  ELR=INCURRIDO_DANOS / PRIMA_DANOS;
  Frecuencia=N_SIN_DANOS / EXPOS_DANOS;
  CosteMedio_F = CosteMedio + IMPORTE_FRANQUICIA;
  if EXPOS_DANOS < 0.0028 then
    delete;
        if CosteMedio > 15000 then delete;
        L_EXPOS = log (EXPOS_DANOS);
        if Frecuencia > 10.15 then delete; /*Eliminación de Datos Atípicos*/
        if N_SIN_DANOS > 1 then delete;
  format PRIMA_DANOS N_SIN_DANOS INCURRIDO_DANOS BurningCost ELR CosteMedio_F  10.2;
run;

DATA tfm.bbddjuan;
set tfm.bbddjuan;
poliza = _N_;
run;
data TFM.bbddjuan (Drop=i);
  set TFM.bbddjuan;
  array todas (37) BurningCost CILINE CosteMedio DENSIDAD_POBLACION ELR
    EXPOS_DANOS FOCUR_SINIESTRO IMPORTE_FRANQUICIA INCURRIDO_DANOS
    INGRESOS_MEDIOS LONMME NUM_POLIZA N_MEDIO_VEHICULOS N_SIN_DANOS PESPOE
    PLAZAS PORC_AUTOS2MANO_SGTOG7 PORC_AUTOSNEW_SGTOG7 PORC_BONUS_APLICADO
    PORC_CARBURANTES_SGTOG7 PORC_GASTO_TRANSPORTE
    PORC_INGRESOS_LIBRE_DE_GASTOS PORC_MANTYREPARVEH_SGTOG7
    PORC_MOTOSYCICLOS_SGTOG7 PORC_SEGUROS_SGTOG12 POTENE PRIMA_DANOS Poliza
    VELMAE ant_car_hab ant_vehic edad_hab pesoveh sumnufamiliar_autos
    valor_vehiculo antiguedad_pol_efec perm_modalidad_vehiculo_ent;
  do i=1 to 37;
    if todas (i)=999999 then todas (i)=.;
   if todas (i)=999999.00 then todas (i)=.;
    if todas (i)=9999 then todas (i)=.;
    if todas (i)=' ' then todas (i)=.;
   if todas (i)=9999999.00 then todas (i)=.;
    if todas (i)=99999.00 then  todas (i)=.;
    if todas (i)=99999 then todas (i)=.;
    if todas (i)="z.Indeterminado" then todas (i)=.;
    if todas (i)="c. Indetermindado" then todas (i)=.;
  end;run;
data TFM.bbddjuan (Drop=i);
  set TFM.bbddjuan;
  array todas (40) CIUDAD_DORMITORIO CODIGO_SSCC_ANUAL COD_CIA COD_PRODUCTO
```

```
    CONDUCTOR_ES_TOMADOR CULPA ECIVIL_CONDUCTOR ESTADO FORMA_PAGO GARAJE
    GRUPOS_MM_DANOS_DIRECTO GRUPOS_MM_RC_DIRECTO IND_MULTAS
    LITERAL_MODO_CONTACTO LITERAL_PROFESION_CONDUCTOR LITERAL_USO_VEHICULO
    LITERAL_km_anual MARCAE MAS_VEHICULOS_UFAMILIAR MATRICULA MODELE MOTORE
    MUNICIPIO NACIONALIDAD NOTA_MOROSIDAD PUERTE SEGMEE SEXO_HABITUAL
    TIPO_MOSAIC VERSIE ZONA_DEHABITABILIDAD codigo_postal comunidad
    grupo_vehic_dir maduracion marca_modelo modalidad_pol provincia
    tipo_vehiculo TIPO_BONUS_ACREDITADO;
  do i=1 to 40;
    if todas (i)=999999 then todas (i)=.;
   if todas (i)=999999.00 then todas (i)=.;
    if todas (i)=9999 then todas (i)=.;
    if todas (i)=' ' then todas (i)=.;
   if todas (i)=9999999.00 then todas (i)=.;
    if todas (i)=99999.00 then  todas (i)=.;
    if todas (i)=99999 then todas (i)=.;
    if todas (i)="z.Indeterminado" then todas (i)=.;
    if todas (i)="c. Indetermindado" then todas (i)=.;
    if todas (i)="c. Indetermindado" then todas (i)=.;
  end;
run;
proc contents data=TFM.bbddjuan;
run;
proc sql ;
  create table CALCULOS_SQL as Select sum(PRIMA_DANOS) as PRIMA_GANADA,
    sum(INCURRIDO_DANOS) as INCURRIDO, sum(N_SIN_DANOS) as N_SINIESTROS,
    (calculated N_SINIESTROS) /sum(EXPOS_DANOS) as FRECUENCIA, (calculated
    INCURRIDO)/(calculated N_SINIESTROS) as CosteMedio, (calculated
    FRECUENCIA)*(calculated CosteMedio) as BC, (calculated INCURRIDO) /
    (calculated PRIMA_GANADA) as LossRatio, (calculated PRIMA_GANADA) /
    71060 as PrimaMediaGanada from TFM.bbddjuan;
quit;
```

/*SEGMENTACIÓN Y UNIÓN DE TABLAS DE VARIABLES INTERNAS Y EXTERNAS*/

/*Segmentación variables internas*/

```
data tfm.new_seg (keep=poliza edad_hab_char ant_carne POTENCIA CILINDRADA
    VELOCIDAD peso_veh valor_veh antig_veh BONUS_APLICADO LONGITUD
    N_autofamilia PESOPOT AUTOS2MANO AUTOSNEW GASTO_CARBURANTES
    GASTO_TRANSPORTE INGRESOS_LIBRE_DE_GASTOS MANTYREPARVEH MOTOSYCICLOS
    SEGUROS INGRESOSMEDIOS DENSIDADPOBLACION MEDIA_VEHICULOS PLAZAS_TX
                    PUERTE FORMA_PAGO GARAJE ECIVIL_CONDUCTOR NOTA_MOROSIDAD
                    LITERAL_USO_VEHICULO GRUPOS_MM_DANOS_DIRECTO MOTORE LITERAL_km_anual modalidad PUERTAS
Antiguedad);

  set TFM.bbddjuan;

  format edad_hab_char ant_carne POTENCIA CILINDRADA VELOCIDAD peso_veh
    valor_veh antig_veh BONUS_APLICADO LONGITUD N_autofamilia PESOPOT
    AUTOS2MANO AUTOSNEW GASTO_CARBURANTES GASTO_TRANSPORTE
    INGRESOS_LIBRE_DE_GASTOS MANTYREPARVEH MOTOSYCICLOS SEGUROS
    INGRESOSMEDIOS DENSIDADPOBLACION MEDIA_VEHICULOS PLAZAS_TX modalidad PUERTAS Antiguedad $50.;
```

```
        select;
                when (modalidad_pol = 'p. TRCF 90') modalidad = '(90-180)';
                when (modalidad_pol = 'p. TRCF 99') modalidad = '(90-180)';
                when (modalidad_pol = 'p. TRCF 120') modalidad = '(90-180)';
                when (modalidad_pol = 'p. TRCF 125') modalidad = '(90-180)';
                when (modalidad_pol = 'p. TRCF 180') modalidad = '(90-180)';
                when (modalidad_pol = 'p. TRCF 200') modalidad = '(200-600)';
                when (modalidad_pol = 'p. TRCF 300') modalidad = '(200-600)';
                when (modalidad_pol = 'p. TRCF 450') modalidad = '(200-600)';
                when (modalidad_pol = 'p. TRCF 600') modalidad = '(200-600)';
        end;

select;
  When (missing(edad_hab)) edad_hab_char="01. LT 37";
  when (edad_hab < 25) edad_hab_char="01. LT 37";
  when (edad_hab < 31) edad_hab_char="01. LT 37";
  when (edad_hab <= 37) edad_hab_char="01. LT 37";
  when (edad_hab <= 44) edad_hab_char="02. 37-44";
  when (edad_hab <= 47) edad_hab_char="03. 44-47";
  when (edad_hab > 47) edad_hab_char="04. MT 47";
  otherwise edad_hab_char="04. MT 47";
end;

select;
  when (missing(ant_car_hab)) ant_carne="01. LT 28";
  when (ant_car_hab <= 28) ant_carne="01. LT 28";
  when (ant_car_hab > 28) ant_carne="02. MT 28";
  otherwise ant_carne="02. MT 28";
end;

select;
  when (missing(POTENE)) POTENCIA="01. LT 74";
  when (POTENE <= 74) POTENCIA="01. LT 74";
  when (POTENE <= 118) POTENCIA="02. 74-118";
  when (POTENE > 118) POTENCIA="03. MT 118";
  otherwise POTENCIA="03. MT 118";
end;

select;
  when (missing(CILINE)) CILINDRADA="01. LT 1895";
  when (CILINE <= 1895) CILINDRADA="01. LT 1895";
  when (CILINE <= 2350) CILINDRADA="02. 1895-2350";
  when (CILINE > 2350 ) CILINDRADA="03. MT 2350";
  otherwise CILINDRADA="03. MT 2350";
end;

select;
  when (missing(VELMAE)) VELOCIDAD="01. LT 178";
  when (VELMAE <= 177) VELOCIDAD="01. LT 178";
  when (VELMAE <= 205 ) VELOCIDAD="02. 178-205";
  when (VELMAE > 205) VELOCIDAD="04. MT 205";
  otherwise VELOCIDAD="04. MT 205";
end;

select;
  when (missing(PESPOE)) PESOPOT="01. 0-10";
  when (PESPOE <= 10) PESOPOT="01. 0-10";
  when (PESPOE <= 16) PESOPOT="02. 10-16";
  when (PESPOE < 22) PESOPOT="03. + 16";
  when (PESPOE < 28) PESOPOT="03. + 16";
```

```
   when (PESPOE < 34) PESOPOT="03. + 16";
   when (PESPOE < 40) PESOPOT="03. + 16";
   when (PESPOE < 46) PESOPOT="03. + 16";
   when (PESPOE < 52) PESOPOT="03. + 16";
   when (PESPOE < 58) PESOPOT="03. + 16";
   when (PESPOE >=58) PESOPOT="03. + 16";
   otherwise PESOPOT="03. + 16";
end;

select;
   when (missing(LONMME)) LONGITUD="01. LT 4337";
   when (LONMME <= 4337) LONGITUD="01. LT 4337";
   when (LONMME <= 4763) LONGITUD="02. 4337-4763";
   when (LONMME > 4763) LONGITUD="03. MT 4763";
   otherwise LONGITUD="03. MT 4763";
end;

select;
   when (missing(pesoveh)) peso_veh="01. LT 1140";
   when (pesoveh <= 1140) peso_veh="01. LT 1140";
   when (pesoveh <= 1331) peso_veh="02. 1140-1331";
   when (pesoveh > 1331) peso_veh="03. MT 1331";
   otherwise peso_veh="03. MT 1331";
end;

select;
   when (missing(valor_vehiculo)) valor_veh="01. LT 13450";
   when (valor_vehiculo <= 13450) valor_veh="01. LT 13450";
   when (valor_vehiculo <= 22000) valor_veh="02. 13450-22000";
   when (valor_vehiculo <=35800 ) valor_veh="03. 22000-35800";
   when (valor_vehiculo > 35800) valor_veh="04. MT 35800";
   otherwise valor_veh="04. MT 35800";
end;

select;
   when (missing(ant_vehic)) antig_veh="01. LT 2 Yrs";
   when (ant_vehic <= 2) antig_veh="01. LT 2 Yrs";
   when (ant_vehic > 2) antig_veh="02. MT 2 Yrs";
   otherwise antig_veh="02. MT 2";
end;

select;
   when (missing(PORC_BONUS_APLICADO)) BONUS_APLICADO="01. LT 40%";
   when (PORC_BONUS_APLICADO < -65) BONUS_APLICADO="01. LT 40%";
   when (PORC_BONUS_APLICADO < -50) BONUS_APLICADO="01. LT 40%";
   when (PORC_BONUS_APLICADO < -35) BONUS_APLICADO="01. LT 40%";
   when (PORC_BONUS_APLICADO < -20) BONUS_APLICADO="01. LT 40%";
   when (PORC_BONUS_APLICADO < -5) BONUS_APLICADO="01. LT 40%";
   when (PORC_BONUS_APLICADO < 10) BONUS_APLICADO="01. LT 40%";
   when (PORC_BONUS_APLICADO < 25) BONUS_APLICADO="01. LT 40%";
   when (PORC_BONUS_APLICADO < 40) BONUS_APLICADO="01. LT 40%";
   when (PORC_BONUS_APLICADO < 55) BONUS_APLICADO="02. 40-55%";
   when (PORC_BONUS_APLICADO >= 55) BONUS_APLICADO="03. MT 55%";
   otherwise BONUS_APLICADO="03. MT 55%";
end;


select;
   when (missing(sumnufamiliar_autos)) N_autofamilia="01. 0 auto";
   when (sumnufamiliar_autos=0) N_autofamilia="01. 0 auto";
   when (sumnufamiliar_autos > 0) N_autofamilia="02. + DE 1 auto";
   otherwise N_autofamilia="02. + DE 1 auto";
```

```
end;

select;
   when (missing(PORC_AUTOS2MANO_SGTOG7)) AUTOS2MANO="01. 0-8.5%";
   when (PORC_AUTOS2MANO_SGTOG7 < 8.50) AUTOS2MANO="01. 0-8.5%";
   when (PORC_AUTOS2MANO_SGTOG7 < 17.00) AUTOS2MANO="02. 8.5-17%";
   when (PORC_AUTOS2MANO_SGTOG7 < 25.50) AUTOS2MANO="03. + de 17%";
   when (PORC_AUTOS2MANO_SGTOG7 < 34.00) AUTOS2MANO="03. + de 17%";
   when (PORC_AUTOS2MANO_SGTOG7 < 42.50) AUTOS2MANO="03. + de 17%";
   when (PORC_AUTOS2MANO_SGTOG7 < 51.00) AUTOS2MANO="03. + de 17%";
   when (PORC_AUTOS2MANO_SGTOG7 < 59.50) AUTOS2MANO="03. + de 17%";
   when (PORC_AUTOS2MANO_SGTOG7 < 68.00) AUTOS2MANO="03. + de 17%";
   when (PORC_AUTOS2MANO_SGTOG7 < 76.50) AUTOS2MANO="03. + de 17%";
   when (PORC_AUTOS2MANO_SGTOG7 >=76.50) AUTOS2MANO="03. + de 17%";
   otherwise AUTOS2MANO="03. + de 17%";
end;

select;
   when (missing(PORC_AUTOSNEW_SGTOG7)) AUTOSNEW="01. 0-8.5%";
   when (PORC_AUTOSNEW_SGTOG7 < 8.50) AUTOSNEW="01. 0-8.5%";
   when (PORC_AUTOSNEW_SGTOG7  < 17.00) AUTOSNEW="02. 8.5-17%";
   when (PORC_AUTOSNEW_SGTOG7  < 25.50) AUTOSNEW="03. 17-34%";
   when (PORC_AUTOSNEW_SGTOG7  < 34.00) AUTOSNEW="03. 17-34%";
   when (PORC_AUTOSNEW_SGTOG7  < 42.50) AUTOSNEW="05. + 35%";
   when (PORC_AUTOSNEW_SGTOG7  < 51.00) AUTOSNEW="05. + 35%";
   when (PORC_AUTOSNEW_SGTOG7  < 59.50) AUTOSNEW="05. + 35%";
   when (PORC_AUTOSNEW_SGTOG7  < 68.00) AUTOSNEW="05. + 35%";
   when (PORC_AUTOSNEW_SGTOG7  < 76.50) AUTOSNEW="05. + 35%";
   when (PORC_AUTOSNEW_SGTOG7  >=76.50) AUTOSNEW="05. + 35%";
   otherwise AUTOSNEW="05. + 35%";
end;

select;
   when (missing(PORC_CARBURANTES_SGTOG7)) GASTO_CARBURANTES="03. lt 30%";
   when (PORC_CARBURANTES_SGTOG7 < 10.00) GASTO_CARBURANTES="03. lt 30%";
   when (PORC_CARBURANTES_SGTOG7  < 20.00) GASTO_CARBURANTES="03. lt 30%";
   when (PORC_CARBURANTES_SGTOG7  < 30.00) GASTO_CARBURANTES="03. lt 30%";
   when (PORC_CARBURANTES_SGTOG7  < 40.00) GASTO_CARBURANTES="04. 30-40%";
   when (PORC_CARBURANTES_SGTOG7  < 50.00) GASTO_CARBURANTES="05. 40-50%";
   when (PORC_CARBURANTES_SGTOG7  < 60.00) GASTO_CARBURANTES="06. + 50%";
   when (PORC_CARBURANTES_SGTOG7  < 70.00) GASTO_CARBURANTES="06. + 50%";
   when (PORC_CARBURANTES_SGTOG7  < 80.00) GASTO_CARBURANTES="06. + 50%";
   when (PORC_CARBURANTES_SGTOG7  < 90.00) GASTO_CARBURANTES="06. + 50%";
   when (PORC_CARBURANTES_SGTOG7  >=90.00) GASTO_CARBURANTES="06. + 50%";
   otherwise GASTO_CARBURANTES="06. + 50%";
end;

select;
   when (missing(PORC_GASTO_TRANSPORTE)) GASTO_TRANSPORTE="02. lt 13%";
   when (PORC_GASTO_TRANSPORTE  < 9.00) GASTO_TRANSPORTE="02. lt 13%";
   when (PORC_GASTO_TRANSPORTE  < 13.00) GASTO_TRANSPORTE="02. lt 13%";
   when (PORC_GASTO_TRANSPORTE  < 17.00) GASTO_TRANSPORTE="03. 13-17%";
   when (PORC_GASTO_TRANSPORTE  < 21.00) GASTO_TRANSPORTE="04. 17-21%";
   when (PORC_GASTO_TRANSPORTE  < 25.00) GASTO_TRANSPORTE="05. + 21%";
   when (PORC_GASTO_TRANSPORTE  < 29.00) GASTO_TRANSPORTE="05. + 21%";
   when (PORC_GASTO_TRANSPORTE  < 33.00) GASTO_TRANSPORTE="05. + 21%";
   when (PORC_GASTO_TRANSPORTE  < 37.00) GASTO_TRANSPORTE="05. + 21%";
   when (PORC_GASTO_TRANSPORTE  < 41.00) GASTO_TRANSPORTE="05. + 21%";
   when (PORC_GASTO_TRANSPORTE  >=41.00) GASTO_TRANSPORTE="05. + 21%";
   otherwise GASTO_TRANSPORTE="05. + 21%";
end;
```

```
select;
    when (missing(PORC_INGRESOS_LIBRE_DE_GASTOS))
      INGRESOS_LIBRE_DE_GASTOS="05. lt -20%";
    when (PORC_INGRESOS_LIBRE_DE_GASTOS  < -80.00)
      INGRESOS_LIBRE_DE_GASTOS="05. lt -20%";
    when (PORC_INGRESOS_LIBRE_DE_GASTOS  < -65.00)
      INGRESOS_LIBRE_DE_GASTOS="05. lt -20%";
    when (PORC_INGRESOS_LIBRE_DE_GASTOS  < -50.00)
      INGRESOS_LIBRE_DE_GASTOS="05. lt -20%";
    when (PORC_INGRESOS_LIBRE_DE_GASTOS  < -35.00)
      INGRESOS_LIBRE_DE_GASTOS="05. lt -20%";
    when (PORC_INGRESOS_LIBRE_DE_GASTOS  < -20.00)
      INGRESOS_LIBRE_DE_GASTOS="05. lt -20%";
    when (PORC_INGRESOS_LIBRE_DE_GASTOS  < -05.00)
      INGRESOS_LIBRE_DE_GASTOS="06. lt-5%";
    when (PORC_INGRESOS_LIBRE_DE_GASTOS  < 10.00)
      INGRESOS_LIBRE_DE_GASTOS="07. lt 10%";
    when (PORC_INGRESOS_LIBRE_DE_GASTOS  < 25.00)
      INGRESOS_LIBRE_DE_GASTOS="08. gt 10%";
    when (PORC_INGRESOS_LIBRE_DE_GASTOS  < 40.00)
      INGRESOS_LIBRE_DE_GASTOS="08. gt 10%";
    when (PORC_INGRESOS_LIBRE_DE_GASTOS  >=40.00)
      INGRESOS_LIBRE_DE_GASTOS="08. gt 10%";
    otherwise INGRESOS_LIBRE_DE_GASTOS="08. gt 10%";
  end;

  select;
    when (missing(PORC_MANTYREPARVEH_SGTOG7)) MANTYREPARVEH="02. lt 16%";
    when (PORC_MANTYREPARVEH_SGTOG7 < 8.00) MANTYREPARVEH="02. lt 16%";
    when (PORC_MANTYREPARVEH_SGTOG7 < 16.00) MANTYREPARVEH="02. lt 16%";
    when (PORC_MANTYREPARVEH_SGTOG7  < 24.00) MANTYREPARVEH="03. 16-24%";
    when (PORC_MANTYREPARVEH_SGTOG7  < 32.00) MANTYREPARVEH="04. + 25%";
    when (PORC_MANTYREPARVEH_SGTOG7  < 40.00) MANTYREPARVEH="04. + 25%";
    when (PORC_MANTYREPARVEH_SGTOG7  < 48.00) MANTYREPARVEH="04. + 25%";
    when (PORC_MANTYREPARVEH_SGTOG7  < 56.00) MANTYREPARVEH="04. + 25%";
    when (PORC_MANTYREPARVEH_SGTOG7  < 64.00) MANTYREPARVEH="04. + 25%";
    when (PORC_MANTYREPARVEH_SGTOG7  < 72.00) MANTYREPARVEH="04. + 25%";
    when (PORC_MANTYREPARVEH_SGTOG7  >=72.00) MANTYREPARVEH="04. + 25%";
    otherwise MANTYREPARVEH="04. + 25%";
  end;



 select;
    when (missing(PORC_MOTOSYCICLOS_SGTOG7)) MOTOSYCICLOS="02. 6-12%";
    when (PORC_MOTOSYCICLOS_SGTOG7  < 6.00) MOTOSYCICLOS="01. 0-6%";
    when (PORC_MOTOSYCICLOS_SGTOG7  < 12.00) MOTOSYCICLOS="02. 6-12%";
    when (PORC_MOTOSYCICLOS_SGTOG7  < 18.00) MOTOSYCICLOS="03. + 12%";
    when (PORC_MOTOSYCICLOS_SGTOG7  < 24.00) MOTOSYCICLOS="03. + 12%";
    when (PORC_MOTOSYCICLOS_SGTOG7  < 30.00) MOTOSYCICLOS="03. + 12%";
    when (PORC_MOTOSYCICLOS_SGTOG7  < 36.00) MOTOSYCICLOS="03. + 12%";
    when (PORC_MOTOSYCICLOS_SGTOG7  < 42.00) MOTOSYCICLOS="03. + 12%";
    when (PORC_MOTOSYCICLOS_SGTOG7  < 48.00) MOTOSYCICLOS="03. + 12%";
    when (PORC_MOTOSYCICLOS_SGTOG7  < 54.00) MOTOSYCICLOS="03. + 12%";
    when (PORC_MOTOSYCICLOS_SGTOG7  >=54.00) MOTOSYCICLOS="03. + 12%";
    otherwise MOTOSYCICLOS="03. + 12%";
  end;

  select;
    when (missing(PORC_SEGUROS_SGTOG12)) SEGUROS="04. lt 35%";
    when (PORC_SEGUROS_SGTOG12  < 14.00) SEGUROS="04. lt 35%";
    when (PORC_SEGUROS_SGTOG12 < 21.50) SEGUROS="04. lt 35%";
```

```
    when (PORC_SEGUROS_SGTOG12 < 29.00) SEGUROS="04. lt 35%";
    when (PORC_SEGUROS_SGTOG12 < 36.50) SEGUROS="04. lt 35%";
    when (PORC_SEGUROS_SGTOG12 < 44.00) SEGUROS="05. 36-44%";
    when (PORC_SEGUROS_SGTOG12 < 51.50) SEGUROS="06. 44-59%";
    when (PORC_SEGUROS_SGTOG12 < 59.00) SEGUROS="06. 44-59%";
    when (PORC_SEGUROS_SGTOG12 < 66.50) SEGUROS="08. + 59%";
    when (PORC_SEGUROS_SGTOG12 < 74.00) SEGUROS="08. + 59%";
    when (PORC_SEGUROS_SGTOG12 >=74.00) SEGUROS="08. + 59%";
    otherwise SEGUROS="08. + 59%";
  end;

  select;
    when (missing(INGRESOS_MEDIOS)) INGRESOSMEDIOS="02. lt 1800";
    when (INGRESOS_MEDIOS  < 1300) INGRESOSMEDIOS="02. lt 1800";
    when (INGRESOS_MEDIOS < 1800) INGRESOSMEDIOS="02. lt 1800";
    when (INGRESOS_MEDIOS < 2300) INGRESOSMEDIOS="03. 1800-2300";
    when (INGRESOS_MEDIOS < 2800) INGRESOSMEDIOS="04. 2300-2800";
    when (INGRESOS_MEDIOS < 3300) INGRESOSMEDIOS="05. + 2800";
    when (INGRESOS_MEDIOS < 3800) INGRESOSMEDIOS="05. + 2800";
    when (INGRESOS_MEDIOS < 4300) INGRESOSMEDIOS="05. + 2800";
    when (INGRESOS_MEDIOS < 4800) INGRESOSMEDIOS="05. + 2800";
    when (INGRESOS_MEDIOS < 5300) INGRESOSMEDIOS="05. + 2800";
    when (INGRESOS_MEDIOS >=5300) INGRESOSMEDIOS="05. + 2800";
    otherwise INGRESOSMEDIOS="05. + 2800";
  end;

  select;
    when (missing(DENSIDAD_POBLACION)) DENSIDADPOBLACION="01. 0-17600";
    when (DENSIDAD_POBLACION < 17600) DENSIDADPOBLACION="01. 0-17600";
    when (DENSIDAD_POBLACION < 35200) DENSIDADPOBLACION="02. 17600-35200";
    when (DENSIDAD_POBLACION < 52800) DENSIDADPOBLACION="03. 35200-52800";
    when (DENSIDAD_POBLACION < 70400) DENSIDADPOBLACION="04. + 52800";
    when (DENSIDAD_POBLACION < 88000) DENSIDADPOBLACION="04. + 52800";
    when (DENSIDAD_POBLACION < 105600) DENSIDADPOBLACION="04. + 52800";
    when (DENSIDAD_POBLACION < 123200) DENSIDADPOBLACION="04. + 52800";
    when (DENSIDAD_POBLACION < 140800)
      DENSIDADPOBLACION="04. + 52800";
    when (DENSIDAD_POBLACION < 158400)
      DENSIDADPOBLACION="04. + 52800";
    when (DENSIDAD_POBLACION >=158400) DENSIDADPOBLACION="04. + 52800";
    otherwise DENSIDADPOBLACION="04. + 52800";
  end;


select;
    when (missing(N_MEDIO_VEHICULOS)) MEDIA_VEHICULOS="02. 0 - 1";
    when (N_MEDIO_VEHICULOS  < 0.5) MEDIA_VEHICULOS="02. 0 - 1";
    when (N_MEDIO_VEHICULOS  < 1) MEDIA_VEHICULOS="02. 0 - 1";
    when (N_MEDIO_VEHICULOS  < 1.5) MEDIA_VEHICULOS="03. 1-1.5";
    when (N_MEDIO_VEHICULOS < 2) MEDIA_VEHICULOS="04. + 1.5";
    when (N_MEDIO_VEHICULOS < 2.5) MEDIA_VEHICULOS="04. + 1.5";
    otherwise MEDIA_VEHICULOS="04. + 1.5";
  end;

 select;
    when (missing(PLAZAS)) PLAZAS_TX="01. LT 5";
    when (PLAZAS  <= 5) PLAZAS_TX="01. LT 5";
                        when (PLAZAS  > 5) PLAZAS_TX="02. MT 5";
    otherwise PLAZAS_TX="02. MT 5";
  end;
```

```
select;
      when (missing(PUERTE)) PUERTAS="01. LT 4";
      when (PUERTE   <= 4) PUERTAS="01. LT 4";
      when (PUERTE > 4) PUERTAS="02. MT 4";
      otherwise PUERTAS ="02. MT 4";
            end;

  select;
      when (missing(antiguedad_pol_efec )) Antiguedad="NB";
      when (antiguedad_pol_efec  = 0) Antiguedad="NB";
      when (antiguedad_pol_efec  <= 5) Antiguedad="1-5 años";
      when (antiguedad_pol_efec  <= 10) Antiguedad="6-10 años";
      when (antiguedad_pol_efec  > 10) Antiguedad="MT 10 años";
      otherwise Antiguedad="MT 10 años";
            end;

if FORMA_PAGO = 'f. MENSUAL' then FORMA_PAGO = 'd. TRIMESTRAL';
if GARAJE = 'P' then GARAJE = 'N';

if ECIVIL_CONDUCTOR = 'D' then ECIVIL_CONDUCTOR = 'S';
if ECIVIL_CONDUCTOR = 'V' then ECIVIL_CONDUCTOR = 'S';
if ECIVIL_CONDUCTOR = 'X' then ECIVIL_CONDUCTOR = 'S';
if ECIVIL_CONDUCTOR = 'O' then ECIVIL_CONDUCTOR = 'S';
if ECIVIL_CONDUCTOR = 'E' then ECIVIL_CONDUCTOR = 'C';

if GRUPOS_MM_DANOS_DIRECTO = 'GRUPO DANOS 0' then GRUPOS_MM_DANOS_DIRECTO = 'GRUPO DANOS 1';
if GRUPOS_MM_DANOS_DIRECTO = 'GRUPO DANOS 6' then GRUPOS_MM_DANOS_DIRECTO = 'GRUPO DANOS 5';

if MOTORE = '.' then MOTORE = 'G';
if MOTORE = 'B' then MOTORE = 'G';
if MOTORE = 'E' then MOTORE = 'G';
if MOTORE = 'L' then MOTORE = 'G';
if MOTORE = 'X' then MOTORE = 'G';
if MOTORE = 'Y' then MOTORE = 'G';

if LITERAL_km_anual='hasta 5.000 Km' then LITERAL_km_anual= 'Menor a 10.000 km';
if LITERAL_km_anual='De 5.001 a 10.000 Km' then LITERAL_km_anual= 'Menor a 10.000 km';
if LITERAL_km_anual='De 10.001 a 15.000 Km' then LITERAL_km_anual= '10.000 - 20.000 km';
if LITERAL_km_anual='De 15.001 a 20.000 Km' then LITERAL_km_anual= '10.000 - 20.000 km';
if LITERAL_km_anual='De 20.001 a 25.000 Km' then LITERAL_km_anual= 'Mas de 20.000 km';
if LITERAL_km_anual='De 25.001 a 30.000 Km' then LITERAL_km_anual= 'Mas de 20.000 km';
if LITERAL_km_anual='De 30.001 a 40.000 Km' then LITERAL_km_anual= 'Mas de 20.000 km';
if LITERAL_km_anual='Desde 40.001' then LITERAL_km_anual= 'Mas de 20.000 km';
if LITERAL_km_anual='más de 50.000 km' then LITERAL_km_anual= 'Mas de 20.000 km';

if LITERAL_USO_VEHICULO='HERRAMIENTA DE TRABAJO' then LITERAL_USO_VEHICULO='Trabajo y Uso Profesional';
if LITERAL_USO_VEHICULO='PARA IR AL TRABAJO. SIN GESTIO' then LITERAL_USO_VEHICULO='Trabajo y Uso Profesional';
if LITERAL_USO_VEHICULO='PARTICULAR CON DESPLAZ.TRABAJO' then LITERAL_USO_VEHICULO='Trabajo y Uso Profesional';
if LITERAL_USO_VEHICULO='PARTICULAR PARA USO PROFESIONA' then LITERAL_USO_VEHICULO='Trabajo y Uso Profesional';
if LITERAL_USO_VEHICULO='PARTICULAR' then LITERAL_USO_VEHICULO='Particular';
if LITERAL_USO_VEHICULO='LLEVAR A LOS NIÑOS AL COLEGIO' then LITERAL_USO_VEHICULO='Particular';
if LITERAL_USO_VEHICULO='RENTING PARTICULARES' then LITERAL_USO_VEHICULO='Particular';
if LITERAL_USO_VEHICULO='PARTICULAR SIN DESPL.TRABAJO' then LITERAL_USO_VEHICULO='Particular';

if NOTA_MOROSIDAD = '.' then NOTA_MOROSIDAD = 'C';
if NOTA_MOROSIDAD = 'F' then NOTA_MOROSIDAD = 'E';
if NOTA_MOROSIDAD = 'G' then NOTA_MOROSIDAD = 'E';
if NOTA_MOROSIDAD = 'H' then NOTA_MOROSIDAD = 'E';
if NOTA_MOROSIDAD = 'I' then NOTA_MOROSIDAD = 'E';


run;
```

```
/*Unión de Tablas*/
/* Union Tablas bbddjuan - tabla variables externas*/

data TFM.bbdd_grp_maestra_externa_sscc (rename=(CODIGO=CODIGO_SSCC_ANUAL));
   set TFM.bbdd_grp_maestra_externa_sscc;
run;
PROC SORT DATA=TFM.bbddjuan;
   By CODIGO_SSCC_ANUAL;
run;
Proc Sort data=TFM.bbdd_grp_maestra_externa_sscc;
   By CODIGO_SSCC_ANUAL;
run;

Data TFM.Basededatos (drop= ECIVIL_CONDUCTOR GRUPOS_MM_DANOS_DIRECTO LITERAL_USO_VEHICULO
LITERAL_km_anual  MOTORE PUERTE FORMA_PAGO GARAJE);
   MERGE TFM.bbddjuan(in=a) TFM.bbdd_grp_maestra_externa_sscc (in=b);
   by CODIGO_SSCC_ANUAL;
   if a;
run;
proc sort data=TFM.Basededatos;
   by poliza;
run;
proc sort data=tfm.new_seg;
   by poliza;
run;
Data tfm.basefinal;
   MERGE TFM.Basededatos tfm.new_seg;
   by poliza;
run;

data tfm.basefinal (Drop=ant_car_hab edad_hab ant_vehic valor_vehiculo pesoveh
      LONMME PESPOE VELMAE CILINE POTENE PORC_BONUS_APLICADO
      sumnufamiliar_autos NUM_POLIZA COD_PRODUCTO COD_CIA MATRICULA COD_DC
      COD_NUM COD_PROV COD_CCAA COD_MUN PORC_AUTOS2MANO_SGTOG7
      PORC_AUTOSNEW_SGTOG7 PORC_CARBURANTES_SGTOG7 PORC_GASTO_TRANSPORTE
      PORC_INGRESOS_LIBRE_DE_GASTOS PLAZAS PORC_MANTYREPARVEH_SGTOG7
      PORC_MOTOSYCICLOS_SGTOG7 PORC_SEGUROS_SGTOG12 N_MEDIO_VEHICULOS
      INGRESOS_MEDIOS DENSIDAD_POBLACION antiguedad_pol_efec );
   retain poliza ESTADO maduracion Antiguedad modalidad_pol modalidad IMPORTE_FRANQUICIA FORMA_PAGO
      FOCUR_SINIESTRO CULPA EXPOS_DANOS L_EXPOS PRIMA_DANOS Frecuencia N_SIN_DANOS
      CosteMedio CosteMedio_F INCURRIDO_DANOS BurningCost ELR BONUS_APLICADO
      grupo_vehic_dir tipo_vehiculo MARCAE MODELE marca_modelo VERSIE
      GRUPOS_MM_RC_DIRECTO GRUPOS_MM_DANOS_DIRECTO MOTORE PUERTE PUERTAS PLAZAS_TX
      PESOPOT POTENCIA CILINDRADA VELOCIDAD peso_veh valor_veh antig_veh
      LONGITUD SEGMEE LITERAL_USO_VEHICULO LITERAL_km_anual GARAJE
      N_autofamilia MAS_VEHICULOS_UFAMILIAR NACIONALIDAD SEXO_HABITUAL
      edad_hab_char ant_carne CONDUCTOR_ES_TOMADOR ECIVIL_CONDUCTOR
      LITERAL_PROFESION_CONDUCTOR LITERAL_MODO_CONTACTO IND_MULTAS provincia
      comunidad ZONA_DEHABITABILIDAD MUNICIPIO CIUDAD_DORMITORIO
      codigo_postal CODIGO_SSCC_ANUAL X Y TIPO_MOSAIC NOTA_MOROSIDAD
      AUTOS2MANO AUTOSNEW GASTO_CARBURANTES GASTO_TRANSPORTE
      INGRESOS_LIBRE_DE_GASTOS MANTYREPARVEH MOTOSYCICLOS SEGUROS
      INGRESOSMEDIOS DENSIDADPOBLACION MEDIA_VEHICULOS;
   set tfm.basefinal;
```

```
run;

/*Análisis Exploratorio*/

/*Análisis Univariable*/

proc means data=TFM.bbddjuan n nmiss min max lclm uclm kurtosis skewness std
      sum mean mode median cv maxdec=2;
   var N_SIN_DANOS Frecuencia CosteMedio INCURRIDO_DANOS BurningCost
      EXPOS_DANOS CosteMedio_F;
run;

proc univariate data=TFM.bbddjuan;
   var Frecuencia CosteMedio INCURRIDO_DANOS BurningCost EXPOS_DANOS CosteMedio_F;
   histogram/kernel;
   probplot;
   ppplot;
run;
proc sgplot data=TFM.bbddjuan;
   hbox CosteMedio / datalabel=Poliza;
run;

/*DETERMINAR LA DISTRIBUCION QUE SIGUE CADA UNA DE ESTAS VARIABLES.*/
/*N_SINIESTROS*/
proc genmod data=TFM.bbddjuan;
   model N_SIN_DANOS= / dist=poisson link=log type3 OFFSET=L_EXPOS ;
run;

proc genmod data=TFM.bbddjuan order=data;
   model N_SIN_DANOS= / dist=negbin link=log type3 OFFSET=L_EXPOS ;
run;

proc genmod data=TFM.bbddjuan order=data;
   model N_SIN_DANOS= / dist=ZIP link=log type3 OFFSET=L_EXPOS ;
   zeromodel/ link=logit;
run;

proc genmod data=TFM.bbddjuan order=data;
   model N_SIN_DANOS= / dist=ZINB link=log type3 OFFSET=L_EXPOS ;
   zeromodel/ link=logit;
run;
/*COSTEMEDIO*/
proc genmod data=TFM.bbddjuan order=data;
   model CosteMedio= / dist=normal link=log;
run;

proc genmod data=TFM.bbddjuan order=data;
   model CosteMedio= / dist=gamma link=log;
run;

proc genmod data=TFM.bbddjuan order=data;
   model CosteMedio= / dist=igaussian link=power(-2);
run;
proc genmod data=TFM.bbddjuan order=data;
   model CosteMedio = / dist=igaussian link=log;
run;
```

```
proc freq data=tfm.new_seg;
   tables edad_hab_char ant_carne POTENCIA CILINDRADA VELOCIDAD peso_veh
      valor_veh antig_veh BONUS_APLICADO LONGITUD N_autofamilia PESOPOT
      PLAZAS_TX AUTOS2MANO AUTOSNEW GASTO_CARBURANTES GASTO_TRANSPORTE
      INGRESOS_LIBRE_DE_GASTOS MANTYREPARVEH MOTOSYCICLOS SEGUROS NOTA_MOROSIDAD
      INGRESOSMEDIOS DENSIDADPOBLACION MEDIA_VEHICULOS PUERTE FORMA_PAGO GARAJE
                        ECIVIL_CONDUCTOR GRUPOS_MM_DANOS_DIRECTO MOTORE LITERAL_km_anual
LITERAL_USO_VEHICULO / scores=table
      plots(only)=freq;
run;


/*Análisis Bivariable*/

%MACRO BIVARIATE (VAR=);

/* Frecvuencia vs. Explicativas */

proc means data = tfm.basefinal_2 noprint;
     class &VAR;
     output out= OUT_STAT /*(drop= _type_ _freq_)*/
       sum(N_SIN_DANOS) = N_SIN_DANOS
         sum(EXPOS_DANOS) = EXPOS_DANOS;
run;

data OUT_STAT_END;
     set OUT_STAT;
  FRECUENCIA = N_SIN_DANOS / EXPOS_DANOS;
     label FRECUENCIA = "Frecuencia"
           EXPOS_DANOS = "Exposicion"
     format FRECUENCIA percent7.4;
run;


proc sgplot data=OUT_STAT_END noautolegend;
     vbarparm  category=&VAR response=EXPOS_DANOS/
               fillattrs=graphdata1
               transparency=0.5
               barwidth=1
              datalabel = _freq_;

     series x=&VAR y=FRECUENCIA /   y2axis markers
         lineattrs=(thickness=2px color = black)
         markerattrs=(color=black size=8px symbol=  circle) datalabel = FRECUENCIA ;

     yaxis OFFSETMAX=0.60 label="Exposicion";
     y2axis  OFFSETMIN= 0.05  OFFSETMAX=0.15 grid
     label="Frecuencia(%)";
     xaxis grid;
               refline 0 / axis=y2;
     title color=pav  "Garantia de Danos";
     title2 color=pav "Frecuencia";
     keylegend / location =inside position = top across=1;
     label EXPOS_DANOS  = "Exposicion"
           FRECUENCIA = "Frecuencia";
run;
```

```
/* Coste Medio vs. Explicativas */

proc means data = tfm.basefinal_2 noprint;
     class &VAR;
     output out= OUT_STAT /*(drop= _type_ _freq_ )*/
       sum(INCURRIDO_DANOS) = INCURRIDO
         sum(N_SIN_DANOS)= N_SIN_DANOS
            sum(EXPOS_DANOS) = EXPOS_DANOS;
run;

data OUT_STAT_END;
     set OUT_STAT;
  COSTE_MEDIO = INCURRIDO / N_SIN_DANOS;
       label COSTE_MEDIO = "Coste Medio"
            EXPOS_DANOS = "Exposicion"
     format COSTE_MEDIO 8.0;
run;

proc sgplot data=OUT_STAT_END noautolegend;
     vbarparm  category=&VAR response = EXPOS_DANOS /
               fillattrs=graphdata1
               transparency=0.5
               barwidth=1
              datalabel = _freq_;

     series x=&VAR y=COSTE_MEDIO /   y2axis markers
          lineattrs=(thickness=2px  color = red)
          markerattrs=(color=black size=8px symbol=  circle) datalabel = COSTE_MEDIO ;

     yaxis OFFSETMAX=0.60 label="Exposicion";
     y2axis  OFFSETMIN= 0.05  OFFSETMAX=0.15 grid
     label="Coste Medio";
     xaxis grid;
     title color=pav  "Garantia de Danos";
     title2 color=pav "Coste Medio";
     keylegend / location =inside position = top across=1;
     label EXPOS_DANOS = "Exposicion"
           COSTE_MEDIO = "Coste Medio";
run;
/* Burning Cost vs. Explicativas */
proc means data = tfm.basefinal_2 noprint;
     class &VAR;
     output out= OUT_STAT /*(drop= _type_ _freq_)*/
       sum(INCURRIDO_DANOS) = INCURRIDO
         sum(EXPOS_DANOS) = EXPOS_DANOS;
run;

data OUT_STAT_END;
     set OUT_STAT;
  BURNING_COST = INCURRIDO  / EXPOS_DANOS;
     label BURNING_COST = "Burning Cost"
              EXPOS_DANOS = "Exposicion"

     format BURNING_COST 8.0;
run;
```

```
proc sgplot data=OUT_STAT_END noautolegend;
    vbarparm  category=&VAR response=EXPOS_DANOS/
              fillattrs=graphdata1
              transparency=0.5
              barwidth=1
             datalabel = _freq_;

    series x=&VAR y=BURNING_COST /   y2axis markers
        lineattrs=(thickness=2px  color = GREEN)
        markerattrs=(color=black size=8px symbol=  circle) datalabel = BURNING_COST ;

    yaxis OFFSETMAX=0.60 label="Exposicion";
    y2axis  OFFSETMIN= 0.05  OFFSETMAX=0.15
    label="Burning Cost";
    xaxis grid;
    title color=pav  "Garantia de Danos";
    title2 color=pav "Burning Cost ";
    keylegend / location =inside position = top across=1;
    label EXPOS_DANOS  = "Exposicion"
          BURNING_COST = "Burning Cost";
run;
%mend BIVARIATE;

%BIVARIATE(VAR=);
```

/*V- DE CRAMER*/

```
/*Analisis de asociacion entre factores de riesgo: V de Cramer*/

%MACRO CRAMER (var1= , var2= );
proc freq data=tfm.basefinal noprint;
            tables  &var1*&var2/ chisq;
            output out=tabla_1 cramv;
run;
proc transpose data=tabla_1
            out=tabla_2
            prefix=column
            name=source
            label=label;
            var _cramv_;
run;
data aux;
  length var1 var2 $ 30.;
  var1 = "&var1"    ;
  var2 = "&var2"    ;
  source = "_CRAMV_";
run;
data fin (drop= label source );
            retain source var1 var2 column1;
            merge aux tabla_2;
            by source;
            rename column1 = Cramer_V;
run;
```

```
proc datasets  lib=work NODETAILS;
 delete tabla_1 tabla_2 aux;
run;

proc append base=final data=fin force getsort;
run;

%MEND CRAMER;

%CRAMER(var1=, var2=) ;

/*STEPWISE: SELECCIÓN DE VARIABLES*/

/*POISSON*/
%MACRO MODEL_Freq (VAR=);/*plots=all*/

ods trace on;
ods output ModelFit (persist=proc) = Estadisticos_F;

proc genmod data=TFM.basefinal_2;
	class &VAR / ORDER=FREQ REF=FIRST;
  model N_SIN_DANOS= &VAR / dist=poisson link=log offset=L_EXPOS type3;
run;

ods _all_ close;
data Estadisticos_F2 (keep=CRITERION VALUE);
	set Estadisticos_F;
run;

PROC TRANSPOSE DATA=Estadisticos_F2
	OUT=WORK.TRNST_F2
	PREFIX=Column
	NAME=Source
	LABEL=Label
;
	ID Criterion;
	VAR Value;
RUN;

data aux;
  length var1  $ 30.;
  var1 = "&var"      ;
  source = "Value";
run;

data fin;
	merge aux WORK.TRNST_F2;
	by source;
run;

proc append base=final data=fin force;
run;

%MEND MODEL_Freq;
```

```
%MODEL_Freq(VAR=);

/*GAMMA*/

%MACRO MODEL_CMe (VAR=);

ods trace on;
ods output ModelFit (persist=proc) = Estadisticos_CMe;

proc genmod data=tfm.basefinal_2; /*plots=all*/
class &VAR / ORDER=FREQ REF=FIRST;
model CosteMedio_F = &VAR / dist=gamma link=log type3;
run;
ods _all_ close;

data Estadisticos_CMe2 (keep=CRITERION VALUE);
	set Estadisticos_CMe;
run;

PROC TRANSPOSE DATA=Estadisticos_CMe2
	OUT=WORK.TRNST_CMe2
	PREFIX=Column
	NAME=Source
	LABEL=Label
;
	ID Criterion;
	VAR Value;
RUN;
data aux;
  length var1  $ 30.;
  var1 = "&var"      ;
  source = "Value";
run;

data fin;
	merge aux WORK.TRNST_CMe2;
	by source;
run;

proc append base=final data=fin force;
run;


%MEND;

%MODEL_CMe(VAR=);
```

```
/*MODELOS GLM Y VALIDACIÓN CRUZADA*/
OPTION COMPRESS=YES;
LIBNAME TFM '/folders/myfolders/TARIFA ACTUARIAL';

proc genmod data=TFM.basefinal_2 order=data plots=all;
        CLASS  modalidad_pol valor_veh IND_MULTAS N_autofamilia Antiguedad GRUPOS_MM_DANOS_DIRECTO
        ZONA_DEHABITABILIDAD BONUS_APLICADO
         TER_PESO_1 TASA_PARO_1 PREC_APR_DIAS_1 LOC_ACTIV_PRC_1 VIENTO_VEL_MED_1 HELADA_DIAS_1
        /order=freq ref=first;
        model N_SIN_DANOS = modalidad_pol valor_veh IND_MULTAS N_autofamilia Antiguedad
GRUPOS_MM_DANOS_DIRECTO
        ZONA_DEHABITABILIDAD BONUS_APLICADO
         TER_PESO_1 TASA_PARO_1 PREC_APR_DIAS_1 LOC_ACTIV_PRC_1 VIENTO_VEL_MED_1 HELADA_DIAS_1
        / dist=poisson link=log type3 offset=L_EXPOS;
        output out=DATOS_F p= Freq_Estimada RESDEV=yresid cooksd=Cook leverage=leverage Betas= Xbeta;
run;
Proc sort data=DATOS_F;
by poliza;
run;
Proc sort data=DATOS_F_sf;
by poliza;
run;

DATA DatosFrecuencias (Keep= poliza EXPOS_DANOS L_EXPOS PRIMA_DANOS
     Frecuencia N_SIN_DANOS CosteMedio INCURRIDO_DANOS BurningCost ELR Freq_Estimada_SinF
                Freq_Estimada Cook leverage Betas yresid) ;
merge DATOS_F_sf DATOS_F;
By poliza;
run;

*=============================================================================;
/* VALIDACIÓN CRUZADA (5-FOLDS) */

* Creación de un número aleatorio de 5 niveles  ;

data A (keep= obs NUM);
call streaminit(123); /* set random number seed */
do obs = 1 to 71057;
  u = rand("Uniform"); /* u ~ U[0,1] */
          a = 1; b = 6;  /* example values */
  NUM = int(a + (b-a)*u);
  output;
end;
run;

* creamos un contador en la tabla oridingal;

data TRANSFER_CV;
        set TFM.basefinal_2;
        obs = _N_;
run;
```

```
* unimos las dos tablas anteriores;

data TRANSFER_CV_ALL;
            merge A TRANSFER_CV;
            by obs;
run;

* creamos 5 conjuntos de datos;

data CV_1 CV_2 CV_3 CV_4 CV_5;
            set TRANSFER_CV_ALL;
                        if NUM = 1 then output CV_1;
            else if NUM = 2 then output CV_2;
            else if NUM = 3 then output CV_3;
            else if NUM = 4 then output CV_4;
            else if NUM = 5 then output CV_5;
run;

data FOLD_1; /*5 FUERA*/
            set CV_1 CV_2 CV_3 CV_4;
run;

data FOLD_2;/*4 FUERA*/
            set CV_1 CV_2 CV_3 CV_5;
run;

data FOLD_3;/*3 FUERA*/
            set CV_1 CV_2 CV_4 CV_5;
run;

data FOLD_4;/*2 FUERA*/
            set CV_1 CV_3 CV_4 CV_5;
run;

data FOLD_5;/*1 FUERA*/
            set CV_2 CV_3 CV_4 CV_5;
run;

%macro CV (train=,val=);

data TRAINING_VALIDATION;
            set &train(in=__ORIG) &val;
            __FLAG=__ORIG;
            __DEP=N_SIN_DANOS;
            if not __FLAG then N_SIN_DANOS= . ;
run;
```

```
**************************************************************************
Puntuamos las observaciones de entrenaiento y de validaciOn. La
variable que recoje la puntuaciOn o prediccIOn es predicted_N_SIN_DANOS
**************************************************************************;

proc genmod data= TRAINING_VALIDATION;
        CLASS  modalidad_pol valor_veh IND_MULTAS N_autofamilia Antiguedad GRUPOS_MM_DANOS_DIRECTO
        ZONA_DEHABITABILIDAD BONUS_APLICADO
         TER_PESO_1 TASA_PARO_1 HELADA_DIAS_1 PREC_APR_DIAS_1 LOC_ACTIV_PRC_1 VIENTO_VEL_MED_1
        /order=freq ref=first;
        model N_SIN_DANOS = modalidad_pol valor_veh IND_MULTAS N_autofamilia Antiguedad
        GRUPOS_MM_DANOS_DIRECTO ZONA_DEHABITABILIDAD BONUS_APLICADO
         TER_PESO_1 TASA_PARO_1 HELADA_DIAS_1 PREC_APR_DIAS_1 LOC_ACTIV_PRC_1 VIENTO_VEL_MED_1
        /dist=poisson link=log offset=L_EXPOS type3;
        output out=PRED_&val p= Freq_Estimada RESDEV=yresid;
run;

data PRED_&val;
        set PRED_&val;
        N_SIN_DANOS = __DEP;
        drop __DEP;
run;

data PRED_&val (keep=Poliza N_SIN_DANOS Freq_Estimada PRIMA_DANOS EXPOS_DANOS);
         set PRED_&val;
         if __FLAG = 0 then output;
run;

%MEND CV;

%CV ( train = FOLD_1, val = CV_5);
%CV ( train = FOLD_2, val = CV_4);
%CV ( train = FOLD_3, val = CV_3);
%CV ( train = FOLD_4, val = CV_2);
%CV ( train = FOLD_5, val = CV_1);

data PRED_ALL;
 set PRED_CV_1 PRED_CV_2 PRED_CV_3 PRED_CV_4 PRED_CV_5;
run;

/* GRAFICO DE LIFT */

/*GRÁFICO LIFT FRECUENCIA*/

proc sort data= PRED_ALL out=Frquency;
        by Freq_Estimada;
run;

proc freq data=Frquency
        order=internal
        noprint
;
        tables Freq_Estimada /         out=work.auxiliar
        scores=table;
run;
```

```
data Frquency;
            set Frquency;
            conteo = _N_;
            porcentaje = conteo / 71057;
run;

data Frquency;
            set Frquency;

    if        porcentaje < = 0.1 then do; clase = 1 ; rating = 1;  end;
else if   0.1  < porcentaje < = 0.2 then do; clase = 2 ; rating = 2;  end;
else if   0.2  < porcentaje < = 0.3 then do; clase = 3 ; rating = 3;  end;
else if   0.3  < porcentaje < = 0.4 then do; clase = 4 ; rating = 4;  end;
else if   0.4  < porcentaje < = 0.5 then do; clase = 5 ; rating = 5;  end;
else if   0.5  < porcentaje < = 0.6 then do; clase = 6 ; rating = 6;  end;
else if   0.6  < porcentaje < = 0.7 then do; clase = 7 ; rating = 7;  end;
else if   0.7  < porcentaje < = 0.8 then do; clase = 8 ; rating = 8;  end;
else if   0.8  < porcentaje < = 0.9 then do; clase = 9 ; rating = 9;  end;
else if   0.9  < porcentaje        then do; clase = 10; rating = 10; end;

run;
proc means data=Frquency noprint;
                        class clase;
                        output out= Classifier_F
          sum(N_SIN_DANOS) = Siniestros
                                    sum(EXPOS_DANOS) = Exposicion
                                    mean(Freq_Estimada) = Frec_Estimada;
            run;
data Classifier_F;
  set Classifier_F;
  Frecuencia = Siniestros / 71057;
  Freq = Siniestros / Exposicion;
  label Frecuencia = "Frecuencia daոos"

run;
**********************************************************************
GRAFICO SCORING POR CLASES - ENTRENAMIENTO/VALIDACION
**********************************************************************;
proc sgplot data=Classifier_F  noautolegend;
            vbarparm category=clase response=_freq_;
            series x=clase y=Frec_Estimada/   y2axis lineattrs=(thickness=.5 pattern=solid color = red);
            series x=clase y=Freq /   y2axis lineattrs=(thickness=.5 pattern=solid color = green);
            scatter x=clase y=Frec_Estimada/ y2axis  /*yerrorlower=lowerwaldcl yerrorupper=upperwaldcl y2axis
                        errorbarattrs=(color=cx445694) */  markerattrs=(color=red size=7px symbol= circle);
            scatter x=clase y=Freq/ y2axis  /*yerrorlower=lowerwaldcl yerrorupper=upperwaldcl y2axis
                        errorbarattrs=(color=cx445694) */  markerattrs=(color=green/*cx445694*/ size=7px symbol=
circle);
            yaxis  values=(0 to 10000 by 100) label="Observaciones" offsetmax=0.70;
            y2axis grid values=(0 to 1.3 by 0.05) label="Porcentaje" offsetmin=0.15;
            xaxis grid;
run;
```

SE SIGUE EL MISMO PROCESO Y CÓDIGO PARA CADA UNO DE LOS MODELOS DE COSTE MEDIO, PRIMA PURA, Y CUALQUIER MODELO QUE SE QUIERA IMPLANTAR CON LAS VARIABLES SELECCIONADAS