UNIVERSIDAD CARLOS III DE MADRID

MÁSTER EN CIENCIAS ACTUARIALES Y FINANCIERAS

TARIFICACIÓN EN ESPACIOS DE ALTA DIMENSIONALIDAD A TRAVÉS DEL APRENDIZAJE AUTOMÁTICO

TRABAJO FIN DE MÁSTER

ALUMNO:

ALEJANDRO CACHÁN BLANCO

TUTORES:

D. JOSÉ MIGUEL RODRÍGUEZ-PARDO DEL CASTILLO

D. JESÚS RAMÓN SIMÓN DEL POTRO

JUNIO 2016

Alejandro Cachán – Junio 2016

Agradecimientos

En primer lugar, mi máximo agradecimiento va dedicado a Marta González Gazagaechevarria por haber sido mi pilar durante estos dos años y medio, y haberme apoyado siempre en todas las decisiones tomadas. Por lo especial que es para mí, ella debe saber que yo siempre la llevo en el corazón.

Agradecer infinitamente a Luis Plaza Campos su continua predisposición y aportación de conocimientos, ya no solo ligada a la tesis, sino también por su apoyo y paciencia en el trabajo diario que me han permitido cumplir el sueño de formar parte de Liberty Seguros. No puedo olvidarme de mencionar la confianza depositada en mí por parte de mis jefes, y la facilidad con la que me he integrado tanto en el departamento de *Pricing* como de *Product Manager Auto* gracias a mis compañeros.

También quiero agradecer a la Universidad Carlos III de Madrid la oportunidad que me concedió hace dos años de poder continuar mi formación en una de las mejores universidades de España y Europa, rodeado de verdaderos expertos; siendo todavía más profundo el agradecimiento a los profesores José Miguel Rodríguez Pardo y Jesús Simón del Potro, por su conocimiento, motivación y participación en la mejora de este proyecto.

En estos reconocimientos personales no pueden faltar mi madre, mis abuelos y mis amigos que a lo largo de mi formación académica siempre han estado ahí ante cualquier contratiempo dándome ánimos para continuar hacia delante.

Solo puedo decir, gracias a todos!!!

Índice

1. Resumen

En la última década, la metodología *GLM* se ha convertido en el recurso más empleado en la modelización de tarifas de *Nueva Producción*. Acompañada durante este tiempo, las aseguradoras han desarrollado importantes plataformas de *"Big Data",* en donde se almacena abundante información, que posteriormente puede ser empleada, entre otras cosas, para mejorar la predicción del riesgo de los clientes.

Este hecho conlleva un *"problema"* de sobredimensionamiento en el proceso de tarificación, que en los próximos años, puede ser notable en la mayoría de empresas aseguradoras.

El objetivo de este estudio ha ido encaminado, casi en su totalidad, a introducir diferentes técnicas estadísticas especializadas en la selección de variables; aplicando estas técnicas como complementos a la modelización GLM.

Con el desarrollo de estas técnicas, se pretende reducir la complejidad de los análisis, maximizando la eficiencia de los recursos empleados durante el proceso de tarificación. Para llevarlo a cabo, será necesario emplear diferentes software estadísticos (SAS, R), que nos facilitaran los resultados del estudio.

Palabras Claves: *Ridge Regression, Lasso, Elastic Net, Stepwise, GLM, SAS, R, Big Data, Tarificación, Machine Learning.*

2. Abstract

In the last decade, the GLM methodology has become one of the most ultilized resource of modelization in the tariff of new business.

During this period, insurers have developed the platforms of *"Big Data" significantly*, which is a wealth of information who can improve the accuracy of risk prediction of customers subsequently.

This methodoloy involves a *"problem"* of oversizing in the pricing process, but in the next few years, will be remarkable in most of the insurance companies.

This study has been totally targeted on applying different statistical techniques, especially in variables selection. All of these techniques can be complements for the GLM modeling process.

With the development of these techniques that can reduce the complexity of analysis, maximizing the efficiency of resources in pricing modeling. Several statistical softwares are necessary and helpful for achieving our study results.

Keywords: *Ridge Regression, Lasso, Elastic Net, Stepwise, GLM, SAS, R, Big Data, Pricing, Machine Learning.*

3. *Introducción*

El ilimitado desarrollo que las áreas de infraestructura analítica están construyendo en sus diferentes plataformas de *"Data Warehouse",* incorpora un valor añadido a esas empresas aseguradoras que las implementan, al disponer éstas de cuantiosa información acerca del negocio que a medio-largo plazo fructificará en un mayor rendimiento y beneficio.

Pero esa ingente cantidad de datos disponibles puede tratarse en algunos casos de un *"problema"*, ya que debe saber capturarse, tratarse y analizarse; para poder experimentar todos sus beneficios.

Nadie duda de que estos recursos continuaran su expansión, viéndonos *"obligados"* a necesitar de técnicas estadísticas que faciliten la compleja manipulación de trabajar con estas cantidades de variables. Sabemos que conocer estas técnicas conllevara un coste de tiempo y esfuerzo tremendo para los actuarios, al no tratarse de metodologías simples, pero esperamos que sean muy valiosas de cara al futuro en los departamentos de pricing.

Ante esta futura demanda del mercado de conocer nuevas técnicas estadísticas que sean capaces de manipular esa cantidad de variables, el objetivo de este estudio ha sido complementar la modelización GLM con nuevas metodologías procedentes de ramas de la *inteligencia artificial* (*Machine Learning*); proponiendo así diversas alternativas que ayuden en la ardua tarea de seleccionar variables (*Ridge Regression, Lasso, Elastic Net*), en escenarios con un alto dimensionamiento de variables.

Por todo lo que acabamos de comentar, se trata de un desarrollo académico novedoso, ante lo reciente que datan las técnicas centrales del trabajo (desarrollándose en torno al año 2005), la necesidad de encontrar nuevas técnicas estadísticas para poder trabajar con la minería de datos procedente de *"Big Data"* y sus escasas implicaciones en el mundo actuarial, las cuales, prácticamente se han visto focalizadas en su totalidad a la rama de vida (estudio sobre el *"Análisis de Supervivencia"* por parte del profesor **Jesús Herranz**).

Para poder llevar a la práctica una tarificación en espacios de alta dimensionalidad, hemos modelizado la Frecuencia de Siniestros RC Material Culpa para la rama de seguros de automóviles. Se trata, junto a la **Frecuencia de Siniestros RC Lesiones**, de la garantía más estudiada por los departamentos de Gestión del Negocio, con la finalidad de determinar cuánto porcentaje de "*malos clientes*" presentan cada una de las carteras.

Aprovechando el gran número de variables externas disponibles, se incorpora un segundo objetivo secundario como es: *"verificar que las variables externas tienen un impacto positivo en la predicción de la Frecuencia de Siniestros RC Material Culpa"*.

Para realizar este estudio ha sido necesario conocer el lenguaje de programación de *"R",* con el que, conseguir obtener la selección de variables óptima, y *"SAS Enterprise Guide"* para realizar: la manipulación, depuración, análisis y categorización de las variables y observaciones originales de la base de datos; además de las modelizaciones GLM definitivas.

El empleo de estas nuevas técnicas estadísticas junto a la mejora predictiva de la incorporación de variables externas, posibilitara a las aseguradoras mejorar la discriminación de riesgos en sus tarifas; a las que se pueden sumar otros muchos *"nuevos"* mecanismos que emplean las aseguradoras en sus tarificaciones, y que no hemos tratado en dicho estudio académico, como pueden ser el empleo del*: Rating, Zonning…*

El presente artículo se ha estructurado de la siguiente manera: en primer lugar, contextualizaremos la historia del Seguro del Automóvil para conocer como hasta no hace muchos años, el funcionamiento del mercado de seguros era completamente opuesto al que hoy en día conocemos. En la actualidad, el mercado del seguro de automóvil se ha vuelto terriblemente competitivo ajustando al máximo sus tarifas; pudiendo ser visible la convergencia, en términos de prima, con las otras ramas aseguradoras de no vida.

Tras presentar ambos escenarios antagónicos, creemos conveniente mostrar un capítulo de tarificación del riesgo, en el que, se detallará diferentes teorías de tarificación, posibles distribuciones a emplear para modelizar la frecuencia o la severidad (aunque como dijimos anteriormente, nosotros únicamente emplearemos las

distribuciones ligadas a la frecuencia) o la desigualdad a la hora de tarificar riesgos procedentes de nueva producción o retenidos en la cartera.

En el capítulo 7, nos detenemos en explicar las peculiaridades de los modelos lineales generalizados (*GLM*), con el que, la mayoría de las aseguradoras tarifican sus riesgos de nueva producción modelizando, creyendo conveniente en este estudio, emplear dicha modelización.

Durante el capítulo 8 y 9, introducimos brevemente el concepto de *"Big Data"* y *"Machine Learning"* por la envergadura que ambos conceptos poseen en el estudio y en el futuro de la tarificación.

Tras los capítulos enumerados anteriormente, nos situaremos en la parte central del proyecto. En primer lugar, realizamos un análisis teórico de los modelos de selección de variables estudiados y aplicados (*Stepwise, Ridge Regression, Lasso, Elastic Net*). Tras estudiar estos complejos métodos de selección estadísticos, intentamos llevar a la práctica lo visto en el capítulo teórico. Para ello, desde un inicio disponemos de una base de datos a la que debemos de aplicarle un correcto tratamiento, para posteriormente comenzar con un primer proceso de depuración de variables. Tras llevarse a cabo esta etapa de tratamiento, se profundiza en el estudio de las variables restantes, realizando para ello un pormenorizado análisis univariante, bivariante y de asociación entre las variables. Este proceso nos permitió categorizar con rigor las variables cuantitativas y recategorizar en determinadas situaciones variables externas previamente categorizadas. Por último, establecemos la distribución que mejor se ajusta a la **frecuencia de RC Material Culpa**.

Por último, en los capítulos 12 y 13 se procederá a llevar a cabo los modelos de selección de variables empleados, comprobando si los resultados obtenidos resultan consistentes con la modelización GLM y pueden emplearse como complemento en los procesos de tarificación.

4. Contextualización Histórica del Seguro del Automóvil en España

El mundo del seguro en España, no solo el del automóvil, ha cambiado mucho en muy poco tiempo. Hasta hace poco más de 3 décadas, las entidades aseguradoras no eran libres de poder fijar sus propias tarifas, sin justificarlo ni comunicarlo previamente a los supervisores de la Dirección General del Seguro. Es más, las restricciones y el control que la administración poseía sobre las aseguradoras era tal, que incluso, si éstas deseaban establecer alguna variación en sus tarifas, debían necesitar de re-aprobaciones constantes que impedían esa libre competencia entre empresas.

Esto hoy en día nos parece algo impensable, ya que constantemente las aseguradoras se encuentran retocando sus tarifas tras haber realizado periódicamente diferentes estudios de seguimiento sobre: el comportamiento de ventas de la tarifa NB[1], la siniestralidad, la competencia…, pero como detallamos hasta hace muy pocos años la situación era muy diferente.

La historia de los seguros españoles se fragmenta en dos períodos temporales que lo delimita el año 1980. Hasta esa fecha, dicho período es conocido por los historiadores como la *"Edad Media"*, mientras que a partir de esa fecha se la denomina *"Edad Moderna"*, caracterizándose éste último por su rápida transición hacia un mercado de libre competencia adecuado a la intención de crear un mercado de empresas privadas, dinámicas e internacionales.

Este es el escenario en el que conviven todos los ramos aseguradores en la actualidad en España, pero el ramo de autos tuvo que sufrir su propio calvario para conseguirlo.

Mientras que en 1980 y 1982, seguros como el de accidentes, enfermedades, transporte o incendios entre otros conseguían liberalizarse gracias al **Real Decreto 1335/1979 de 1980**, y a la **Orden Ministerial del 22 de Octubre de 1982**; otros muchos seguros como el de automóvil tuvieron que esperar hasta 1990, para poder fijar con

[1] Nueva Producción.

total libertad todas las coberturas que conforman el seguro de autos, bajo el decreto **ley 21/1990**.

Con la **ley 21/1990** se liberaliza el resto de seguros distintos de vida, adaptándose a la normativa europea que marcaba la **Directiva 88/357/CEE de 1988**.

Con la libertad plena por parte de las aseguradoras para fijar la tarifa en autos, aparecía un problema que en realidad siempre ha existido pero que ahora se acrecienta, la **tarificación del riesgo**.

Desde el inicio del mercado asegurador de automóviles en España, las aseguradoras no han sido nunca capaces (hasta ahora) de cuantificar correctamente el riesgo de sus clientes ante la falta de:

- Conocimientos.
- Experiencia.
- Libertad plena en la toma de decisiones.

Este problema todavía se vio agigantado con la obligatoriedad de disponer de un seguro de responsabilidad civil a partir de 1965 como fijaba la **Ley de Uso y Circulación de Vehículos a Motor 122/1962,** provocando que se duplicaran el número de asegurados en las carreteras, acrecentando el problema de las aseguradoras españolas.

La mayoría de empresas aseguradoras tuvieran resultados negativos durante amplios períodos de tiempo como consecuencia de la volatilidad de la siniestralidad, que ha *"oscilado históricamente entre un 75% y un 90%"* **(Embid. I, 2011).** Solo una gran empresa pudo escapar y crecer a pesar de la dificultad para las aseguradoras en los años 60 y 70, como es: la **Mutua Madrileña,** que con su política de costes bajos[2] y ausencia de accionistas[3], consiguió situarse entre en el top de aseguradoras.

Este problema todavía se vio agigantado con la obligatoriedad de disponer de un seguro de responsabilidad civil a partir de 1965 como fijaba la **Ley de Uso y Circulación**

[2] Esta política pudo ser posible al disponer únicamente de un canal de venta, **Directo** el cuál, no trabaja con agentes, los cuales, perciban comisiones.
[3] La empresa al estar constituida como una mutua no debía a los accionistas, pudiendo así aplicar mayores descuentos a los mutualistas.

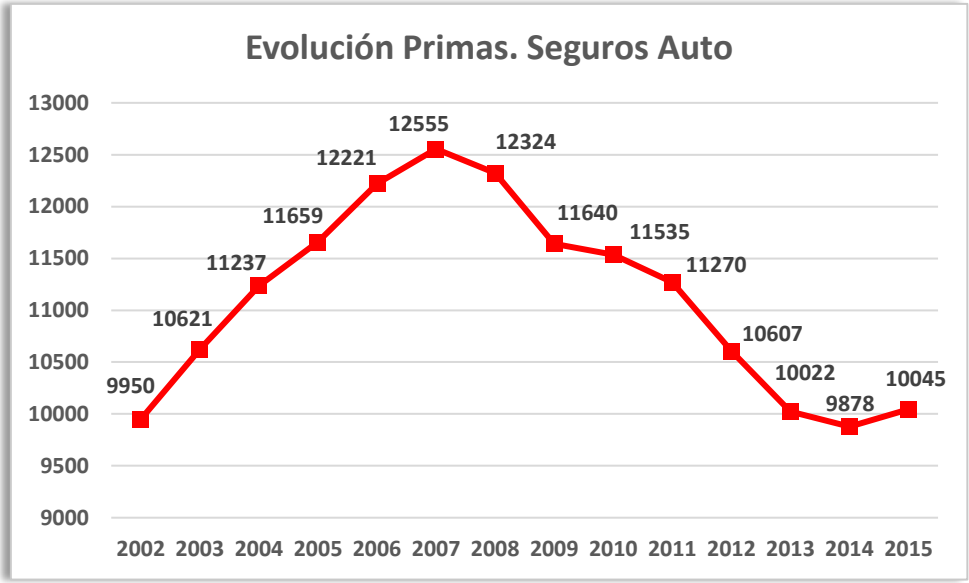
de Vehículos a Motor 122/1962, provocando que se duplicaran el número de asegurados en las carreteras, acrecentando el problema de las aseguradoras españolas.

De ahí la importancia de crear tarifas de NB con las que poder competir en el mercado sin poner en peligro la viabilidad contable de las empresas.

5. Mercado de Seguros de Automóviles Actual

Tras la crisis económica del período 2007-2014, todos los datos económicos hacen indicar que el sector asegurador del automóvil vuelve a despegar, en términos de primas.

Grafica 5.1. Evolución Primas en millones de euros. Seguros Auto



Fuente: **Serie Histórica. Icea. Gráfico elaborado por el autor.**

Como vemos en la gráfica 2.1, tras una larga y pronunciada caída de los niveles de prima total desde 2007 hasta situarnos por debajo de la prima de 2002 en 2014, las primas totales vuelven a crecer un 1.69% entre el período 2014-2015. Como explica Ramón Nadal en su artículo *"La eficiencia en precios y costes, como elementos clave para mantener la competitividad en el seguro del automóvil"*, las noticias que llegan desde el mercado presagian que la guerra de precios ha terminado.

Otros indicadores como el número de matriculaciones de coches o el aumento del uso del combustible corroboran lo expuesto anteriormente.

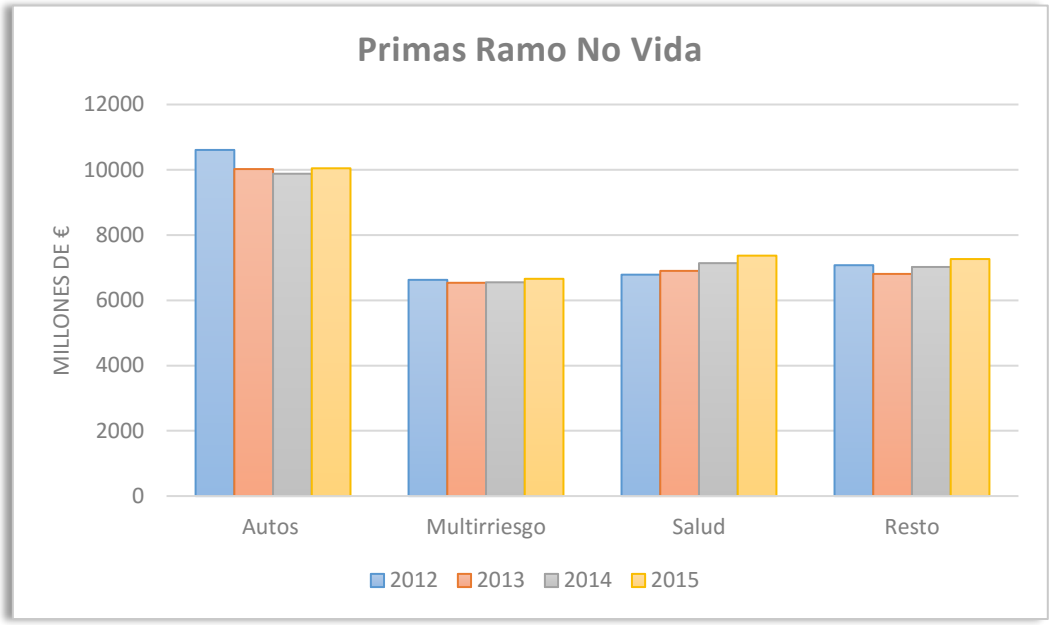
Pero las expectativas de futuro no únicamente se sitúan por la parte de los ingresos, por la parte de los gastos se apunta a que los costes aumentaran como consecuencia de la reforma del Baremo[4] del 1 enero de 2016.

[4] Según la **Ley 35/2015**: *"se reforma la valoración de daños y perjuicios causados a las personas en accidentes de circulación"*.

Esta incertidumbre, en términos de ingresos y costes, provoca que muchas aseguradoras estén reaccionando de manera muy diferente, ya sea, modificando sus tarifas para impedir la entrada de clientes con malos riesgos, bajando primas para obtener mayor cuota de nuevos clientes...

Aún con la caída de primas de los últimos 7 años, el ramo en seguros de no vida que encabeza la facturación sigue siendo los automóviles; aunque durante este período de crisis económica unida a la guerra de precios entre aseguradoras, el resto de ramos han convergido hacia los niveles de facturación de autos, disminuyendo la brecha de facturación existente entre el resto de ramos aunque sin alcanzarles. En la actualidad, el peso que representa los seguros de vida es algo más de un tercio, en contraposición a la mitad que representaba el año 2002 **(Nadal. R, 2014)**.

Grafica 5.2. Primas Ramo No Vida



Fuente: Serie Histórica. Icea. Gráfico elaborado por el autor.

6. Tarificación de Seguros de Autos

6.1. Teoría del Riesgo Individual vs Colectivo

La tarificación en Seguros de No Vida consiste en predecir el coste esperado de cada póliza para un período temporal, que por lo general, suele ser de un año. Para ello, existen dos teorías que intentan medir el coste agregado de una cartera:

- **Teoría del Riesgo Individual.**
- **Teoría del Riesgo Colectivo.**

La **Teoría del Riesgo Individual** parte de la premisa de identificar el coste esperado de cada póliza por separado agregando posteriormente todas las pólizas y estableciendo el coste agregado de la cartera.

$$Y = \sum_{i=1}^{N} S_i$$

La teoría del Riesgo Individual sería lo ideal pero no deja de ser algo utópico en la actualidad, ya que para ello necesitas disponer de una *"gran masa de información"* que no dispones cuando presupuestas a un potencial cliente. Por eso, empleamos la Teoría del Riesgo Colectivo.

La teoría del Riesgo Colectivo consiste en estudiar el coste agregado *Y* de una cartera. Dicha cartera debería tener riesgos lo más homogéneo posible y ser independientes entre sí.

Tarificar la prima de un seguro es un proceso estocástico que refleja dos variables a predecir: el **número de siniestros** y la **severidad de cada siniestro**.

$$\boldsymbol{S_t = X_1 + X_2 + \cdots + X_n}$$

Siendo la esperanza del coste agregado de toda la cartera:

$$\boldsymbol{E(Y) = E(N) * E(S)}$$

6.2. Distribuciones Modelizadoras de la Frecuencia Siniestral

Para modelizar estadísticamente la variable *"Número de Siniestros"* empleamos **distribuciones de conteo**, aunque la más utilizada de todas es la ***distribución de Poisson*,** siendo su distribución de probabilidad y sus momentos de orden uno y dos los siguientes:

$$\boldsymbol{P(N = x) = e^{-\lambda}\frac{\lambda^x}{x!}}$$

$$\boldsymbol{E(N) = Var(N) = \lambda}$$

La distribución Poisson presenta tres anomalías muestrales muy comunes que hay que tenerlas en cuenta a la hora de analizar la variable respuesta como son: el *contagio, la sobre-dispersión* y el *inflado de ceros.*

- El *contagio* denota que los siniestros tienen una propensión o unos tiempos de espera diferente.
- La *sobre-dispersión* se da cuando la varianza muestral es diferente a la media muestral.
- El *inflado de ceros* se observa cuando la muestra presenta una frecuencia elevada de ceros. El inflado de ceros provoca sobre-dispersión.

Para solventar la anomalía del inflado de ceros y la sobre-dispersión se puede aplicar un ***modelo de distribución Poisson inflado de ceros***. Este modelo asigna únicamente dos probabilidades, uno para los *"0"* y otros para el resto de observaciones muestrales distintas de 0, tratándose de una ***distribución de Poisson truncada***. Expresándose sus probabilidades como:

$$\boldsymbol{P(Y_i = 0) = q_i + (1 - q_i) * e^{-\lambda_i}}$$

$$\boldsymbol{P(Y = y_i > 0) = (1 - q_i) * \frac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!}}$$

El problema que tiene este modelo es que al estar confeccionado con la distribución de Poisson, no se sabe realmente si las observaciones se distribuyen como una *Poisson* o como una *Poisson inflada de ceros*.

Otra de las distribuciones más utilizadas para modelar la frecuencia de siniestros es la distribución **Binomial Negativa.** La distribución Binomial Negativa se puede obtener por dos procedimientos diferentes:

- Como una composición entre una Poisson y una Logarítmica.
- Como una mixtura entre una Poisson y una Gamma.

Se suele emplear dicha distribución cuando la propensión o la intensidad de los siniestros de una Poisson "λ" no son constantes. Esto puede darse por la heterogeneidad de una cartera.

Su probabilidad y sus momentos de orden uno y dos son:

$$\boldsymbol{P(Y = x) = p_x = \binom{x - r - 1}{r - 1} (p)^r (1 - p)^x}$$

$$\boldsymbol{E(Y|X_i) = r * \frac{1 - p}{p}}$$

$$\boldsymbol{Var(Y|X_i) = r * \frac{1 - p}{p^2}}$$

Si la muestra con la que trabajamos presentará muchos ceros, sería conveniente emplear una distribución truncada, que en este caso se denomina, **Binomial Negativa inflada de ceros.** Al igual que la distribución truncada para una Poisson, la distribución truncada para una Binomial Negativa asigna únicamente dos probabilidades: una para los *"0"* y otra para todo lo que no sean *"0"*, expresándose las probabilidades y sus momentos como:

$$\boldsymbol{P(Y_i = 0) = q_i + (1 - q_i)(p)^r}$$

$$\boldsymbol{P(Y_i > 0) = (1 - q_i) \binom{x - r - 1}{r - 1} (p)^r (1 - p)^x}$$

No se trata de la distribución de conteo más empleada por las aseguradoras, pero muchas veces pueden ser los más apropiados para emplearse, sobre todo, si la variable respuesta presenta muchos ceros en sus observaciones.

Por lo tanto, hay que tener mucho cuidado con la distribución a escoger para modelizar el número de siniestros de una garantía, ya que las observaciones pueden presentar muchos ceros obligando a emplear distribuciones truncadas.

6.3. Distribuciones Modelizadoras de la Severidad Siniestral

Para modelizar la severidad de los siniestros de una muestra empleamos otras distribuciones completamente diferentes a las analizadas anteriormente para modelizar la frecuencia de los siniestros. A priori la distribución más empleada para modelar la severidad es la **Gamma**.

La gamma es la distribución ideal para modelizar siniestros de cola corta (valores no catastróficos). Su función de densidad y sus momentos vienen dada por las siguientes expresiones:

$$f(S = x) = \lambda e^{-\lambda x} \frac{(\lambda x)^k}{r(k)}$$

$$E(x) = \frac{k}{\lambda}$$

$$Var(x) = \frac{k}{\lambda^2}$$

Otra distribución frecuente para modelizar la severidad de la siniestralidad es la **Log-Normal.** Se trata de una distribución más destinada a emplearse para siniestros con colas grandes. Su función de densidad y sus momentos vienen dada por las siguientes expresiones:

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

$$E(x) = e^{\mu+\sigma^2/2}$$

$$Var(x) = (e^{\sigma^2} - 1)e^{2\mu+\sigma^2}$$

Y por último, una distribución que igual no es tan frecuente emplearla para analizar la severidad es la **Pareto.** Dicha distribución se emplea fundamentalmente para modelizar riesgos catastróficos. Su función de densidad y sus momentos vienen dado por las siguientes expresiones:

$$F(S = x) = 1 - \left(\frac{\beta}{x + \beta}\right)^{\alpha}$$

$$E(x) = \frac{\beta}{\alpha - 1}$$

$$Var(x) = \frac{\alpha\beta^2}{(\alpha - 1)(\alpha - 2)!}$$

6.4. Características de los Riesgos a Cuantificar

Para realizar una nueva tarifa lo ideal sería disponer de una:

- Una **cartera grande**, debido a que agrupando riesgos individuales podemos diversificar el riesgo global de una aseguradora.
- La cartera debería presentar **riesgos individuales** entre sí, debido a que las pérdidas que sufra un individuo no deberían de afectar al resto de pólizas de la cartera.
- **Riesgos homogéneos**. Si la cartera a tarificar fuera heterogénea, la compañía estaría expuesta a la **selección adversa**, ya que algunos clientes se les estaría asignando una prima no equivalente a su riesgo, lo que propiciaría que aquellos individuos con un riesgo menor a lo que corresponde su prima abandonaran la cartera, permaneciendo dentro de ella aquellos individuos con riesgos más mayores.

  Si ocurriera la situación de que la cartera fuese heterogénea, la entidad aseguradora debería **subdividir la cartera en subcarteras homogéneas** para tarificar los riesgos sin selección adversa.

6.5. Sistemas de Tarificación

Existen dos enfoques de tarificación diferentes para los clientes: por una parte, la que se conoce como el **sistema de tarificación a priori (class rating)** y el **sistema de tarificación a posteriori (experience rating)**.

6.5.1. Sistema de Tarificación a Priori (Class Rating)

El sistema de tarificación a priori permite poder asignar una prima a una póliza de nueva producción sobre la que no se tienen datos propios sobre su siniestralidad, lo que dificulta asignar una prima a un riesgo desconocido.

Por ello, previamente a ese individuo se le pregunta una serie de factores de riesgo que permiten poder ir asignándole en diferentes grupos de riesgo, y que con ello se le pueda establecer una prima para la siniestralidad y el coste esperado.

Por lo tanto, cuando realizamos una tarificación a priori, la parte esencial de este análisis es la experiencia de la cartera. Con la experiencia de dicha cartera debemos conformar los niveles de riesgo significativos que predigan la siniestralidad esperada de esos grupos de riesgo homogéneos, siendo el GLM el modelo predictivo más utilizado.

6.5.2. Sistemas de Tarificación a posteriori (Experience-Rating)

En contraposición a lo especificado anteriormente, el sistema de tarificación a posteriori parte ya de una cierta experiencia sobre la siniestralidad de ese individuo, y por tanto, dispone también de una prima inicial que se irá modificando en cada período de renovación según la siniestralidad presentada por ese individuo. Para modificar la prima de tarificación a priori, las empresas aseguradoras emplean diferentes mecanismos permitiendo ajustar mejor la prima de la cartera. Estos mecanismos entre otros pueden ser:

- **Optimización.** Proceso de re-tarificación de un cliente cuando se procede a estudiar su renovación. En este proceso, se aplica una simulación de posibles escenarios en los que los clientes pueden caerse de la cartera y analizando la cartera según la elasticidad-precio de ese período temporal.
- **Scoring Técnico.** Técnica según la cual, se identifica el perfil de cliente y según su perfil se le asigna un rango de una clasificación previamente diseñada, por la cual, se le bonificará o se le recargará.

- **Saneamiento.** En los casos *"extremos",* en los que, se trate de un cliente con muchos siniestros a lo largo de un período temporal, la empresa puede someterle a un proceso de saneamiento, por el que, se le anulará el contrato.

Ahora las clases de riesgo que nos encontramos ya no son homogéneas como antes. Esos individuos puede que hayan tenido una siniestralidad diferente a la esperada, habiendo sido agrupados incorrectamente al considerar mal los factores de riesgo, o lo más lógico, no haber tenido en cuenta otra serie de factores desconocidos para la compañía aseguradora como pueden ser: la **forma de conducción** o el **diferente comportamiento de los individuos**, entre otros factores. Aunque las aseguradoras intentan disponer de datos acerca de estos factores de riesgo, implementando diferente *know-how* tecnológico que proporcione información acerca de la forma de conducción del individuo, como puede ser el *OnStar* creado por General Motors, *k2k* creado por Telefónica y empleado por Generali Seguros o el *UBI* fabricado por Vodafone con la ayuda de Tower Watson, o analizando el comportamiento de los individuos en las redes sociales o estudiando la inteligencia computacional de los individuos. De todos modos, para que estas medidas se pongan en funcionamiento todavía vamos a tener que esperar unos cuantos años pero posiblemente serán el futuro.

6.5.3. Fichero Histórico del Seguro del Automóvil (SINCO)

Como hemos dicho, la información sobre la siniestralidad histórica de un individuo es muy importante para tarificar lo mejor posible la entrada de una póliza a nuestra cartera. Para ello, las aseguradoras disponen de un *"Fichero Histórico del Seguro del Automóvil"* facilitado por TIREA[5], con el que es posible ajustar mejor la prima del cliente, según su historial siniestral de los últimos 5 años. Este sistema se denomina *Bonús-Malus*, y consiste en bonificar o recargar la prima pura del individuo según sus antecedentes. Mencionar que las entidades aseguradoras reciben los datos de siniestralidad de los individuos, y según sus criterios propios fijaran las bonificaciones o los recargos que impondrán a sus nuevos clientes.

[5] Tecnología de la Información y Redes para las Entidades Aseguradoras.

Este sistema de bonus se emplea en las tarificaciones a priori, ya que en las renovaciones aplicaremos un sistema a posteriori, en el que, tendríamos en cuenta todos los mecanismos comentados en el apartado anterior que permiten un mejor valoración del riesgo.

Pero este método de tarificación empleado en la actualidad por las aseguradoras tiene varios problemas:

- El fichero común SINCO solo recoge los datos de siniestralidad que los conductores asegurados han sufrido durante los últimos 5 años para la garantía ***"RC Material Culpable"***. Esto significa que, si por ejemplo, un individuo que durante los últimos años ha declarado varios siniestros por daños propios y no haya tenido ningún siniestro "RC Material Culpable" aparezca con un 0 en siniestralidad aunque como sabemos si ha generado siniestralidad para la antigua empresa aseguradora, no permitiendo recoger toda la información deseable para la nueva empresa aseguradora.
- En segundo lugar, no todas las entidades aseguradoras están adheridas al fichero común de siniestros SINCO, lo que provoca una cierta reticencia del resto de empresas aseguradoras a aceptar esos riesgos, al desconocer que clientes estas aceptando en tu cartera[6].

[6] Hay muchas empresas que ante este problema, solicitan al cliente que la otra empresa aseguradora les facilite el historial siniestral de ese cliente durante su vinculación con esa empresa aseguradora.

7. Modelos Lineales Generalizados

Hasta 1972 todo análisis predictivo se realizaba a través del "Modelo Lineal General", hasta que en dicho año dos estadísticos británicos llamados **Nelder** y **Wedderburn** revolucionan el mundo estadístico actual del que se nutre la tarificación no vida, desarrollando una extensión del anterior modelo, conocido como: Modelo Lineal Generalizado (más conocido por sus siglas, GLM).

Dicho modelo presenta varias virtudes en comparación con el Modelo Lineal General permitiendo solventar distintas situaciones tan comunes en el mundo actuarial como pueden ser:

- Modelizar distribuciones con errores no normales.
- Transformar la variable dependiente, linealizandola a través de lo que se conoce como: la **Función Vínculo, Enlace o de Ligadura**. Ésta función nos permite describir relaciones no lineales entre la variable dependiente y las variables explicativas del modelo original. Esta falta de relación de identidad entre los valores ajustados y el predictor era impensable en los Modelos Lineales Generales.
- Presentar modelos con varianzas heterocedasticas.

Los Modelos Lineales Generalizados se construyen a través de tres componentes:

- Una **Componente Aleatoria** que hace referencia a una distribución de probabilidad de la variable respuesta *"y".*
- La **Componente Sistémica** que está conformada por una matriz de variables explicativas a las que se las asigna parámetros, pudiéndose describir en términos matriciales como $X\beta$.
- La **Función Link**, la cual, es capaz de combinar linealmente la componente aleatoria y sistémica. Las Funciones Link más empleadas en la modelización de pricing no vida suele proceder de distribuciones pertenecientes a la familia de la exponencial conociéndose estas funciones link con el término de **Funciones de Enlace Canónicas**. En el próximo cuadro, vamos a detallar las funciones canónicas más empleadas:

Función	Función Link	Formulación
Normal	Identidad	μ
Binomial	Logit	$Log\left(\frac{\mu}{n-\mu}\right)$
Poisson	Logarítmica (Log)	$Log(\mu)$
Gamma	Recíproca	$\frac{1}{\mu}$

Un excelente ejercicio sería comparar las expresiones matemáticas del Modelo Lineal General frente al Modelo Lineal Generalizado con la finalidad de comparar las diferencias esenciales mencionadas entre ellos:

Modelo Lineal General	Modelo Lineal Generalizado
$y_i = \beta_0 + \beta_1 * x_{1i} + \beta_2 * x_{2i} + \cdots + \varepsilon_i$	$y_i = \beta_0 + \beta_1 * x_{1i} + \beta_2 * x_{2i} + \cdots + \varepsilon_i$
$\eta_I = \mu_I$	$\eta_I = g(\mu_I)$
$\varepsilon_i \sim (0, \sigma^2)$	$\varepsilon_i \sim (0, \sigma_i^2)$

Los departamentos de pricing no vida emplean está técnica estadística con la finalidad de:

- ❖ Modelar sus nuevas tarifas para cada producto y ramo de la compañía. Con esta técnica intentamos discriminar al máximo el precio en función del riesgo del cliente.
- ❖ Realizar un detallado control de la clase de asegurados de los que está compuesta la cartera de ese negocio, clasificando a los clientes a través de *Scoring Técnicos* y recargando o descontando las primas por renovación de éstos.

Pero siempre hay que tener en cuenta que para realizar un modelo GLM se debe disponer y evaluar los siguientes puntos que a continuación detallaremos:

- El cimiento sobre el que se debe sustentar toda modelización GLM es la **base de datos**. La disposición de una buena base de datos acompañada de un correcto empleo es la parte fundamental del estudio. Con la base de datos tratada es necesario analizar los datos disponibles, con el objetivo de conocer los datos con los que se va a trabajar a través, por ejemplo, de gráficos

univariantes/bivariantes o matrices de correlación. Con este tipo de análisis, seremos consecuentes antes de realizar el modelo, sobre la necesidad de transformar las variables, analizar la posible presencia de multicolinealidad entre las variables…

- Tras un primer análisis de las variables a emplear, es necesario escoger la **estructura de errores y las funciones de vínculo** que mejor predicen la variable dependiente. Para ello, puede ser buena idea comparar distintas funciones de vínculo para comprobar cuál de ellas se ajusta mejor al modelo.
- Como en cualquier otro modelo estadístico, cuando fijamos nuestra hipotética regresión es necesario observar la cantidad de variabilidad explicada por el modelo, conociéndose esta terminología en los *"Modelos Lineales Generalizados"* como "*Devianza*" (D). Mencionar que para cada una de las posibles distribuciones de la variable respuesta, el cálculo de la *devianza* difiere matemáticamente, aunque en términos genéricos se expresa como la diferencia de un modelo residual con todos los parámetros y un modelo estimado sin ninguna variable.

$$D = -2\log[L(\hat{\mu}; y) - L(y; y)]$$

En términos generales, a mayor deviance, más pobre será el ajuste del modelo

Distribución	**Devianza**
Normal	$\sum_i (y_i - \hat{\mu}_i)^2$
Poisson	$2 * \sum_i [y_i log\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i]$
Binomial	$2 * \sum_i w_i m_i [y_i log\left(\frac{y_i}{\mu_i}\right) + (1 - y_i) * log\left(\frac{1-y_i}{1-\mu_i}\right)]$
Gamma	$2 * \sum_i w_i [-log\left(\frac{y_i}{\mu_i}\right) + \left(\frac{y_i - \mu_i}{\mu_i}\right)]$
Binomial Negativa	$2^* \sum_i [y_i log\left(\frac{y}{\mu}\right) - \left(\frac{y+w_i}{k}\right) log\left(\frac{\frac{(y+w_i)}{k}}{\frac{(\mu+w_i)}{k}}\right)$

Mientras que, la significación de los estimadores vendrá dada por el coeficiente:

$$z = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)}$$

Donde:

$$se(\hat{\beta}_i) = \sqrt{x_i * w_{ii}}$$

$$w_{ii} = \frac{1}{Var(y_i)} * \left(\frac{\partial u_i}{\partial y_i}\right)^2$$

- Unido con el punto anterior, podemos utilizar diferentes **criterios para evaluar** los diferentes **modelos** propuestos. Los dos criterios estadísticos más empleados para examinar la idoneidad del modelo son: el **Criterio de Información de Akaike** (AIC) y el **Criterio de Información Bayesiano** (BIC). Dichos modelos califican los modelos considerando la incorrelación residual de las variables y la cantidad de parámetros empleados. Como vemos en las expresiones siguientes, el criterio BIC penaliza en mayor grado el empleo de un mayor número de variables[7], interesándonos que los modelos sean los modelos sean lo más simplificados posibles siguiendo el principio de parsimonia.

  $$\boldsymbol{AIC = 2\,K - 2\,ln(L)} \qquad\qquad \boldsymbol{BIC = -2\,ln(L) + K\,ln(n)}$$

  Donde los parámetros anteriores hacen referencia a:

  - *K:* número de parámetros del modelo.
  - *L* es el máximo de la función máximo verosimilitud.

- Por último y no menos importante, es muy interesante **estudiar** los **residuos** obtenidos del modelo estimado. En muchas situaciones nos encontraremos que la estructura de errores y la función de vínculo escogida no son las idóneas para los datos presentes, debiendo modificar el modelo propuesto. Además con este estudio, podemos identificar atípicos no detectados en la primera etapa (análisis inicial de los datos) que puedan distorsionar la modelización.

[7] Puede incluso que en algunas ocasiones, la valoración del número de variables a través de la metodología BIC este demasiado penalizada por la inclusión de variables. En esos casos extremos, tomaremos como referencia el **Criterio de Información de Akaike** para seleccionar el modelo de variables internas y externas.

8. *Big Data*

Hoy en día, el activo más importante para una empresa aseguradora son sus datos.

Esto ha generado que en la última década, los equipos de *"Analytics"* tomen mucha importancia dentro de las empresas aseguradoras ante la expansión del *"Big Data"*. El *"Big Data"* no deja de ser más que la recolección y el tratamiento de un repositorio de datos que los individuos generan diariamente con sus comportamientos. Estos departamentos permiten a los actuarios de pricing disponer de un gran volumen de información tratada, y en muchos casos, reportada que posibilitan: la incorporación de nuevas variables discriminantes a la modelización GLM de las tarifas, la rápida reacción de los equipos de gestión de negocio frente a posibles desviaciones del plan de la compañía (como puede ser, en términos, de primas, siniestralidad, posicionamiento…), aprovecharnos de oportunidades que nos ofrezca el mercado frente a la competencia….

Se trata de una tarea ardua y pesada, ya que, recopila información muy densa y dispar, pero como vemos, los beneficios que repercuten a las empresas son muy importantes en su actividad diaria.

Es más, el "Big Data" sigue en continuo crecimiento dentro de las aseguradoras mundiales, habiéndose creado en la mayoría de empresas grandes, diferentes estructuras, como pueden ser el *"Data Warehouse"*, que canalizan los flujos de datos, recopilando la información y derivando en *"Data Marts*[8]*"*. Entre la información tan peculiar que podemos encontrar destacan: el **comportamiento de los individuos en las redes sociales, hábitos de navegación** o incluso **variables cognitivas**[9]; que en un futuro cercano servirá a las aseguradoras para completar la tarificación de los individuos.

La industria aseguradora no se detiene en la innovación de nuevas técnicas avanzadas actuariales. Posiblemente, el futuro del pricing no vida se centre en la modelización a través de los "Modelos de Aprendizaje Automático" o también conocidos como, "Machine Learning", que sea capaz de gestionar eficazmente la enorme masa de información de la que se dispondrá.

[8] Estructuras que realizan las áreas analíticas.
[9] Conjunto de factores que influyen en el aprendizaje de una persona.

De ahí que, nuestro estudio académico este centrado en la complementación de herramientas de estadísticas avanzada en la modelización GLM ante la alta dimensionalidad de variables de las que disponemos en una base de datos cuando realizamos una tarifa de NB.

9. *Machine Learning (Aprendizaje Automático)*

Como sabemos, en el área de seguros no vida la competencia entre las compañías es atroz y la identificación del *"buen riesgo"* es muy importantísimo para los resultados de la compañía, lo que obliga a que las compañías de seguros deban estar en continúa renovación en cuanto a sus tarifas se refieren. Las técnicas avanzadas actuariales *"Machine Learning"* empiezan a introducirse lentamente en el mundo actuarial, aunque su implantación en el mercado asegurador español es prácticamente nula.

El amplio desarrollo del *"Big Data"* junto con las técnicas *"Machine Learning"* han permitido a un amplio abanico de modelos predictivos, aunque la mayoría de ellos son *"especies de cajas negras a la espera de ser abiertas"* e investigadas en profundidad para entender su funcionamiento.

Podemos definir el *"Machine Learning"* como: una rama de la inteligencia artificial capaz de automatizar comportamientos aplicando diferentes algoritmos de aprendizajes.

Entre las diferentes técnicas que componen el mundo del "Machine Learning" aparte de selección que detallaremos en el siguiente capítulo, destacamos:

- **Redes Neuronales Artificiales (RNA):** Se trata del principal exponente del *"Machine Learning."* Aunque se empezó a estudiar en la analogía biológica con el cerebro, ha sido capaz de extrapolarse al mundo actuarial de vida y no vida. Su modelación es muy compleja, pero puede simplificarse como la suma de *"n"* inputs, cada uno de ellos ponderados e interconectados entre sí creando un modelo capaz de transformar los inputs en una salida, ya sea binaria, de rango [-1,1] o continuas. En pricing no vida, los primeros estudios han estado encaminados a predecir una posible caída de cartera de los asegurados.
- **Máquinas de Vectores de soporte (SVM):** Es capaz de supervisar el análisis de los datos y reconocer diferentes patrones, pudiendo tanto clasificar como realizar predicciones de los datos, ya sean éstas, continuas o categóricas.
- **Árboles de Decisión (CHAID):** En este estudio se han empleado brevemente esta técnica para conformar clústers procedentes de variables numéricas

esencialmente, aunque también puede emplearse en las variables categóricas (como la marca del automóvil). Su funcionamiento se basa en agrupar según unos algoritmos[10].

[10] Esta metodología será explicada más detalladamente en el capítulo de *"Categorización de Variables"*

10. Modelos de Selección de Variables a través del Aprendizaje Automático

Dentro de lo que concierne al objetivo central del estudio, vamos a pasar a elaborar un estudio inicialmente teórico/matemático, aunque posteriormente práctico con el que, poder seleccionar variables a través de técnicas estadísticas extrapoladas al mundo actuarial. Se trata de plantear una solución a un problema bastante serio y costoso dentro de las aseguradoras, acrecentado en estos últimos años por la incorporación del uso del Big Data, sistema por el cual, podemos disponer de un mayor número de variables con las poder tarificar un producto recogiendo mejor el riesgo de los individuos; a sabiendas de que gran parte de las nuevas variables disponibles serán redundantes con el resto de variables o irrelevantes no viéndose aumentado en exceso el número de variables significativas dentro de las tarifas.

Cuando realizamos un modelo es importante tener en cuenta que la modelización debe tener el número óptimo de variables con las que poder predecir el número de siniestros y la siniestralidad de la cartera, ya que realizar un modelo con una gran cantidad de variables puede conllevar aumentar la cantidad de porcentaje explicado respecto a la variable dependiente pero a costa de aumentar la varianza de las variables y perder precisión en cada variable predicha, **sobre-ajustando** por tanto la modelización. En caso contrario, si seleccionamos un pequeño número de variables con las que pretendemos predecir el riesgo de los asegurados, esto propiciará que el modelo se encuentre **sub-ajustado** al poder predecir un menor porcentaje explicado de la variable dependiente que si introdujéramos alguna otra variable significativa, presentando cada una de las variables un mayor sesgo a costa de presentar una varianza muy pequeña. Por lo tanto, lo ideal sería disponer de un modelo que te sea capaz de seleccionar el número de variables óptimas con las que poder modelizar una tarifa.

Cuatro son las técnicas estadísticas, que vamos a estudiar y que permiten seleccionar y regularizar variables con el objetivo de modelizar en este caso una tarifa multivariante para el ramo de autos:

10.1. Selección Stepwise

Selección Stepwise: Se trata del método tradicional de selección de variables, en el que, se busca encontrar las variables que mejor expliquen la variable respuesta a través de una secuencia gradual de incorporaciones o extracciones de variables del modelo. Dicha técnica puede dividirse en las siguientes dos metodologías:

Según el método *Forward-Stepwise*, el proceso comenzaría prediciendo la variable respuesta a través de un único coeficiente, el intercepto, y a partir de ahí introduciríamos paulatinamente variables analizando la mejora del modelo.

Mientras que el método *Backward-Stepwise* sería el proceso contrario al anterior, donde partiríamos de un modelo compuesto por todas las variables seleccionadas en el estudio y a partir de ese punto, descartaríamos variables no relevantes para este modelo.

Esta técnica se desarrolla aplicando aplicando "*algoritmos greedy"*, la cual consiste en intentar buscar un óptimo global a través de los sucesivos óptimos locales encontrados:

Forward-Stepwise: $1 + (1 + 1) + (1 + 2) + \cdots + (p - 2 + 1) + (p) = p(p + 1)/2$

Backward-Stepwise: $p + (p - 1) + (p - 2) + \cdots + (p - p - 2) + (1) = p(p + 1)/2$

Dicho algoritmo presenta varios inconvenientes que hacen que la técnica *stepwise* pueda no seleccionar correctamente las variables a modelizar por las siguientes razones:

1. **El algoritmo no aplica una exploración combinada completa** con todas las variables sino que va realizándole secuencialmente según las variables que había escogido anteriormente.
2. **Problemas de selección con variables correlacionadas**. Si dispones de variables muy correlacionadas entre sí, puede que éstas no entren dentro del modelo a pesar de predecir mejor que la otra variable que permanece en el modelo.

3. Se trata de un **algoritmo muy inestable ante pequeños cambios en las observaciones**, pudiendo realizar grandes cambios en el modelo inicialmente fijado.

Con todos los inconvenientes mencionados en los tres puntos anteriores, es muy posible que no se alcance el óptimo global y por lo tanto, no sea el mejor modelo predictivo. Es más, es relativamente fácil que aplicando ambas metodología, Forward-Stepwise y Backward-Stepwise, no alcancemos el mismo modelo "óptimo" entre sí. Por esta razón, no es recomendable emplear esta técnica para seleccionar variables, aunque a pesar de ello, sea el modelo más conocido y empleado por las aseguradoras.

De ahí, la necesidad de búsqueda de otras técnicas de selección de variables actuariales avanzadas. Las técnicas que vamos a estudiar a continuación se enmarcan dentro de los modelos conocidos como *"Shrinkage"* o de *"Regularización".*

10.2. Modelos Shrinkage

Los modelos Shrinkage son métodos estadísticos capaces de regularizar (penalizar) coeficientes. Los modelos Shrinkage, más conocido en la actualidad dentro del mundo estadístico son: el modelo *Ridge Regression,* el modelo *Lasso* y por último, los modelos *Elastic Net*.

Pueden tratarse de modelos más interesantes y eficientes que las técnicas stepwise debido a que son capaces de regularizar las estimaciones de los coeficientes según sus varianzas. Además, se trata de técnicas no tan restrictiva como los modelos de regresión clásicos. El modelo de regresión clásico está basado en una serie de supuestos muy restrictivos conocidos como los supuestos de Gauss-Markov, que a continuación enumeraremos brevemente:

Supuesto 1 → Linealidad de los Parámetros

En un modelo poblacional, la variable dependiente está relacionada linealmente con las variables independientes x_p y el error ε_i

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 y_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

Supuesto 2 → Muestreo Aleatorio

Muestra aleatoria de tamaño n, $\{x_{pi}, y_i | i = 1, \dots, n\}$

Supuesto 3 → Variación Muestral en las Variables Explicativas

Las variables independientes no son iguales entre sí:

$$\{x_{pi} : i = 1, \dots, n\}$$

Supuesto 4 → Media Condicional Cero

El valor explicado del error ante cualquier variable dependiente es siempre cero.

$$E(\mu/x) = 0$$

Supuesto 5 → Homocedasticidad

El error μ tiene la misma varianza ante cualquier variable explicativa.

$$Var(\mu/x) = \sigma^2$$

Uno de los problemas clásico existente entre los modelos predictivos es la **existencia de Multicolinealidad.**

La **multicolinealidad** es la existencia de una fuerte correlación entre varias variables explicativas del modelo propuesto. Se trata de un problema muy grave para la predicción, demostrando a través de la siguiente ecuación matemática:

$$Var(\hat{\beta}_j) = \frac{\sigma^2}{STC_i(1 - R_j^2)}$$

Donde:

$$\sigma^2 \longrightarrow Var(y/x)$$

$STC \longrightarrow$ Suma Total de Cuadrados, es decir, es la medida de la variación muestral total.

$$\sum_{i=1}^{n} (x_{ij} - \bar{x}_i)^2$$

$(1 - R_j^2) \longrightarrow$ es el R^2 de x_j frente al resto de variables independientes. Es decir, R^2 denota las relaciones líneales entre las variables independientes.

Si R_j^2 se encuentra cercano a 1, eso quiere decir que esa variable independiente puede ser explicada en gran parte por la variable independiente i, estando ambas fuertemente correlacionadas.

La multicolinealidad entre variables provoca que sí:

$$R_j^2 \approx 1 \Longrightarrow Var(\hat{\beta}_j) \longrightarrow \infty$$

El hecho que haya multicolinealidad si somos estrictos, no viola el supuesto 3 de Gauss-Markov, pero provoca un aumento demasiado grande de la varianza.

La ventaja de los modelos *Shrinkage* es que *"fuerzan"* la reducción de la varianza de los estimadores aumentando el sesgo, y por tanto, penalizando los coeficientes de las variables analizadas. Con estas estimaciones, conseguiremos coeficientes más eficientes y una mejor construcción de "y" sobre "x_j".

Este proceso se consigue gracias a un parámetro de ajuste regularización, conocido en la literatura estadística inglesa como *"tunning"* (λ). Cuanto mayor es el λ, mayor será la penalización sobre los coeficientes de regresión. Por lo tanto, parte fundamental del problema será encontrar el λ óptimo. Esto se puede conseguir a través de múltiples procesos pero el más sencillo y visual será a través de un proceso de *validación cruzada* o *bootstrap*, el cual simula una secuencia de λ con la finalidad de encontrar el menor error de la predicción esperado.

Dentro de los modelos Shrinkage podemos diferenciar los modelos de esparsidad (*Lasso* y *Elastic Net*) y los modelos de no espasidad (*Ridge Regression*).

La diferencia entre este tipo de modelos es que los primeros son capaces de seleccionar variables, pudiéndose encargarse de solucionar el problema de construcción de un modelo regresivo, sobre todo ante un espacio de alta dimensionalidad de variables.

10.2.1. Ridge Regression

El método *Ridge Regression* (RR) es el *"método shrinkage"* más famoso de todos. Dicho método fue inventado por el matemático soviético, André Tikhonov a mediados del siglo XX ante la necesidad de encontrar solución a los *"problemas bien definidos"*. Con este modelo, Tikhonov desarrollo, lo que anteriormente hemos denominado, la regularización, pero no fue hasta los años 70 cuando Hoerl y Kennard la dieron a conocer en el mundo estadístico.

Aunque no se trata de un modelo de selección de variables, se estudia dentro de los modelos esparsivos al tratarse de la base matemática de la que se sustenta dichos modelos estadísticos avanzados.

Se trata de un modelo en el que todos sus predictores son penalizados, en mayor o menor grado, por el *tunning* según la varianza estimada.

Como sabemos, el mayor grado de varianza de una variable (o mejor dicho, una menor precisión de en la estimación) puede venir dado entre las principales razones por:

- Baja exposición de esa variable dentro de la base de datos.
- La multicolinealidad entre dos variables, como lo hemos comprobado anteriormente.

Esto quiere decir que el modelo Ridge Regression penalizará los coeficientes resultantes de la estimación lineal, minimizando los "β_j" según la varianza y el tunning seleccionado, como puede verse en la expresión matemática:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

Donde:

$\boldsymbol{\lambda}$: Parámetro capaz de controlar la cantidad de variables en un modelo a través de penalizaciones a los coeficientes.

Se puede expresar en términos matriciales como:

$$\hat{\beta}^{ridge} = (X'X + \lambda I)^{-1}X'y$$

En términos generales, los modelos estimados por Ridge Regression producen resultados más precisos que los obtenidos a partir de un Modelo Lineal Clásico, a menos que el menor error estimado en la validación cruzada se encuentre en $\lambda = 0$, con el que, no penalizarías las "β" permaneciendo las estimaciones igual que en la modelización MCO. La idea es penalizar la suma de cuadrados de los parámetros si hay correlación entre las variables.

Una forma equivalente de escribir el problema al método lagrangiano anterior sería:

$$\hat{\beta}^{ridge} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

$$s.a. \sum_{j=1}^{P} \beta_j^2 \leq s$$

El método Ridge Regression minimiza las "β_j", pero en su cálculo no se incluyen la minimización del intercepto "β_0", como deseo de querer mostrar la dependencia original existente entre origen y la variable dependiente.

10.2.2. Lasso

El modelo de selección *Lasso ("Least Absolute Shrinkage and Selection Operator")* es otro "*métodos shrinkage*" como el modelo *Ridge Regression* pero con importantes diferencias entre sí. En primer lugar, se trata de un modelo bastante más reciente que el modelo *Ridge Regression*. Data de 1996 cuando Tibshirani fue capaz de introducir dicho método en el mundo estadístico actual.

Se trata de un modelo que al igual que el modelo Ridge Regression permite penalizar los regresores, pero a que además a diferencia de lo que ocurre con el modelo explicado anterior, *Lasso* puede anular o excluir predictores del modelo.

Como en el caso del modelo *Ridge Regression*, cuanto mayor sea el "λ" mayor contración de los "β_j" hacia 0, y a su vez como consecuencia, mayor número de

exclusiones. Los estadísticos dicen que con el método Lasso "se produce una estimación de parámetro y selección de variables simultanea".

Matemáticamente, el modelo *Lasso* es muy similar al modelo *Ridge Regression*, con la diferencia esencial de que se incluye el intercepto dentro de la minimización del problema como puede verse en la expresión *Lasso*:

$$\hat{\beta}^{lasso} = \underset{\beta_0,\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Si preferimos desarrollarlo a través de un problema de optimización con restricciones

$$\hat{\beta}^{lasso} = \underset{\beta_0,\beta}{\operatorname{argmin}} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

$$s.a. \sum_{j=1}^{P} |\beta_j| \leq s$$

La penalización del problema anterior produce que la estimación de las variables independientes frente a la variable dependiente no sea lineal, y por lo tanto, no se pueda conseguir una expresión matemática del $\hat{\beta}_j^{lasso}$ como se había detallado con el $\hat{\beta}_j^{ridge}$ al no existir un mínimo global al que se tienda.

Como en el modelo anterior, aplicamos *"validación cruzada"* para seleccionar la estimación que minimiza el error predicho y que nos permitirá hacer una selección de variables según el parámetro de regularización, "λ". Esta metodología de fijar el "λ" seguramente sea la alternativa más eficiente para estimar los "β_j", ya que se soluciona a través de una solución lineal por tramos **(Efron et at, 2004)**, siendo la expresión matemática del algoritmo aplicado:

$$\hat{\beta}^{LASSO}(\lambda) = \hat{\beta}^{LASSO}(\lambda_k) + (\lambda - \lambda_k)\xi_k \qquad \lambda_k \leq \lambda \leq \lambda_{k+1}$$

Aplicando este algoritmo repetidamente conformaremos un camino completo de todas las soluciones de $\hat{\beta}^{LASSO}(\lambda)$, si $0 \leq \lambda \leq \infty$, aunque no serán soluciones óptimas.

10.2.3. Elastic Net

Se trata del modelo de regularización y selección de variables más reciente de todos los estudiados, en el que se consigue mezclar en un mismo modelo las ventajas del modelo *Ridge Regression* y del *Lasso,* siendo este último el modelo fundamental sobre el que se sustenta. El modelo *Elastic Net* fue propuesto por los estadísticos Zou y Hastie en 2005 como una variante mejorada del método *Lasso.*

Se trata de un modelo, que a diferencia de los dos anteriores, cumple las tres condiciones deseables que un modelo de regularización y selección de variables debe cumplir, según lo que detalla en su estudio **Fan y Li (2001**:

1. *Esparsidad.* El modelo debe efectuar selección automática de variables.
2. *Continuidad* en sus datos para ofrecer estabilidad a la predicción.
3. *Insesgadez.* El sesgo debe de ser bajo para valores altos de los coeficientes.

Si recordamos el estudio de los dos modelos anteriores recordamos que el modelo *Ridge Regression* no es un modelo de selección de variables, ya que no es capaz de anular coeficientes, no cumpliendo por tanto, la condición de *esparsidad;* mientras el modelo *Lasso* si seleccionaría automáticamente las variables pero a costa de incurrir en un elevado sesgo en los coeficientes β_j, no cumpliendo en este caso la propiedad de *insesgadez*.

El estimador del modelo *Elastic Net* puede resolverse minimizando la siguiente expresión matemática, que combina como vemos las penalizaciones del *Ridge Regression* y del *Lasso*:

$$\hat{\beta}^{elastic\ net} = \underset{\beta_0,\beta}{\operatorname{argmin}} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \phi_\alpha(\beta) \right\}$$

$$\phi_\alpha(\beta) = \frac{(1-\alpha)}{2} \sum_{j=1}^{p} \beta_j^2 + \alpha \sum_{j=1}^{p} |\beta_j|$$

10.3. Extensión a Modelos Lineales Generalizados

Dichas técnicas pueden extenderse como vamos a aplicar nosotros en este estudio a los Modelos Lineales Generalizados, pudiéndose aplicar a todas las distribuciones, en el que, la variable respuesta pertenezca a la familia de la de la exponencial, entre las que encontraremos las siguientes distribuciones: *logísticas*, *multinomial*, *poisson*, *gamma*, *binomial negativa* o *normal/gaussiana*.

Matemáticamente los modelos de selección de variables aplican la regularización en los Modelos Lineales Generalizados igual que con los Modelos Lineales Clásicos, cuando se resuelve el problema, que en este caso vendrá dada por la resolución de la minimización de la función de verosimilitud:

$$-\sum_{i=1}^{n} \log f\big((y_i|g(X'_i\beta)\big) + n\sum_{j=1}^{p} \phi_j(\beta_j)$$

Donde:

$\phi_j(\beta_j)$: $denota\ el\ diferente\ modelo\ de\ selección\ de\ variables\ empleado.$

11.BBDD, Variables y Herramientas Empleadas

El objetivo central de mi trabajo consiste en emplear diferentes técnicas de selección de variables con la finalidad de modelizar la frecuencia del número de siniestros RC culpa a través de la Modelización Lineal Generalizada. Para ello, dispongo de una base de datos *real* procedente de una cartera de autos que recopila información sobre un período de 3 **Rolling Year** comprendidos entre los años 2009 y 2012. A pesar de disponer de una base de datos original con más de 4.000.000 de observaciones, hemos decidido después de sumarizar registros con distintos movimientos de riesgo y filtrar la BBDD con pólizas procedentes del canal directo, quedarnos aleatoriamente con 500.000 de registros.

En dicha base de datos, se presentan unas 230 variables con las que poder trabajar, aunque tras un primer análisis hemos decidido desechar variables que no nos aportan nada, quedándonos en una primera instancia con los siguientes factores de riesgo agrupados de la siguiente manera según:

Variables Internas

Factores Relativos al Conductor

Profesión Conductor	Sexo Conductor Habitual
Edad Conductor Habitual	Sexo Conductor Ocasional
Antigüedad Vehículo	Conduce Otros Automóviles
Porcentaje Bonús	Nacionalidad
Estado Civil Conductor	Siniestros Últimos 5 Años Asegurado
Número Conductores Ocasionales	Nota Morosidad
Unidad Familiar	Cliente Recomendado
Indicador Multas	

Factores Relativos al Vehículo

Grupo Vehículo	Marca-Modelo
Marca	Peso-Vehículo
Motor	Puertas
Potencia	Categ. Tip. Clas. Vehículo
Cilindrada	Segmentación
Peso-Potencia	Velocidad Máxima
Plazas Vehículo	Categ. Clas. Vehículo & Segmentación
Valor Vehículo	Garaje
Antigüedad Vehículo	Nº de Automóviles en la Familia

Factores Relativos a la Póliza

Uso Vehículo	Número Años Compañía Anterior
Km Anuales	Forma Pago
Modalidad Póliza	Antigüedad Póliza

Variables Externas[11]

Factores Sociodemográficos

Tasa Paro	Condición Económica
Ingresos Medios	Tasa de Actividad
Porcentaje Peso Sector Terciario	Porcentaje Peso Sector no Terciario
Población Edad Media	Población Edad Mediana
Población Desvianza Típica Edad	Población Rango Intercuartilico Edad
Población con más 20 años	Población con menos 20 años
Tamaño Medio Familia	Solteros
Casados	Viudos
Separado	Divorciado
Parejas de Hecho	Hijos Medio Familia

[11] Las variables externas se encontraban en su mayor parte ya categorizadas, debido a que están referenciadas al código censal donde el cliente habita.

Total Viviendas	Total edificios
Indicación Contaminación	Áreas Mal Comunicadas
Áreas Poco Limpias	Áreas Verdes
Áreas Ruidosas	Área Pob. Española
Área Pob. Extranjera	Área Pob. Europea
Área Pob. No Europea	Estudios 3º Grado
Nivel Estudios Pob. 30-39 años	Estudios Preobligatorios
Estudios Postobligatorios	Población Residente
Indicador Juventud	Dependencia Juventud
Dependencia Senil	Tasa Autoctonía
Indicador Dependencia	Porc. Ingresos Libre de Gastos
Porc. Gastos Transporte	Nº Medio Vehículos
Porc. Estudios 2º y 3º Grado	Porc. Delincuencia
Porc. Autos Nuevos	Porc. Autos 2 Mano
Porc. Motos	Porc. Vehículos Motor
Porc. Mantenimiento y Reparación	Porc. Carburantes
Porc. Transporte Público	Porc. Seguros de Motor

Factores Meteorológicos

Altitud	Temperatura Media
Temperatura Media Máxima	Temperatura Media Mínima
Temperatura Máxima Absoluta	Temperatura Mínima Absoluta
Temperatura Inferior Máxima	Temperatura Superior Mínima
Precipitaciones Totales	Precipitaciones Máximas Diarias
Días Precipitación	Días Precipitación Superior a 10 ml
Días Lluvia	Días Nieve
Días Granizo	Días Helada
Rachas Viento	Velocidad Rachas Viento
Rachas Viento sup 55 km	Rachas Viento sup 91 km
Velocidad Media Viento	Insolación Media
Porcentaje Insolación	Presión Atmosférica
Presión Atmosférica Máxima	Presión Atmosférica Mínima
Presión Atmosférica Nivel Mar	Cuenca Nival

Zona Inundable	Días Temp. Máxima menor a 1 Grado
Días Temp. Máxima menor a 1 Grado	Congelación
Precipitaciones Abundantes	Nº Días Temp. Max. Mayor al percentil 50
Nº Días Temp. Max mayor al percentil 75	Desv. Típica Temp. Max y Min. Diaria
Nº Días Rango Temp Max y Min mayor al 5%	Nº Días Rango Temp Max y Min mayor al 95%

11.1. Sentencia Test-Achats

De las 131 factores de riesgo que hemos detallado anteriormente como posibles factores de riesgo de riesgo influyentes en la tarificación de nueva producción, dos de éstas variables no pueden ser empleadas en la modelización desde el 21 de diciembre del 2012, fecha en la que se hizo vigente la sentencia europea *Test-Achats*.

"El artículo 5 de la Directiva de la Directiva 2004/113/CE del Consejo, relativa a la aplicación del principio de igualdad entre hombres y mujeres en el acceso a bienes y servicios y su suministro regula el uso de factores actuariales basados en el sexo en la prestación de servicios financieros afines. (…).

A partir del 21 de diciembre de 2012, la norma de independencia del sexo que figura en el artículo 5, apartado 1, deberá aplicarse sin ninguna excepción para el cálculo de las primas y prestaciones a efectos de los nuevos contratos. (…)". **(Artículo 5, Directiva 2004/113/CE).**

Dicha sentencia impide usar a las entidades aseguradoras como factor de riesgo las variables de sexo al tratarse según esta sentencia de un *"problema de discriminación indirecta".*

Desde hace unos años siguiendo con la línea de la sentencia **Test-Achats** de no discriminar al cliente, se ha prohibido tarificar la nacionalidad del cliente como posible factor de riesgo para ese individuo.

11.2. Variable Respuesta y CICOS

La variable que vamos a predecir es la frecuencia de la garantía RC Material Culpable. Se trata de una variable muy analizada dentro del área gestora de autos, debido a que nos indica el grado de malos conductores que tiene una cartera. De ahí la elección de esta variable como factor a predecir. Dentro de esta garantía podemos subdividir esta garantía, entre aquellos siniestros que se resuelven a través de un convenio y los que no se encauzan a través de un convenio.

Desde el 1 de enero de 1988 en España existe el denominado *"Convenio de Indemnización Directo Español"* por el que, se intenta dar una solución ágil a todos aquellos siniestros cuyos clientes no han sido culpables, habiéndose fijado previamente un *módulo*[12] .

Con este convenio se intenta simplificar el trámite de los siniestros entre las aseguradoras, mejorando y reduciendo los tiempos de espera de las indemnizaciones a los asegurados.

El funcionamiento de este convenio es muy simple:

- Cuando se produce un siniestro, cada conductor comunica su siniestro a su compañía aseguradora.
- La compañía "inocente" comunica el siniestro al sistema CICOS.
- La compañía "culpable" puede aceptar o no el siniestro.
    - Si le acepta, pagará el modulo.
    - Si no lo acepta, deberá comunicárselo a la otra aseguradora, conllevando consigo, un proceso de resolución más largo.

Al cabo del año, una compañía puede encontrarse con dos situaciones:

❖ Tener ganancias debido a que el importe fijado en el módulo y recibido sea superior al coste real de todos los siniestros acaecidos por sus clientes sin culpa.

[12] Actualmente dicho modelo se encuentra fijado en 882€, pero anualmente dicha cuantía es susceptible a variaciones.

- Tener pérdidas debido a que el coste real de todos los siniestros acaecidos por sus clientes sin culpa sean superiores al importe fijado en el módulo.

Hoy en día para que veamos la importancia que tiene este convenio con datos, más del 70%[13] de los siniestros que se generan se resuelven a través del convenio CICOS, ahorrándose en muchos de los casos entrar en posibles litigios entre las compañías aseguradoras. Por lo tanto, se trata de un buen sistema para los clientes y un buen sistema para las aseguradoras.

Ante el gran porcentaje que representa el convenio dentro del sector asegurador, hemos decido únicamente modelizar la frecuencia y no la severidad de esta variable.

11.3. Herramientas

Tres van a ser los programas informáticos estadísticos utilizados en el estudio, aunque dos de ellos serán los softwares fundamentales en el desarrollo del trabajo: el *SAS Enterprise Guide* y el *R Studio*. Se trata de dos de los softwares más empleados por las aseguradoras en su día a día (sobre todo el primero de ellos).

El primero de ellos le emplearemos para el tratamiento previo de la base de datos original, el análisis descript ivo de los datos y la definitiva modelización de la frecuencia a través de un Modelo Lineal Generalizado. Mientras que, el R le utilizaremos especialmente en la modelización de la selección de variables, ya que disponemos de un potente repositorio de paquetes especializados en la selección de variables. También, emplearemos el R por su sencillez y desarrollo en la elaboración de gráficos.

Por último, emplearemos el software estadístico de IBM *SPSS 22,* con la única finalidad de definir los *clúster* de las variables cuantitativas procedentes de factores detallados por el cliente o por las características de los automóviles asegurados extraídos de **base SIETe**[14].

[13] Lo que en términos absolutos asciende a 2.000.000 de siniestros al año.
[14] Sistema informativo que emplean las entidades aseguradoras en la identificación y conocimiento de las características de los vehículos.

12.Análisis Exploratorio de los Datos

12.1. Variable Respuesta y Ajuste de Distribución.

Como imaginábamos, existe poca frecuencia de siniestros catalogada dentro de la garantía *RC Material Culpable*, asumiendo un porcentaje de la cartera de algo más del 2%.

Figura 12.1.1. Frecuencia de Siniestros RC Material Culpa

N_SIN_RC_MAT_CULPA	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
0	488405	97.68	488405	97.68
1	11595	2.32	500000	100.00

Gráfico elaborado por el autor.

Figura 12.1.2. Gráfico de Distribución del Número de Siniestros RC Material Culpa

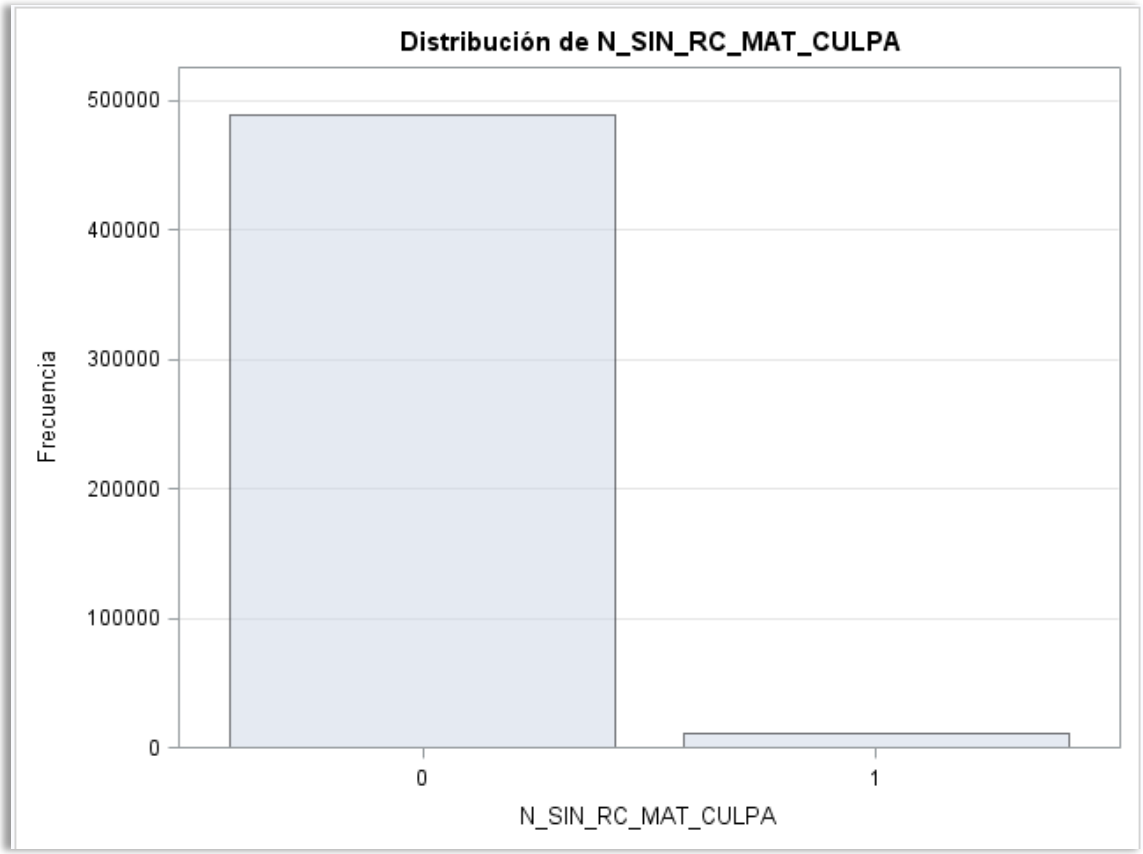


Gráfico elaborado por el autor.

Figura 12.1.3. Momentos de la variable explicada *"Número de Siniestros RC Material Culpa"*.

Momentos			
N	500000	Sumar pesos	500000
Media	0.02319	Observ suma	11595
Desviación std	0.15050671	Varianza	0.02265227
Asimetría	6.33608844	Curtosis	38.1461694
SC no corregida	11595	SC corregida	11326.112
Coef. variación	649.015559	Media error std	0.00021285

Gráfico elaborado por el autor.

Como hemos explicado anteriormente en el capítulo 6, la modelización de la frecuencia se ajusta a través de distribuciones discretas o de conteo. Vamos a considerar las cuatro distribuciones enumeradas anteriormente para modelizar la frecuencia estudiando la probabilidad teórica que mejor se ajusta a la variable *"Número de Siniestros RC Material Culpa".*

Figura 12.1.4. Probabilidad Observada y Estimada para una Distribución *"Poisson"*

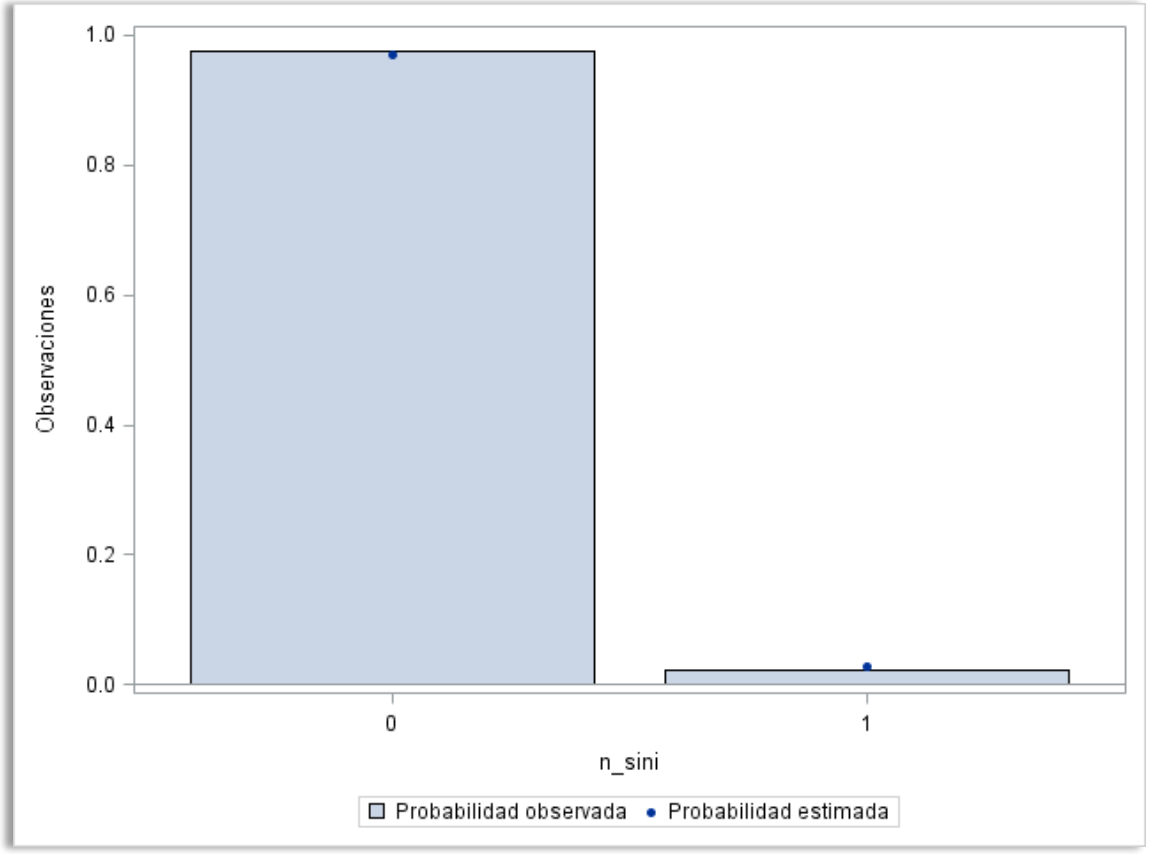


Gráfico elaborado por el autor.

Tabla 12.1.5. Probabilidad Observada y Estimada para una distribución *"Poisson"*

Número siniestros	Número obs.	Porcentaje obs.	Prob. estimada Poisson	Prob. observada
0	488405	97.681	0.97119	0.97681
1	11595	2.319	0.02835	0.02319
.	.	.	0.00045	.
.	.	.	0.00001	.
.	.	.	0.00000	.
.	.	.	0.00000	.
.	.	.	0.00000	.
.	.	.	0.00000	.
.	.	.	0.00000	.
.	.	.	0.00000	.
.	.	.	0.00000	.

Tabla elaborada por el autor.

Figura 12.1.6. Probabilidad Observada y Estimada para una Distribución *"Binomial Negativa"*

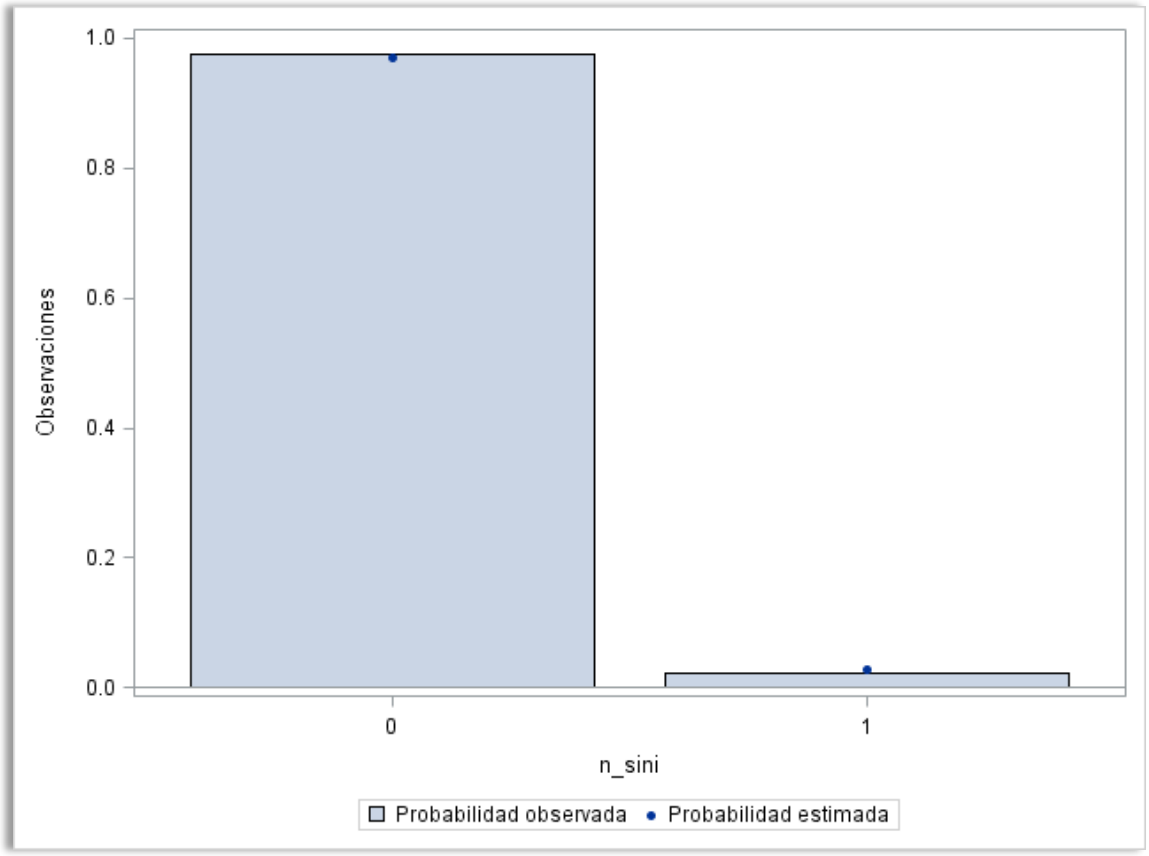


Gráfico elaborado por el autor.

Tabla 12.1.7. Probabilidad Observada y Estimada para una Distribución *"Binomial Negativa"*

Número siniestros	Número obs.	Porcentaje obs.	Prob. estimada Binomial Negativa	Prob. observada
0	488405	97.681	0.97132	0.97681
1	11595	2.319	0.02763	0.02319
.	.	.	0.00066	.
.	.	.	0.00012	.
.	.	.	0.00009	.
.	.	.	0.00006	.
.	.	.	0.00004	.
.	.	.	0.00003	.
.	.	.	0.00002	.
.	.	.	0.00001	.
.	.	.	0.00001	.

Tabla elaborado por el autor

Figura 12.1.8. Probabilidad Observada y Estimada para una Distribución *"Poisson con Exceso de Ceros"*

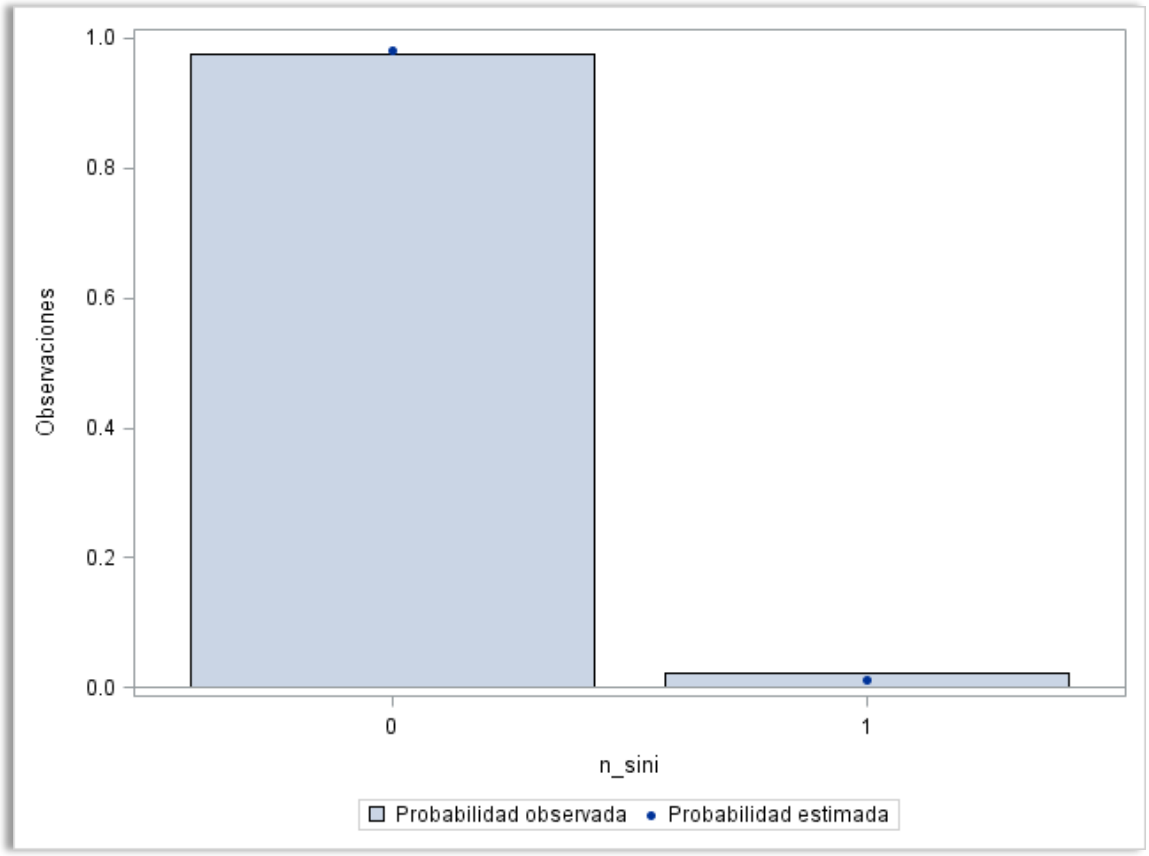


Gráfico elaborado por el autor

Tabla 12.1.9. Probabilidad Observada y Estimada para una Distribución *"Poisson con Exceso de Ceros"*

Número siniestros	Número obs.	Porcentaje obs.	Prob. estimada Poisson con exceso de ceros	Prob. observada
0	488405	97.681	0.98177	0.97681
1	11595	2.319	0.01050	0.02319
.	.	.	0.00522	.
.	.	.	0.00183	.
.	.	.	0.00051	.
.	.	.	0.00012	.
.	.	.	0.00003	.
.	.	.	0.00001	.
.	.	.	0.00000	.
.	.	.	0.00000	.
.	.	.	0.00000	.

Tabla elaborado por el autor

Tabla 12.1.10. Probabilidad Observada y Estimada para una Distribución *"Binomial Negativa con Exceso de Ceros"*

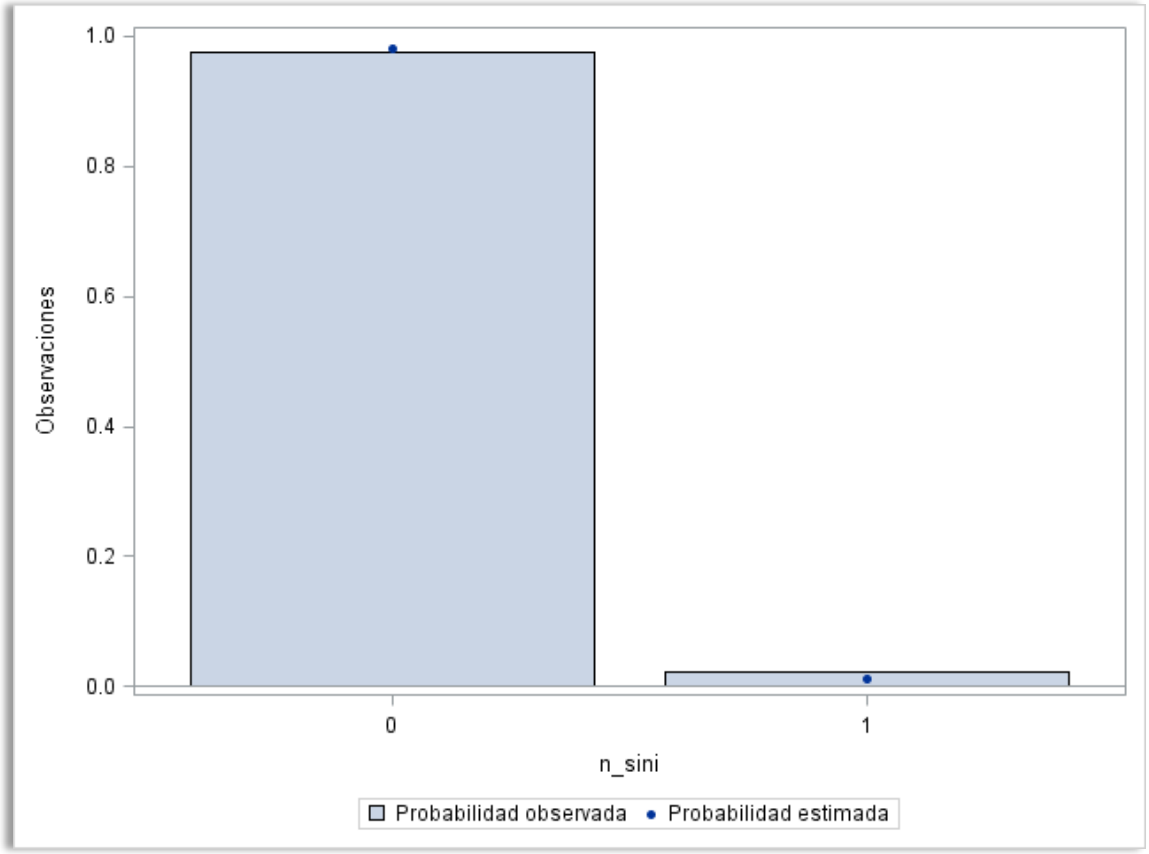


Gráfico elaborado por el autor.

Tabla 12.1.11. Probabilidad Observada y Estimada para una Distribución *"Binomial Negativa con Exceso de Ceros"*

Número siniestros	Número obs.	Porcentaje obs.	Prob. estimada Binomial Negativa con exceso de ceros	Prob. observada
0	488405	97.681	0.98177	0.97681
1	11595	2.319	0.01050	0.02319
.	.	.	0.00522	.
.	.	.	0.00183	.
.	.	.	0.00051	.
.	.	.	0.00012	.
.	.	.	0.00003	.
.	.	.	0.00001	.
.	.	.	0.00000	.
.	.	.	0.00000	.
.	.	.	0.00000	.

Tabla elaborada por el autor.

Mostrando las distribuciones teóricas de las posibles modelizaciones para la variable dependiente, vemos que cualquiera de las cuatro distribuciones analizadas pueden predecir la variable respuesta, pero debemos escoger alguna de ellas. Por ello, descartamos aquellas distribuciones infladas de cero debido a que su probabilidad estimada difiere más que en las otras dos, además de infravalorar las carteras con siniestros RC Material Culpa.

Entre las otras dos distribuciones restantes, hemos decidido modelizar a través de una Poisson ya que a pesar de que se ajusta un poco mejor la Binomial Negativa, en el mundo asegurador para modelizar la frecuencia está más convencionalizado modelar a través de una distribución Poisson y no va a existir, visto los resultados anteriores, un erróneo ajuste de distribución de la variable respuesta.

12.2. Análisis Descriptivo de las Variables Predictivas

En los tiempos en los que corren tras ver cómo ha avanzado el pricing no vida en el ramo de autos es impensable modelizar una tarifa competitiva sin realizar previamente

un *Análisis Multivariante* que permita estudiar las relaciones existentes entre las variables. Por eso, nosotros hemos realizado un exhaustivo y preciso análisis de las variables independientes que anteriormente hemos enumerado. Este análisis le consideramos vital en nuestro estudio, ya que nos ha permitido: desechar algunas variables por su baja exposición que impedía discriminar correctamente entre asegurados (ej: indicador de multas, cliente recomendado), corregir categorizaciones iniciales que habían distorsionado la esencia de la variable, obtener una intuición de aquellas variables que según el análisis gráfico podrían ser significativas en la modelización, conocer la dependencia entre variables ante un posible problema de multicolinealidad que aumente la varianza e imprecisión de las variables…

Destacar la importancia que el análisis gráfico[15] tiene sobre la tramificación de variables. Para categorizarlas, se ha empleado distintas técnicas específicas dependiendo de la variable tratada.

12.2.1. Análisis Clúster

Desde hace muchos años, en tarificación no vida se emplea métodos de *análisis de clúster,* con el objetivo de formar grupos homogéneos de individuos que presenten un riesgo similar.

En la actualidad existen dos metodologías de clúster, que se dividen según sea:

- **Una Estructura Jerárquica.**
- **Una Estructura no Jerárquica.**

La estructura jerárquica intenta agrupar observaciones de una variable según algoritmos basados en una matriz de distancias mientras que la estructura no jerárquica agrupa sus observaciones optimizando sus datos según el objetivo funcional fijado a priori, como puede ser haber fijado de antemano el número de clúster.

[15] El capítulo del **análisis gráfico univariante y bivariante** de las variables predictivas va a ser expuesto en el apéndice debido a la abundancia de variables de la que disponemos y la materialidad del estudio.

Dentro de la estructura jerárquica existen dos tipos de técnicas:

- **Técnicas Aglomerativas.**
- **Técnicas Divisivas.**

La diferencia esencial entre las técnicas aglomerativas y las divisivas es que las primeras parten de una escenario sin conglomerados, y paulatinamente van produciéndose los clúster entre sí, haciéndose cada vez más mayores a través de subclústers; mientras que las técnicas divisivas parten de un escenario opuesto, en el que, todas las observaciones en principio se encuentran dentro de un conglomerado y a medida que se va desarrollando el algoritmo, el conjunto inicial de datos conformados en un clúster van subdividiéndose en nuevas aglomeraciones más particionadas.

En nuestro estudio, nosotros vamos a emplear una técnica de segmentación divisiva como es la **CHAID** (*Chi-Square Automatic Interaction Detection*). Su intención es dividir las observaciones en subgrupos creando así variables categóricas a través de un algoritmo de diagrama de árbol que maximice la diferencia entre cada clúster y reduzca la parte de varianza no explicada midiéndolo a través de la Chi-Cuadrado. Como hemos explicado anteriormente en el capítulo 9, esta metodología a través de **"Árboles de Decisión"** se enmarca dentro de la teoría del Aprendizaje Automático.

Se trata de un método muy intuitivo y correcto estadísticamente al seleccionar los grupos más homogéneos, presentando cada grupo unos valores similares respecto a la variable respuesta. Con esta técnica, conseguiremos categorizar correctamente los diferentes colectivos según su riesgo.

Con esta metodología hemos categorizado todas las variables procedentes de Base SIETe referentes a las características de los automóviles, al igual que aquellas variables propias de los individuos como puedan ser la edad, número de vehículos por familia…

En cambio, las variables externas ya se encontraban categorizadas de la base de datos original, por lo que, únicamente en algunas variables se ha producido pequeñas recategorizaciones con la finalidad de disponer de grupos con mayor exposición que no se vean tan afectados por problemas de varianza.

12.2.2. Grado de Asociación entre Variables

Como hemos remarcado al principio de este capítulo es muy importante cuantificar el grado de dependencia entre las variables para no concebir problemas de multicolinealidad durante las estimaciones.

Al encontrarse todos los factores de riesgo tramificadas en diferentes niveles de riesgo, se debe emplear medidas de asociación destinadas a variables categóricas nominales. Entre las medidas disponibles para analizar la asociación entre variables hemos escogido la *"V-Crammer"* por su sencillez a la hora de calcular e interpretar resultados.

Esta medida no deja ser una corrección del coeficiente Chi-Cuadrado, que permite medir a través de un índice el grado de asociación entre variables. Su formulación matemática es la siguiente:

$$\boldsymbol{\nu} = \sqrt{\frac{\boldsymbol{\chi^2}}{\boldsymbol{N} * \boldsymbol{m}}}\ , \quad \mathbf{0} \leq \boldsymbol{\nu} \leq \mathbf{1}$$

Donde:

N: Número total de Observaciones en la Tabla.

m: $\min(f-1, c-1)$

A continuación vamos a mostrar una tabla con una selección de variables que presentan un elevado grado de asociación (cercano al 1).

Figura 12.2.2.1. Grado de asociación de variables a través de la V-Crammer

Variable 1	Variable 2	V-Crammer
Porc. Pob. Española	Porc. Pob. Extranj	0.98678635
Presión Media	Presión Abs. Max	0.94358716
Temperatura Mínima	Temp. Media Min.	0.93515022
Presión Media	Presión Abs. Min	0.88074071
Presión Abs. Max	Presión Abs. Min	0.84111777
Precip. Totales	Días Precip. Más 10 ml	0.77662251
Edad Conductor Ocas	Antig. Carnet Ocas	0.76044476
Congelación	Cuenca Nival	0.75557298
Días Precipitación	Precip. Totales	0.74567154

Porc. Insolación	Precip. Totales	0.74345183
Cuenca Nival	Temperaturas Media	0.74133319
Velocidad Media Viento	Cuenca Nival	0.72777997
Temp. Media Min	Cuenca Nival	0.72766709
Temp. Media Max	Porc. Insolación	0.72718757
dias_granizos222	Cuenca Nival	0.72613681
Rachas Viento Mas 91 km	Cuenca Nival	0.72499923
Temperatura Mínima	Cuenca Nival	0.7231489
Zona Inundable	precipit_totales222	0.72292065
Precipitaciones Totales	Cuenca Nival	0.71486891
Cuenca Nival	Rachas Viento Mas 55 km	0.71449665
Temp. Inf. Min.	Zona Inundable	0.71435105
Precipitaciones Máximas	Cuenca Nival	0.71409986
Rural Urbano 5%	Zona Inundable	0.71390342
Días Precipitación	Cuenca Nival	0.71345726
Zona Inundable	Ingresos Medios	0.71273952
Zona Inundable	Peso Activ. No Terciario	0.71190427
Días Precipitación	Temperaturas Media Max	0.71182676
Rural Urbano 5%	Cuenca Nival	0.71082872
Zona Inundable	Tasa Paro	0.70740121
Días Precipitación	Temp. Inferior Max	0.69687879
Temp. Media Min.	Temp. Inferior Max	0.6859432
Temp. Media Max.	Días Lluvia	0.68550166
Temperaturas Medías	Días Precip. Más 10 ml	0.68181689
Días Precipitación	Temp. Media Min.	0.67837289
Velocidad Media Viento	Porc. Insolación	0.67770012
Días Precipitación	Porc. Insolación	0.67708459
Días Granizo	dias_precipit_sup10ml22	0.67237389
Porc. Insolación	Pres. Atmosfér. Niv. Mar	0.67193573

Tabla elaborada por el autor.

Este análisis nos ha servido a lo largo del trabajo para cerciorarnos de la alta correlación que presentan las variables externas entre sí, algo por otra parte totalmente lógico. Esto provocará que el modelo de regresión final no contenga muchas variables externas sin incurrir en multicolinealidad y aumentar consecuentemente la varianza de la regresión.

13. Modelización Variables Internas

13.1. Ridge Regression

Vamos a poner en práctica el objetivo de dicho estudio, mostrando en ella las diferentes técnicas de regularización y selección de variables explicadas durante el capítulo 10. Para ello, hemos empleado el paquete glmnet del software estadístico R para poder llevar a cabo la regularización y selección variables.

Empezamos analizando el modelo de penalización más primitivo de todos, el modelo *Ridge Regression*, asumiendo que nuestra variable respuesta presenta una **estructura de link logarítmica**.

Con el objetivo de identificar el λ^{opt}, el primer paso será realizar la **validación cruzada**[16]**,** mostrando con ello la verosimilitud de cada λ evaluada a través de múltiples particiones a través de la función cv.glmnet en R.

Figura 13.1.1. Validación Cruzada modelización Ridge Regression

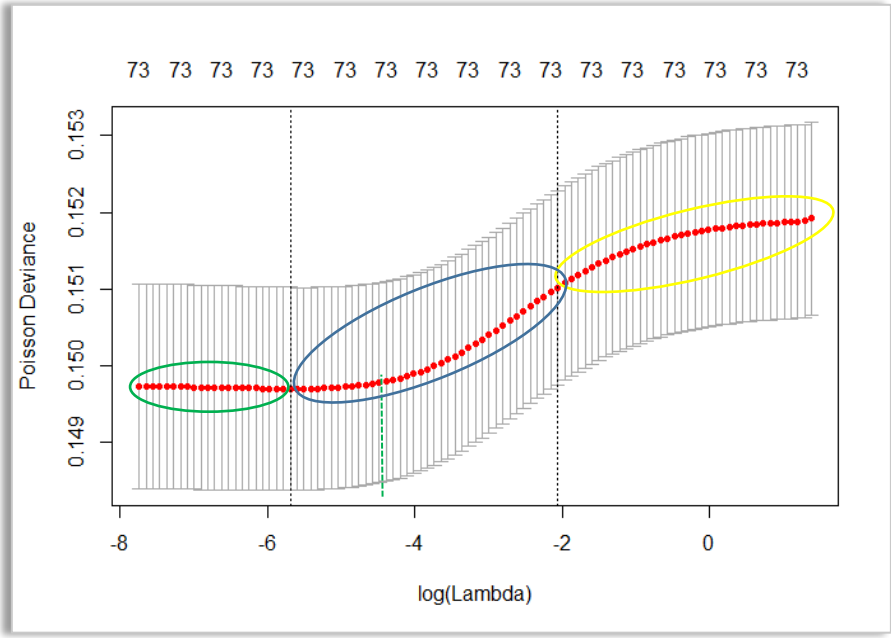


Tabla elaborada por el autor.

[16] Para llevar a cabo la validación cruzada con la secuencia de lambdas empleo la función cv.glmnet().

En dicha gráfica se puede identificar perfectamente los cambios de tendencia que se produce con cada partición.

Cuando se aplica poca penalización ($\lambda\ pequeña$), la Poisson Deviance decrece hasta llegar a su mínimo global (línea serpenteada situada más a la izquierda), marcando esta línea el valor óptimo de la regresión. Posteriormente medida que se penaliza más el modelo, la Poisson Deviance crece[17] paulatinamente, aumentando su ritmo de crecimiento a partir de una $\log(\lambda) \leq -4.3$ aproximadamente.

La segunda línea serpenteada (situada a la derecha de la figura) corresponde con el valor del λ más una desviación típica. Con esta λ se desea aplicar un modelo *ridge regression* muy penalizado que distorsionará bastante los coeficientes resultantes de un modelo GLM. A partir de este punto, la Deviance crece más lenta consecuencia de las escasas variables relevantes con esos niveles de λ (área redondeada de amarillo).

De todo el proceso de validación cruzada extraído, para la modelización *Ridge Regression* voy a estudiar el que contenga la menor Deviance (punto óptimo), que en este caso equivaldrá a un λ^{opt} de 0.003387337, o lo que es lo mismo, a un $log(\lambda) = -5.68771121,$ situándose su Poisson Deviance en un valor de 0.1496993.

Comparando los modelos Ridge Regression y GLM en una tabla, podemos apreciar la penalización aplicada a cada variable dummy con el λ^{opt} obtenido en la selección cruzada:

Tabla 13.1.2. Coeficientes Ridge Regression vs GLM

C. Num	Variables	Dummy	Coef. RR	Coef. GLM	Dif. %
1	Intercepto		-4.2418	-4.4131	-3.88%
2	Nº años cia anter	<4	0.0455	0.0333	36.64%
3	Nº años cia anter	Indeterminado	0.2202	0.2944	-25.20%
4	Potencia	<=68	-0.0251	-0.1251	-79.94%
5	Potencia	(68;89]	-0.0434	-0.0623	-30.34%
6	Potencia	(89;99]	-0.0060	0.0092	-165.22%
7	Potencia	>122	-0.0093	-0.0013	615.38%
8	Cilindrada	<=1351	-0.0526	-0.0262	100.76%
9	Cilindrada	(1868;1896]	0.0552	0.0568	-2.82%
10	Cilindrada	(1896;2188]	0.0376	0.0379	-0.79%
11	Cilindrada	>2188	0.1089	0.1324	-17.75%
12	Peso Potencia	<=10.47	-0.0570	-0.0886	-35.67%

[17] Área circulada en color azul.

13	Peso Potencia	>13.44	0.0631	0.0889	-29.02%
14	Valor Vehículo	<=11450	-0.0436	-0.0064	581.25%
15	Valor Vehículo	(11450;15485.50]	-0.0263	-0.0133	97.74%
16	Valor Vehículo	(13370.85;17565]	0.0611	0.0795	-23.14%
17	Ant. Vehículo	(12;38]	-0.0691	-0.0886	-22.01%
18	Ant. Vehículo	9999	0.4181	0.2522	65.78%
19	Peso Vehículo	<=979.92	-0.0594	-0.0469	26.65%
20	Peso Vehículo	>1619.76	0.0404	0.0666	-39.34%
21	Velocidad	<=167	-0.0112	-0.0123	-8.94%
22	Velocidad	(187;209]	-0.0479	-0.0364	31.59%
23	Velocidad	>209	-0.0758	-0.0824	-8.01%
24	E. Cond. Hab.	[18;28]	0.2280	0.1365	67.03%
25	E. Cond. Hab.	(53;74]	0.0749	0.1166	-35.76%
26	Ant. Carn. Hab.	[0;10]	0.0168	0.0093	80.65%
27	Ant. Carn. Hab.	[23;25]	-0.0948	-0.1447	-34.49%
28	Ant. Carn. Hab.	(25;78]	-0.0390	-0.0749	-47.93%
29	E. Cond. Ocas.	[18;30]	0.2083	0.5984	-65.19%
30	E. Cond. Ocas.	(30;53)	-0.1091	0.3424	-131.86%
31	E. Cond. Ocas.	(53;87]	-0.0330	0.3985	-108.28%
32	Ant. Carn. Ocas.	[0;7]	0.1885	-0.2337	-180.66%
33	Ant. Carn. Ocas.	[8;38]	-0.1016	-0.5694	-82.16%
34	Ant. Carn. Ocas.	[39;57]	0.3798	0	.
35	Modalidad	Terc. Básicos	0.0492	0.0719	-31.57%
36	Modalidad	Terc. Ampliados	0.0152	0.0174	-12.64%
37	Modalidad	TRSF	0.0252	-0.0161	-256.52%
38	Forma de Pago	Semestral	0.1364	0.1492	-8.58%
39	Forma de Pago	Trimestral-Mensual	0.1855	0.1737	6.79%
40	Uso Vehículo	Trab y Uso Prof	0.0174	-0.0038	-557.89%
41	Km Anuales	Hasta 5000 Km	-0.0300	-0.0323	-7.12%
42	Km Anuales	De 10001 a 15000 Km	0.0691	0.0248	178.63%
43	Km Anuales	Desde 15001 Km	0.0140	0.0167	-16.17%
44	Motor	Gasolina	-0.0615	-0.0378	62.70%
45	Motor	Otros/Desconocido	0.1040	-0.0804	-229.35%
46	Plazas	>5	0.0350	0.0120	191.67%
47	Plazas	Otros	-1.6019	-15.2500	-89.50%
48	Bonificación	Bonus Bajo	-0.5002	0.3262	-253.34%
49	Bonificación	Bonus Medio	-0.2705	-0.5530	-51.08%
50	Puertas	3	-0.0424	-0.0606	-30.03%
51	Puertas	Indefinido	-1.1236	-14.4756	-92.24%
52	Unidad Familiar	Sí	0.0968	0.0596	62.42%
53	Nº Auto Familia	0	0.2897	-0.0871	-432.61%
54	Nº Auto Familia	Más de 2	0.0725	0.0749	-3.20%
55	Nº Auto Familia	Indefinido	0	-14.6957	-100.00%
56	Cond. Otros Veh.	Sí	-0.1116	-0.0619	80.29%
57	Morosidad	A-B	-0.0977	-0.0837	16.73%
58	Morosidad	D-E	0.0383	0.0286	33.92%
59	Morosidad	F	0.2635	0.2614	0.80%
60	Morosidad	Desconocido	0.1582	0.1582	0.00%
61	Grupo Marca	1	0.0589	0.0602	-2.16%

62	Grupo Marca	3	-0.1363	-0.1228	10.99%
63	Grupo Marca	4	0.1720	0.1394	23.39%
64	Grupo Marca	5	0.2513	0.2556	-1.68%
65	Grupo Veh.	Tur. Famil/Monov	-0.0010	-0.0309	-96.76%
66	Grupo Veh.	Todot./Veh. Sobrep.	0.0978	0.0584	67.47%
67	Grupo Veh.	Otros Grupos Vehículos	-0.0490	-0.0594	-17.51%
68	Estado Civil	Divorciado/Separado	0.0763	0.1246	-38.76%
69	Estado Civil	Soltero	-0.0434	-0.0740	-41.35%
70	Estado Civil	Otros	0.0050	0.1434	-96.51%
71	Garaje	Garaje Individual	0.0214	0.0259	-17.37%
72	Garaje	Sin Garaje	0.0429	0.0923	-53.52%
73	Profesión	Grupo 1	-0.0213	-0.0279	-23.66%
74	Profesión	Grupo 2	-0.0321	-0.0299	7.36%
75	Profesión	Grupo 4	0.0616	0.0198	211.11%

Tabla elaborada por el autor.

Como vemos, el modelo Ridge Regression regulariza las variables con el objetivo de disminuir la varianza que estas presentan (ya sea consecuencia de multicolinealidad, poca exposición de las variables…) y consecuentemente penalizar los coeficientes de las variables. Es decir, que no todas las variables se regularizan por igual según la λ^{opt}.

Gráfico 13.1.3. Regularización de las variables modelizadas a través de Ridge Regression

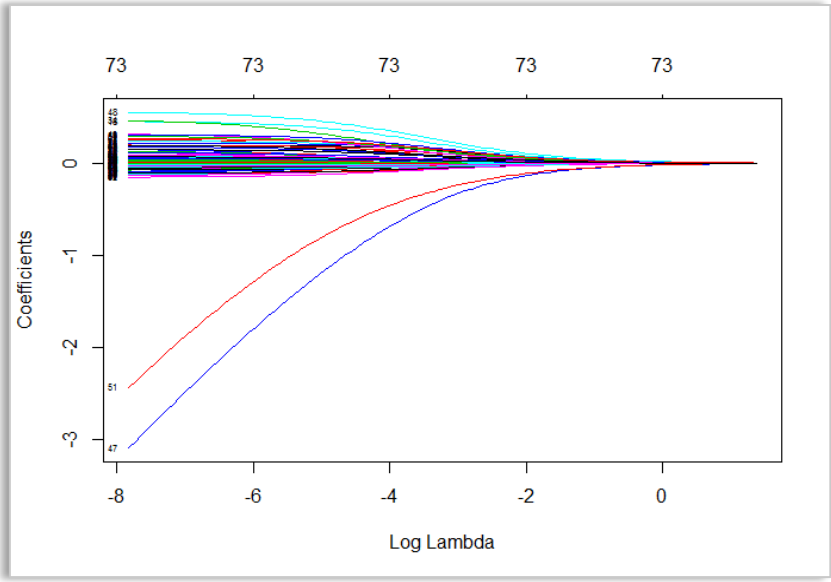


Gráfico elaborado por el autor

Como vemos en el gráfico anterior, todas las variables convergen a 0 a medida que la λ crece, no existiendo una selección de variables.

El siguiente gráfico tiene una explicación en la misma línea. En este caso, muestra como a medida que se va aumentando las λ, los coeficientes de las variables disminuyen y la cantidad de Deviance que en conjunto todas las variables son capaces de explicar también disminuyen.

Figura 13.1.4. Cantidad de Deviance explicada por el conjunto de todas variable modelización *Ridge Regression*.

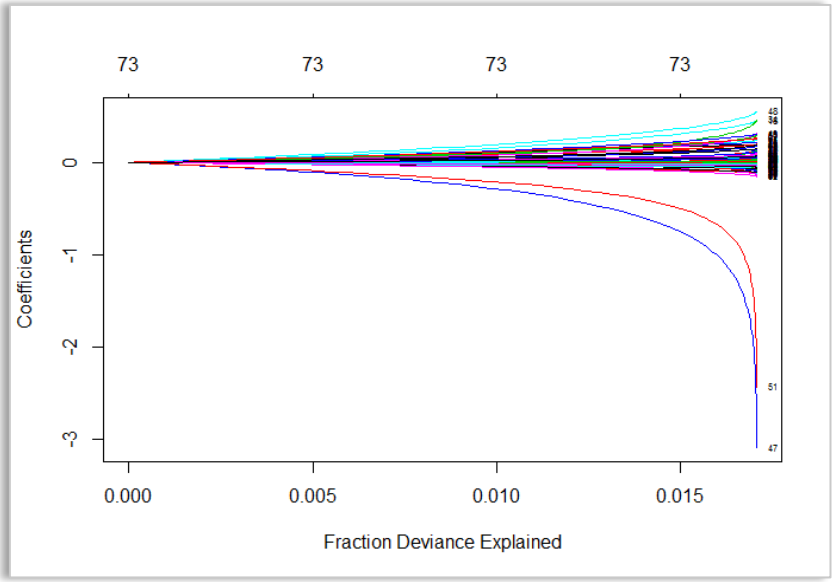
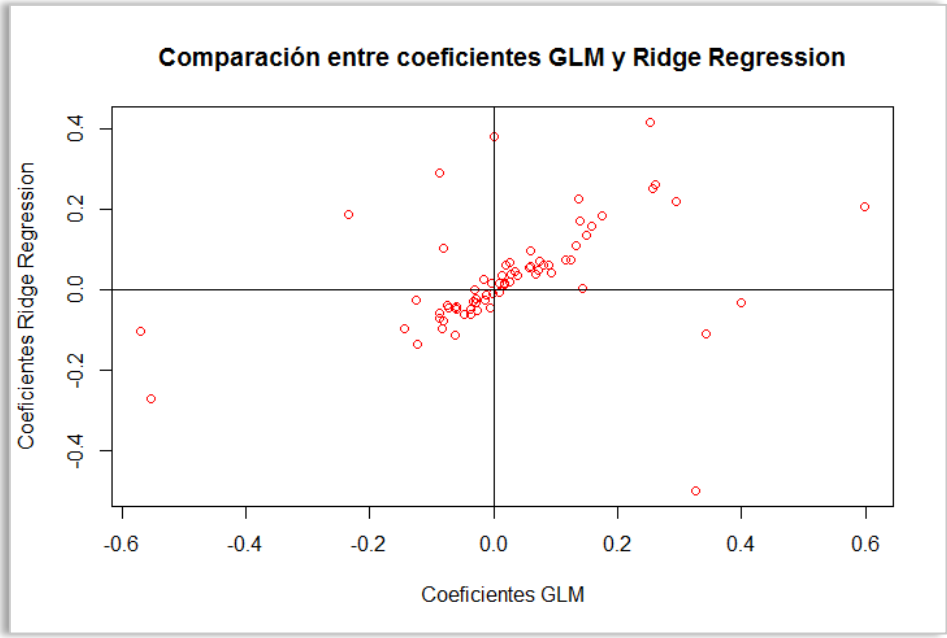


Gráfico elaborado por el autor.

Si hacemos una comparativa a través de un gráfico de dispersión entre los coeficientes modelizados a través de un GLM y *Ridge Regression*, eliminando aquellas variables con unos coeficientes muy elevados que distorsionen cualquier análisis (ya sea, procedentes de una baja exposición o tras haber eliminado el intercepto que marca la relación entre la variable dependiente y el origen).

Figura 13.1.5. Gráfica de Dispersión comparativa de coeficientes GLM y *Rigde Regression*



Gráfica elaborada por el autor.

En dicho gráfico podemos ver como las variables más alejadas del origen tendrán en los próximos modelos una mayor relevancia en el estudio al tratarse de verdaderas variables discriminantes siendo seleccionadas en los modelos de espasividad posteriores. Esto no quiere decir que inmediatamente todas aquellas variables que se encuentran cercanas al origen no sean seleccionadas en próximos modelos, pero con seguridad no serán variables tan relevantes como la mayoría de estas y en su mayoría no se trataran de variables significativas/seleccionadas.

13.2. Lasso

A continuación, vamos a detallar los resultados obtenidos a través del modelo de selección *Lasso*. Para ello al igual que hicimos para el modelo *Ridge Regression*, lo primero que vamos a hacer es analizar a través de la validación cruzada cual sería el λ^{opt}.

Figura 13.2.1. Validación Cruzada modelización *Lasso*

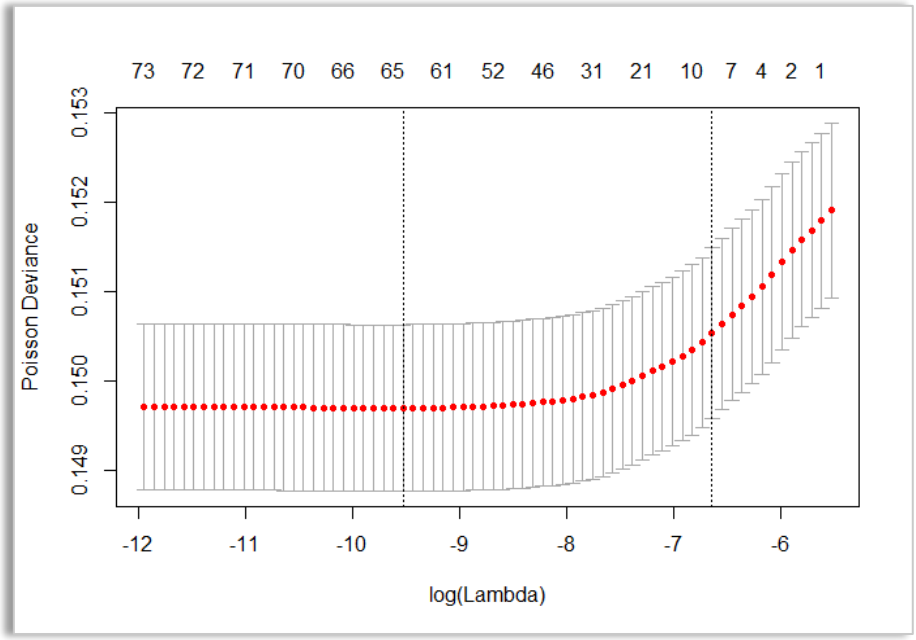


Gráfico elaborado por el autor.

Como es lógico, la validación cruzada para el modelo Lasso presenta una mayor pendiente y tendencia que el modelo Ridge Regression respecto a la Poisson Deviance. Hasta que no seleccionamos 64 de las 75 posibles variables, la Poisson Deviance se mantiene prácticamente constante. Desde la primera línea serpenteada, la tendencia de la Poisson Deviance cambia aunque cuando realmente aumenta es a partir de $\log(\lambda) = -8$. A partir de esa penalización, la aplicación de una λ más restrictiva provoca que esa menor selección de variables incremente la Poisson Deviance.

Los λ para los modelos *Shrinkage* no son comparables entre sí. Los λ de los modelos de esparsividad son más pequeños que los modelos *Shrinkage* únicamente regularizadores.

El λ^{opt} que indica la validación cruzada como óptimo corresponde a un λ de 0.00007297796, o lo que es lo mismo, un $log(\lambda) = -9.52535308$, situándonos en una Poisson Deviance de 0.1497042.

A continuación, mostraremos los coeficientes del modelo *Lasso* haciendo una comparación en la misma tabla con los modelos Ridge Regression y GLM, habiendo visto previamente en el gráfico de validación cruzada que la esparsividad de variables va a ser muy pequeña (únicamente 11 variables son desechadas).

Figura 13.2.2. Coeficientes *Lasso*, *Ridge Regression* y *GLM*.

C. Num	Variables	Dummy	Coef. Lasso	Coef. RR	Coef. GLM	Dif % L-RR	Dif % L-GLM
1	Intercepto		-4.2658	-4.2418	-4.4131	0.57%	-3.34%
2	Nº años cia anter	<4	0.03904	0.0455	0.0333	-14.20%	17.24%
3	Nº años cia anter	Indeterminado	0.1629	0.2202	0.2944	-26.02%	-44.67%
4	Potencia	<=68	-0.0209	-0.0251	-0.1251	-16.73%	-83.29%
5	Potencia	(68;89]	-0.0412	-0.0434	-0.0623	-5.07%	-33.87%
6	Potencia	(89;99]	.	-0.0060	0.0092	.	.
7	Potencia	>122	.	-0.0093	-0.0013	.	.
8	Cilindrada	<=1351	-0.0517	-0.0526	-0.0262	-1.71%	97.33%
9	Cilindrada	(1868;1896]	0.0531	0.0552	0.0568	-3.80%	-6.51%
10	Cilindrada	(1896;2188]	0.0365	0.0376	0.0379	-2.93%	-3.69%
11	Cilindrada	>2188	0.1231	0.1089	0.1324	13.04%	-7.02%
12	Peso Potencia	<=10.47	-0.0612	-0.0570	-0.0886	7.37%	-30.93%
13	Peso Potencia	>13.44	0.0623	0.0631	0.0889	-1.27%	-29.92%
14	Valor Vehículo	<=11450	-0.0273	-0.0436	-0.0064	-37.39%	326.56%
15	Valor Vehículo	(11450;15485.50]	-0.0149	-0.0263	-0.0133	-43.35%	12.03%
16	Valor Vehículo	(13370.85;17565]	0.0654	0.0611	0.0795	7.04%	-17.74%
17	Ant. Vehículo	(12;38]	-0.0691	-0.0691	-0.0886	0.00%	-22.01%
18	Ant. Vehículo	9999	0.2699	0.4181	0.2522	-35.45%	7.02%
19	Peso Vehículo	<=979.92	-0.0627	-0.0594	-0.0469	5.56%	33.69%
20	Peso Vehículo	>1619.76	0.0187	0.0404	0.0666	-53.71%	-71.92%
21	Velocidad	<=167	-0.0072	-0.0112	-0.0123	-35.71%	-41.46%
22	Velocidad	(187;209]	-0.0438	-0.0479	-0.0364	-8.56%	20.33%
23	Velocidad	>209	-0.0787	-0.0758	-0.0824	3.83%	-4.49%
24	E. Cond. Hab.	[18;28]	0.1983	0.2280	0.1365	-13.03%	45.27%
25	E. Cond. Hab.	(53;74]	0.0878	0.0749	0.1166	17.22%	-24.70%
26	Ant. Carn. Hab.	[0;10]	.	0.0168	0.0093	.	.
27	Ant. Carn. Hab.	[23;25]	-0.1022	-0.0948	-0.1447	7.81%	-29.37%
28	Ant. Carn. Hab.	(25;78]	-0.0450	-0.0390	-0.0749	15.38%	-39.92%
29	E. Cond. Ocas.	[18;30]	0.2979	0.2083	0.5984	43.01%	-50.22%
30	E. Cond. Ocas.	(30;53)	-0.0235	-0.1091	0.3424	-78.46%	-106.86%
31	E. Cond. Ocas.	(53;87]	.	-0.0330	0.3985	.	.
32	Ant. Carn. Ocas.	[0;7]	0.0886	0.1885	-0.2337	-53.00%	-137.91%
33	Ant. Carn. Ocas.	[8;38]	-0.1729	-0.1016	-0.5694	70.18%	-69.63%
34	Ant. Carn. Ocas.	[39;57]	0.2974	0.3798	0	-21.70%	.
35	Modalidad	Terc. Básicos	0.0363	0.0492	0.0719	-26.22%	-49.51%
36	Modalidad	Terc. Ampliados	0.0010	0.0152	0.0174	-93.42%	-94.25%
37	Modalidad	TRSF	0.0125	0.0252	-0.0161	-50.40%	-177.64%
38	Forma de Pago	Semestral	0.1452	0.1364	0.1492	6.45%	-2.68%
39	Forma de Pago	Trimestral-Mensual	0.1864	0.1855	0.1737	0.49%	7.31%
40	Uso Vehículo	Trab y Uso Prof	0.0130	0.0174	-0.0038	-25.29%	-442.11%
41	Km Anuales	Hasta 5000 Km	-0.0272	-0.0300	-0.0323	-9.33%	-15.79%
42	Km Anuales	De 10001 a 15000 Km	0.0725	0.0691	0.0248	4.92%	192.34%
43	Km Anuales	Desde 15001 Km	0.0050	0.0140	0.0167	-64.29%	-70.06%
44	Motor	Gasolina	-0.0717	-0.0615	-0.0378	16.59%	89.68%
45	Motor	Otros/Desconocido	0.0304	0.1040	-0.0804	-70.77%	-137.81%

46	Plazas	>5	0.0324	0.0350	0.0120	-7.43%	170.00%
47	Plazas	Otros	.	-1.6019	-15.2500	.	.
48	Bonificación	Bonus Bajo	0.5569	-0.5002	0.3262	-211.34%	70.72%
49	Bonificación	Bonus Medio	0.3134	-0.2705	-0.5530	-215.86%	-156.67%
50	Puertas	3	-0.0458	-0.0424	-0.0606	8.02%	-24.42%
51	Puertas	Indefinido	.	-1.1236	-14.4756	.	.
52	Unidad Familiar	Sí	0.1183	0.0968	0.0596	22.21%	98.49%
53	Nº Auto Familia	0	0.3070	0.2897	-0.0871	5.97%	-452.47%
54	Nº Auto Familia	Más de 2	0.0693	0.0725	0.0749	-4.41%	-7.48%
55	Nº Auto Familia	Indefinido	.	0	-14.6957	.	.
56	Cond. Otros Veh.	Sí	-0.1233	-0.1116	-0.0619	10.48%	99.19%
57	Morosidad	A-B	-0.0954	-0.0977	-0.0837	-2.35%	13.98%
58	Morosidad	D-E	0.0441	0.0383	0.0286	15.14%	54.20%
59	Morosidad	F	0.2845	0.2635	0.2614	7.97%	8.84%
60	Morosidad	Desconocido	0.1696	0.1582	0.1582	7.21%	7.21%
61	Grupo Marca	1	0.0615	0.0589	0.0602	4.41%	2.16%
62	Grupo Marca	3	-0.1330	-0.1363	-0.1228	-2.42%	8.31%
63	Grupo Marca	4	0.1789	0.1720	0.1394	4.01%	28.34%
64	Grupo Marca	5	0.2558	0.2513	0.2556	1.79%	0.08%
65	Grupo Veh.	Tur. Famil/Monov	.	-0.0010	-0.0309	.	.
66	Grupo Veh.	Todot./Veh. Sobrep.	0.1095	0.0978	0.0584	11.96%	87.50%
67	Grupo Veh.	Otros Grupos Vehículos	-0.0246	-0.0490	-0.0594	-49.80%	-58.59%
68	Estado Civil	Divorciado/Separado	0.0638	0.0763	0.1246	-16.38%	-48.80%
69	Estado Civil	Soltero	-0.0397	-0.0434	-0.0740	-8.53%	-46.35%
70	Estado Civil	Otros	.	0.0050	0.1434	.	.
71	Garaje	Garaje Individual	0.0160	0.0214	0.0259	-25.23%	-38.22%
72	Garaje	Sin Garaje	0.0356	0.0429	0.0923	-17.02%	-61.43%
73	Profesión	Grupo 1	-0.0108	-0.0213	-0.0279	-49.30%	-61.29%
74	Profesión	Grupo 2	-0.0247	-0.0321	-0.0299	-23.05%	-17.39%
75	Profesión	Grupo 4	0.0584	0.0616	0.0198	-5.19%	194.95%

Tabla elaborada por el autor.

A pesar de tratarse de un λ^{opt} bastante pequeño y poco relevante, se puede intuir viendo el comportamiento de los coeficientes, las variables que no serán seleccionadas si aumentamos un poco el *tunning* en la modelización *Elastic Net*. Todo hace indicar que variables como: la **profesión**, el **uso del vehículo**, la **edad del conductor ocasional,** no serán entre otras variables si decimos aumentar el *tunning*.

Por lo contrario, también podemos intuir que variables serán significativas en el modelo de regresión de variables internas definitivo. Hemos apreciado que varias variables cuando se produce la regularización aumentan sus coeficientes, llevándonos a la lógica que si estas variables han reducido su varianza y aumentan sus estimadores serán variables significativas. Entre estas variables que hemos detectado con el

aprendizaje, podemos enumerar: el **intercepto**, la **morosidad**, la **bonificación**, la **unidad familiar**, la **forma de pago** y el **peso vehículo**.

En el siguiente gráfico veremos la regularización y selección de variables dependiendo del λ seleccionado durante el proceso de validación cruzada.

Gráfico 13.2.3. Regularización de las variables modelizadas a través de *Lasso.*

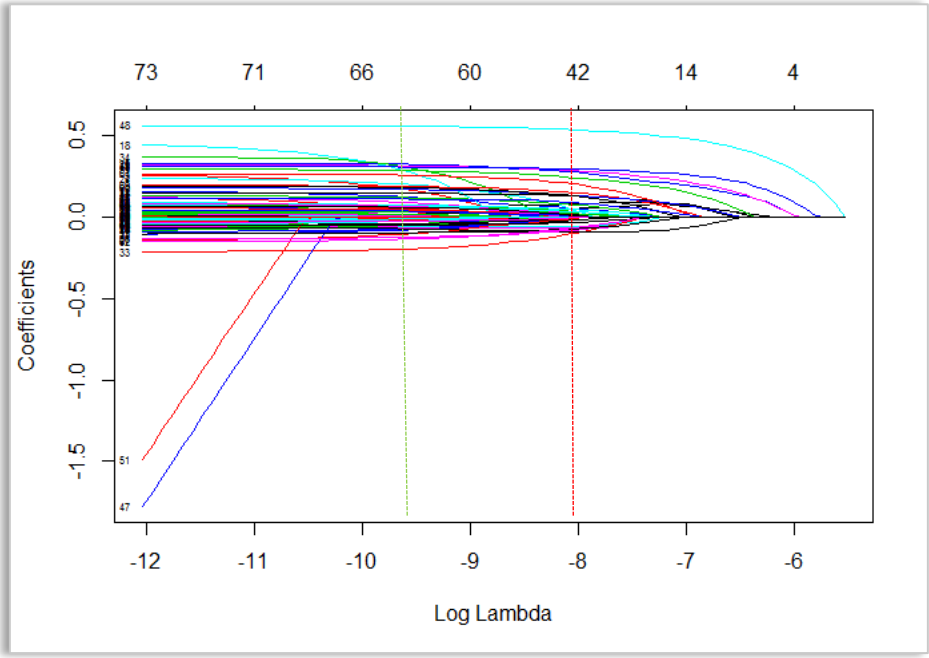


Gráfico elaborado por el autor.

La línea serpenteada verde indica aproximadamente donde se encuentra el λ^{opt} mientras que la línea serpenteada roja indica aproximadamente un $\log(\lambda)$ de 8. Como vemos, la selección de variables en ese tramo comprendido es más intensa desechándose aproximadamente 20 variables. Se aprecia también como las variables 47 y 51 (correspondientes a: puertas indefinidas y plazas indefinidas) presentan unos coeficientes muy elevados por su baja exposición pero serán eliminadas con la introducción de un pequeño tunning[18].

[18] Si habría un menor número de variables, la nitidez de la regularización de las variables sería más clara.

El gráfico siguiente va muy en línea de lo explicado en el párrafo anterior pero enmarcado dentro de la cantidad de Deviance explicada por el conjunto cada variable según el λ y estimador en cada tramo de la validación cruzada.

Figura 13.2.4. Cantidad de Deviance explicada por el conjunto de todas variable modelización *Lasso*.

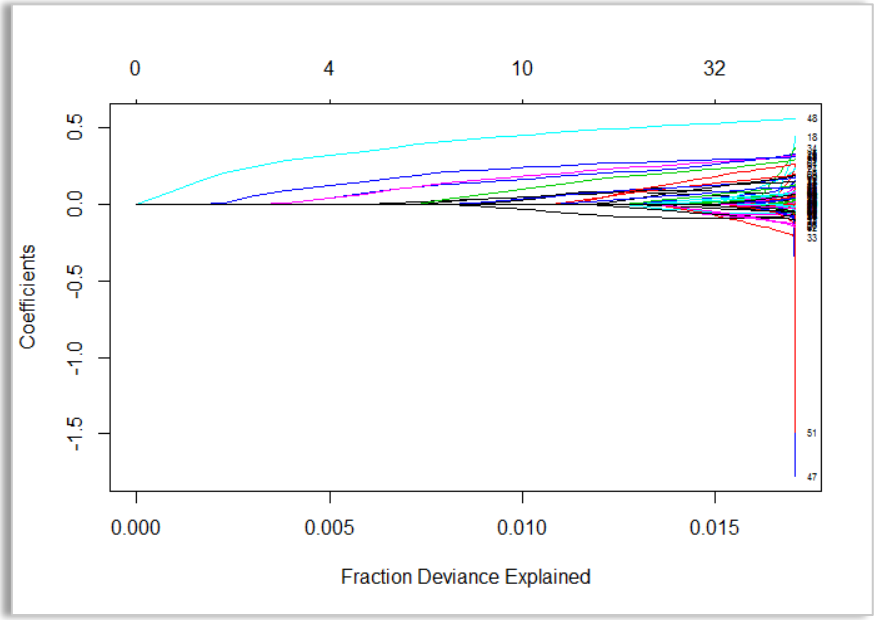


Gráfico elaborado por el autor.

Igual que para la modelización *Ridge Regression*, voy a comparar a través de un gráfico de dispersión los estimadores resultantes a través del modelo *Lasso* y *GLM.* Eso sí, antes de nada voy a eliminar el intercepto al distorsionar el gráfico[19].

[19] En este caso, no nos ha hecho falta eliminar del gráfico todas aquellas variables con poca exposición al haberse desechado en su mayoría en la modelización *Lasso*.

Figura 13.2.5. Gráfica de Dispersión comparativa de coeficientes GLM y *Lasso.*

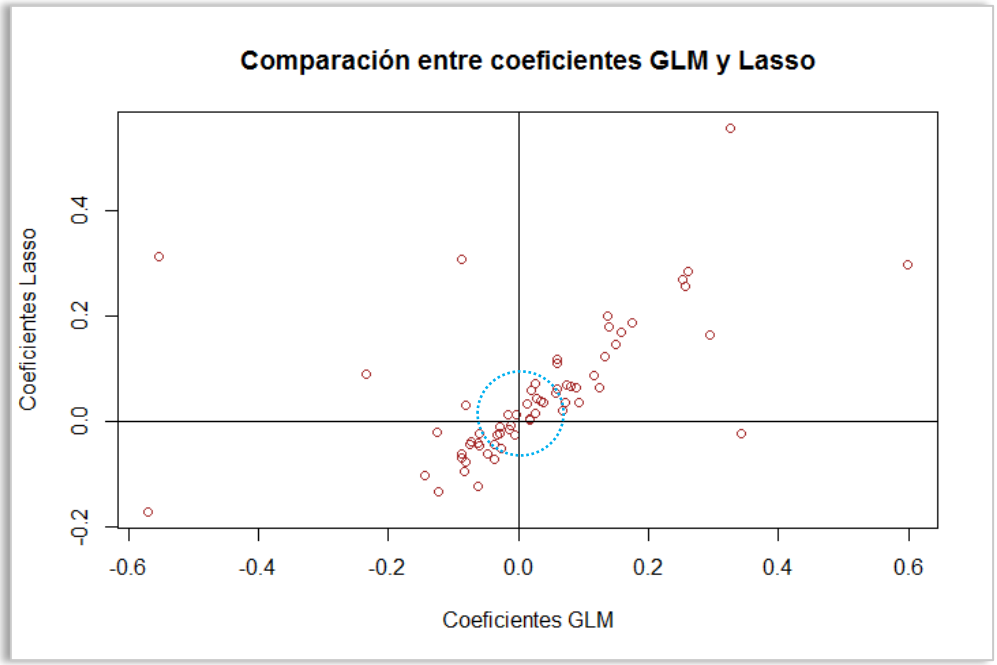


Gráfico elaborado por el autor.

Lo primero que destacar es la reducción del *"racimo de puntos"* que se sitúan en torno al origen en comparación con la figura 13.1.5, tras un primer proceso de selección. Todas estas variables que se sitúan en torno al origen no tienen porque al final no ser significativas y además, no serán muy discriminantes a la hora de identificar la diferencia de clientela. Por esta razón, lo más seguro es que la penalización del modelo *Elastic Net* sea superior y más restrictiva que la presentada en este capítulo.

13.3. Elastic Net

Como dijimos con anterioridad, un modelo Elastic Net tiene la potestad de emplear un *"*$\alpha \in [0.1; 0.9]$*"*. Para detectar el α empleado ejecutaremos una macro con 9 validaciones cruzadas, estudiando en cada uno de los casos la Poisson Deviance de cada validación cruzada para cada λ^{opt}, decidiendo así escoger aquella $\alpha^{elastic\ net}$ que presente la Poisson Deviance mínima.

Tras realizar la serie de validaciones cruzadas para cada α, encontramos que la menor Poisson Deviance se da para un $\alpha = 0.6$. Con ese α, nos situamos en una Poisson

Deviance igual a 0.1305854, siendo su $\lambda^{opt} = 0.0001213549$, o lo que es lo mismo, un $\log(\lambda) = -9.01679125$.

En la gráfica siguiente, se puede comprobar los diferentes Poisson Deviance que se conseguían para cada λ^{opt}.

Figura 13.3.1. Selección del α^{opt} para la modelización *Elastic Net.*

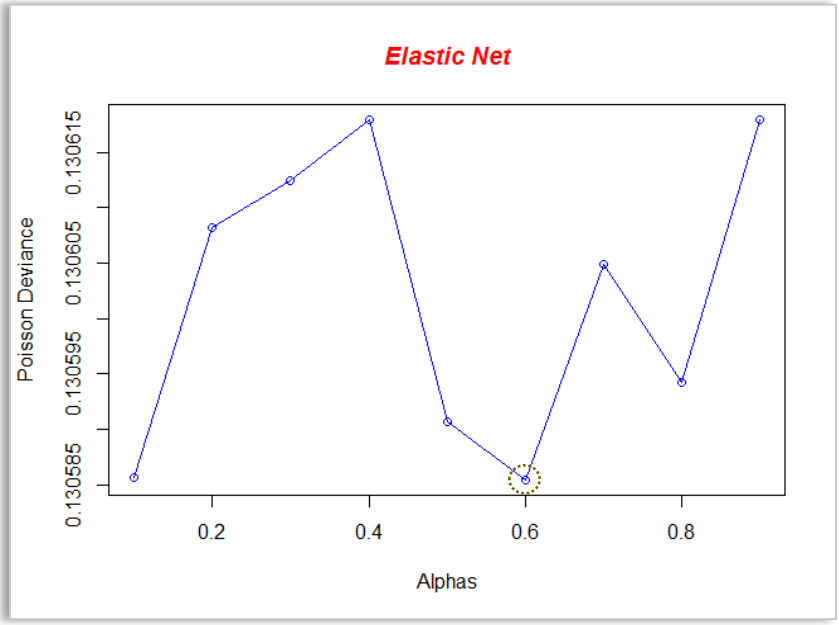


Gráfico elaborado por el autor.

Como puede verse, el modelo *Elastic Net* presenta una función idéntica a la modelización *Lasso,* pero como comentamos en el capítulo 8 este modelo no incurre en un coste de oportunidad de sesgo por varianza, tratándose de un modelo mejorado respecto al anterior. Esta mejora también se puede extrapolar a la práctica, ya que siempre los modelos óptimos *Elastic Net* presentan una Poisson Deviance menor que los modelos *Lasso* o *Ridge Regression*.

Si empezamos estudiando la modelización *Elastic Net* a través de la validación cruzada, igual que lo hicimos con los anteriores modelos, vemos que ésta presenta una menor pendiente en términos de Poisson Deviance que la del modelo *Lasso*. Esto quiere decir que, para conseguir una mayor penalización de las variables se necesita un mayor valor de λ, siendo los modelos *Elastic Net* menos restrictivos que la modelización *Lasso*.

Figura 13.3.2. Validación Cruzada[20] modelización *Elastic Net.*

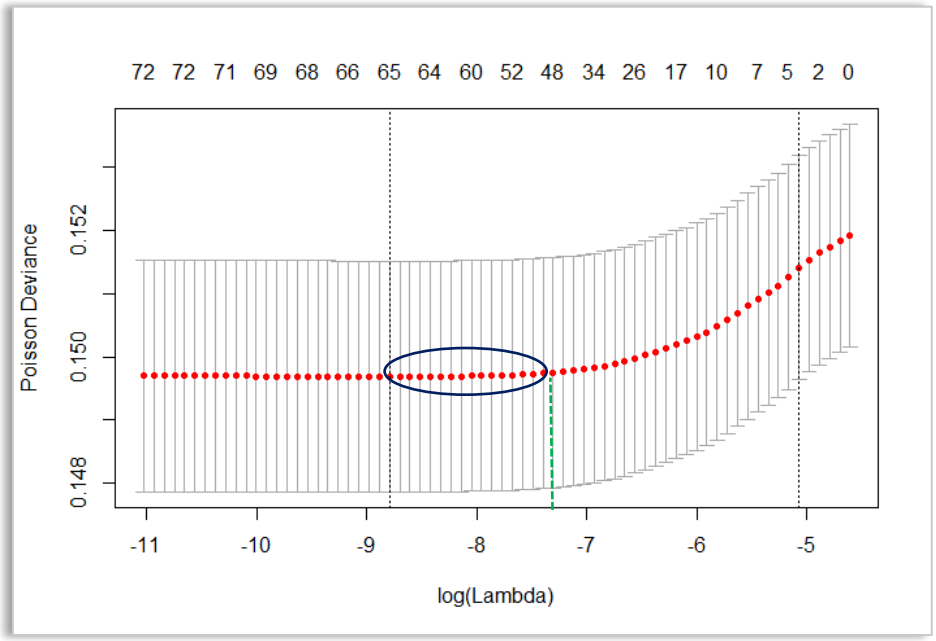


Gráfico elaborado por el autor.

Al tratarse de una modelización menos restrictiva, el λ^{opt} donde se minimiza la Poisson Deviance para el modelo Elastic Net selecciona un mayor número de variables que el modelo Lasso, por lo que en este caso es conveniente sacrificar Poisson Deviance a cambio de mayores niveles de penalización. Como vemos en la gráfica, estaríamos penalizando muy poco de Poisson Deviance obteniendo una selección de variables mucho más ajustada, situándonos para realizar el estudio en la línea serpenteada de color verde.

Por resituarnos rápidamente, hemos seleccionado un $\lambda = 0.0006711$, o lo que equivalente, a un $\log(\lambda) = -7.3$. Con estos niveles conseguimos seleccionar 48 de las 75 posibles variables dummies establecidas para el estudio de las variables internas, mientras que con el λ^{opt} dispondríamos de 65 variables.

En este caso, hemos realizado una tabla similar al de los otros dos casos pero con la diferencia mostrando los estimadores resultantes del modelo Elastic Net y GLM, aunque con el modelo establecido de variables internas que prediga la frecuencia de los siniestros RC Material Culpa.

[20] En el anexo se adjuntará las λ extraídas de la validación cruzada, con la finalidad de escoger un λ que se ajuste a nuestro objetivo.

Figura 13.3.3. Coeficientes *Elastic Net* y modelización *GLM* definitiva.

C. Num	Variables	Dummy	Coef. Elastic	Coef. GLM	% Dif
1	Intercepto		-4.1877	-4.3256 (0.0285***)	-3.19%
2	Nº años cia anter	<4	-0.0241	.	.
3	Nº años cia anter	Indeterminado	.	.	.
4	Potencia	<=68	.	.	.
5	Potencia	(68;89]	-0.0007	.	.
6	Potencia	(89;99]	.	.	.
7	Potencia	>122	.	.	.
8	Cilindrada	<=1351	-0.0300	.	.
9	Cilindrada	(1868;1896]	0.0072	.	.
10	Cilindrada	(1896;2188]	.	.	.
11	Cilindrada	>2188	0.0613	.	.
12	Peso Potencia	<=10.47	-0.0565	-0.0633 (0.0333**)	-10.74%
13	Peso Potencia	>13.44	0.0278	0.0545 (0.0271***)	-48.99%
14	Valor Vehículo	<=11450	-0.0008	.	.
15	Valor Vehículo	(11450;15485.50]	.	.	.
16	Valor Vehículo	(13370.85;17565]	0.0430	.	.
17	Ant. Vehículo	(12;38]	-0.0116	.	.
18	Ant. Vehículo	9999	.	.	.
19	Peso Vehículo	<=979.92	-0.0673	-0.1218 (0.0387***)	-44.75%
20	Peso Vehículo	>1619.76	0.0455	0.2 (0.0330***)	-77.25%
21	Velocidad	<=167	.	-0.0538 (0.0315*)	.
22	Velocidad	(187;209]	-0.0016	-0.0429 (0.0255*)	-96.27%
23	Velocidad	>209	-0.0163	-0.0826 (0.0453*)	-80.27%
24	E. Cond. Hab.	[18;28]	0.0814	.	.
25	E. Cond. Hab.	(53;74]	0.0291	.	.
26	Ant. Carn. Hab.	[0;10]	.	.	.
27	Ant. Carn. Hab.	[23;25]	-0.0206	.	.
28	Ant. Carn. Hab.	(25;78]	.	.	.
29	E. Cond. Ocas.	[18;30]	0.2534	.	.
30	E. Cond. Ocas.	(30;53)	-0.0136	.	.
31	E. Cond. Ocas.	(53;87]	.	.	.
32	Ant. Carn. Ocas.	[0;7]	0.1063	0.3423 (0.0338***)	-68.95%
33	Ant. Carn. Ocas.	[8;38]	-0.1021	-0.1816 (0.0664***)	-43.78%
34	Ant. Carn. Ocas.	[39;57]	.	-0.4033 (0.2776)	.
35	Modalidad	Terc. Básicos	.	.	.
36	Modalidad	Terc. Ampliados	.	.	.
37	Modalidad	TRSF	.	.	.
38	Forma de Pago	Semestral	0.1320	0.1584	-16.67%

				(0.0218***)	
39	Forma de Pago	Trimestral-Mensual	0.1572	-0.1829 (0.0444***)	-185.95%
40	Uso Vehículo	Trab y Uso Prof	.	.	.
41	Km Anuales	Hasta 5000 Km	-0.0074	.	.
42	Km Anuales	De 10001 a 15000 Km	0.0507	.	.
43	Km Anuales	Desde 15001 Km	.	.	.
44	Motor	Gasolina	-0.0914	-0.0719 (0.0222***)	27.12%
45	Motor	Otros/Desconocido	.	-0.1425 (0.2190)	.
46	Plazas	>5	0.0259	.	.
47	Plazas	Otros	.	.	.
48	Bonificación	Bonus Bajo	0.5324	0.5723 (0.0328***)	-6.97%
49	Bonificación	Bonus Medio	0.2854	0.3356 (0.0243***)	-14.96%
50	Puertas	3	-0.0363	-0.0607 (0.0242**)	-40.20%
51	Puertas	Indefinido	.	-15.6382 (13356.12)	.
52	Unidad Familiar	Sí	0.0811	0.0602 (0.0210***)	34.72%
53	Nº Auto Familia	0	0.2947	-0.0782 (0.0284***)	-476.85%
54	Nº Auto Familia	Más de 2	0.0161	0.0852 (0.0433**)	-81.10%
55	Nº Auto Familia	Indefinido	.	-15.6873 (139.4904)	.
56	Cond. Otros Veh.	Sí	-0.0877	-0.0626 (0.0227***)	40.10%
57	Morosidad	A-B	-0.0765	-0.0755 (0.0324**)	1.32%
58	Morosidad	D-E	0.0107	0.0309 (0.0234)	-65.37%
59	Morosidad	F	-0.2485	0.2778 (0.0383***)	-189.45%
60	Morosidad	Desconocido	0.1388	0.1617 (0.0281***)	-14.16%
61	Grupo Marca	1	0.0375	0.0575 (0.0198***)	-34.78%
62	Grupo Marca	3	-0.0801	-0.1350 (0.0657**)	-40.67%
63	Grupo Marca	4	0.1387	0.1895 (0.0569***)	-26.81%
64	Grupo Marca	5	0.2138	0.2988 (0.0834***)	-28.45%
65	Grupo Veh.	Tur. Famil/Monov	.	.	.
66	Grupo Veh.	Todot./Veh. Sobrep.	0.0094	.	.
67	Grupo Veh.	Otros Grupos Vehículos	.	.	.
68	Estado Civil	Divorciado/Separado	0.0088	0.1363	-93.54%

				(0.0563**)	
69	Estado Civil	Soltero	-0.0119	-0.0760	-84.34%
				(0.0302**)	
70	Estado Civil	Otros	.	-0.1625	.
				(0.0775**)	
71	Garaje	Garaje Individual	.	.	.
72	Garaje	Sin Garaje	0.0122	.	.
73	Profesión	Grupo 1	.	.	.
74	Profesión	Grupo 2	-0.0087	.	.
75	Profesión	Grupo 4	0.0222	.	.

Tabla elaborada por el autor.

Figura 13.3.4. Criterios para la Bondad de Ajuste del modelo *GLM*.

Criterios para la Bondad de Ajuste	
Criterio	Valor
Deviance	88179.90
Chi-Cuadrado de Pearson	545332.60
Verosimilitud log	-55684.95
AIC	111439.90
BIC	111829.18

Tabla elaborada por el autor.

Como vemos, la modelización interna seleccionada para predecir la variable respuesta, **Número de Siniestros RC Material Culpa**, no va muy desencaminada con la seleccionada a través del modelo *Elastic Net.* Las variables que conforman la regresión de variables internas son las siguientes: **Peso Potencia, Peso Vehículo, Velocidad, Antigüedad Carnet Conductor Ocasional, Forma de Pago, Motor, Bonificación, Puertas, Unidad Familiar, Número de Autos por Familia, Morosidad, Grupo Marca y Estado Civil.**

Es maravilloso atender a la columna de modelización GLM y ver la segregación que las dummies realizan sobre los grupos de riesgo, en el podemos destacar a las variables: la nota de morosidad, la antigüedad del carnet del conductor ocasional, peso potencia, peso vehículo...

Comparando ambos modelos, vemos que la mayoría de las variables dummies han sido seleccionadas por el modelo Elastic Net y las que no han sido seleccionadas se debe principalmente a la baja exposición[21] que presentan.

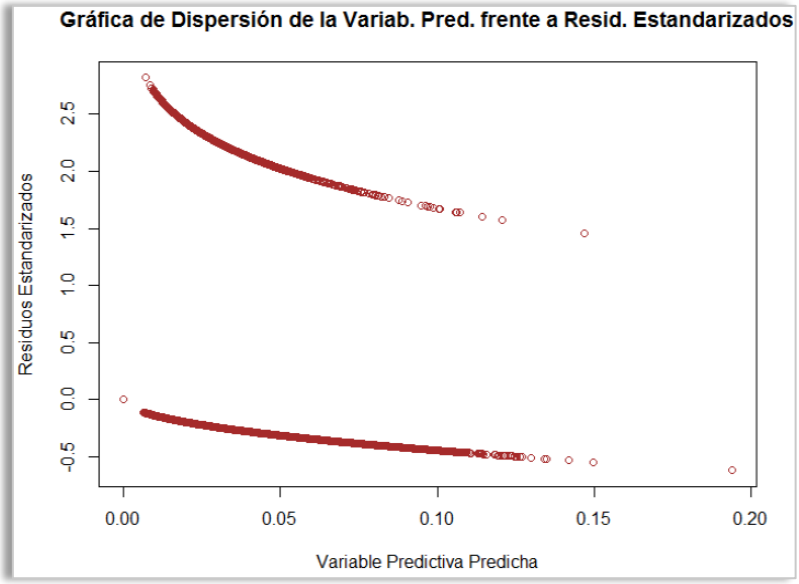
Mencionar que todas las variables seleccionadas por el modelo GLM son significativas, en términos generales, al 10% de significación individual y colectiva. En todos los casos, la incorporación de una de estas variables al modelo consiguen reducir los criterios de bondad de ajuste (ya sea: AIC, BIC y Deviance).

Tras estimar el modelo que mejor predice el número de siniestros RC material culpa, el siguiente paso será validar los residuos frente a los valores observados de *"y"*, y por otra parte, visualizar la existencia de datos atípicos que puedan distorsionar el estudio, determinando para ello la existencia de posibles puntos palanca que puedan ser puntos influyentes. Para poder detectarlos emplearemos dos metodologías estadísticas como son: el estudio del **efecto palanca** (Robustez a Priori) y la **Distancia de Cook** (Robustez a Posteriori). Adelantamos que el Modelo Lineal Generalizado trabajado no presenta valores outliers influyentes a un nivel de significación del 5%. Aun así, el estudio de las observaciones influyentes se encuentra detallado en el apéndice.

Validando los residuos a través de una gráfica de dispersión de la variable predictiva frente a los residuos estandarizados, vemos que como esperábamos, los residuos se sitúan en rachas en torno a los valores de la variable predictiva (en este caso, entre el 0 y 1).

[21] Estas variables que no han sido seleccionadas a través de las *Elastic Net* son: **puertas - indefinida**, **número de autos por familia – indefinido**, **motor – otros/desconocido, antigüedad carnet conductor ocasional [39,57], velocidad - <=167, estado civil – otros.**
En el anexo, pueden comprobarse sus niveles de exposición en los gráficos bivariantes.

Figura 13.3.5. Criterios para la Bondad de Ajuste del modelo *GLM*.



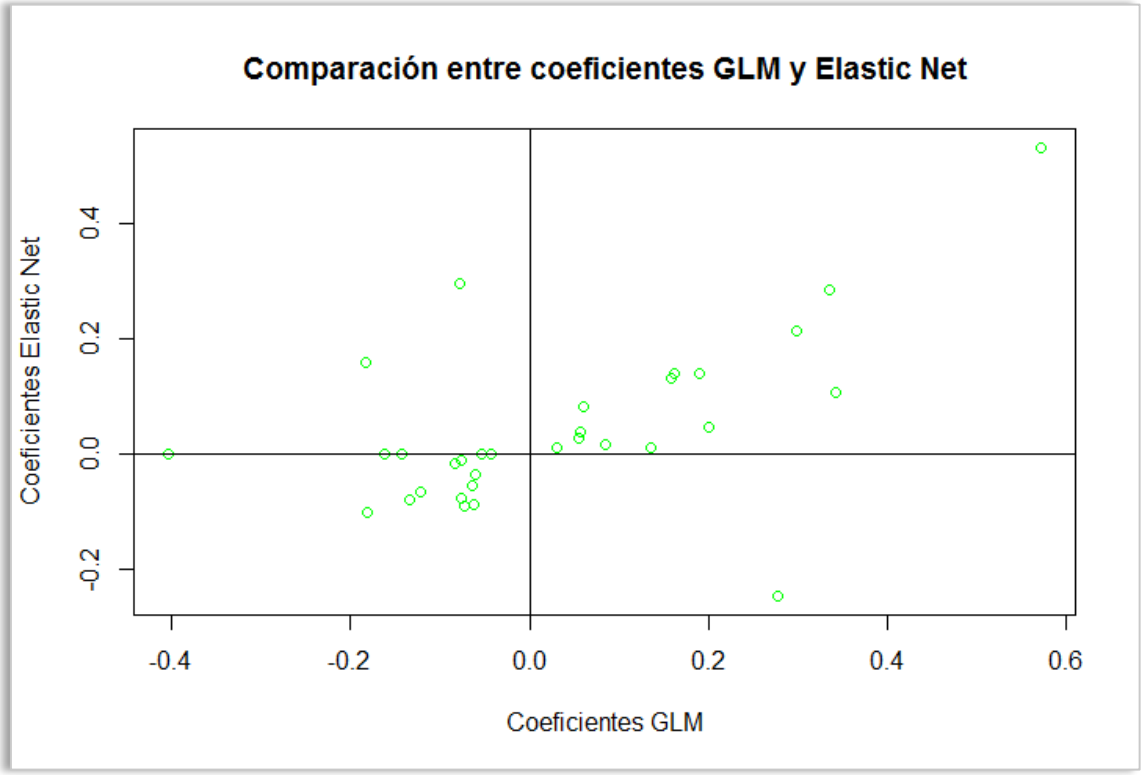
Gráfica elaborada por el autor.

Mencionar a continuación que no vamos a mostrar los gráficos de: **la regularización de las variables a través de un modelo *Elastic Net* y la cantidad de Deviance explicada por el conjunto de variables bajo las *Elastic Net*[22]**; debido gráfico que nos ilustra lo mismo que el modelo Lasso pero para otros niveles de $log\ (\lambda)$.

Por último, para finalizar el estudio de la modelización de variables internas voy a construir un gráfico de dispersión entre los coeficientes GLM del modelo final y los coeficientes del modelo Elastic Net.

[22] Vamos a introducir estos gráficos en el apéndice.

Figura 13.2.5. Gráfica de Dispersión comparativa de coeficientes GLM y *Elastic Net*[23].



Gráfica elaborada por el autor.

Lo primero que podemos comentar sobre este gráfico es que las variables que se sitúan cerca del origen son una minoría. Los coeficientes resultantes del GLM se parecen mucho a los coeficientes *Elastic Net*, marcándose una clara línea de tendencia creciente entre los puntos (que conlleva la excelente discriminación de riesgos según la categorización).

Seguramente los coeficientes que están próximos al eje de coordenadas serán los siguientes coeficientes regularizados si aumentásemos la λ.

Tras este análisis de la modelización interna, una primera y rápida conclusión que se puede extraer sobre lo visto en la práctica, es que parece que estamos corroborando el objetivo del estudio: verificar que podemos emplear técnicas de selección de variables como complemento para la modelización GLM. Aun así, vamos a esperar a realizar la modelización externa para extraer unas conclusiones certeras y definitivas.

[23] Hemos decidido no introducir el valor del intercepto para que no distorsione la interpretación del gráfico de dispersión.

14.Modelización Variables Externas

Tras realizar el análisis de las variables internas, vamos a intentar estudiar si las variables externas consiguen mejorar la predicción de la variable respuesta: **Número de Siniestros RC Material Culpa**. Para ello, ante las numerosas variables externas de las que disponemos en la base de datos original, hemos decidido realizar una batida de posibles variables significativas con la finalidad de no mostrar un estudio de variables sobredimensionado y perder con ello la noción del estudio. Hemos decidido introducir en el estudio de la modelización de variables externas, las 14 variables internas que consiguen predecir la variable respuesta según lo analizado en el capítulo anterior y 27 variables externas que hemos seleccionado para el modelo, tras la elaboración de un detallado estudio gráfico y estadístico univariable y bivariante de cada una de las variables externas disponibles.

Con dichas variables igual que en la modelización de variables internas, hemos ejecutado por este orden, los modelo *Ridge Regression*, *Lasso* y *Elastic Net*.

14.1. Ridge Regression

Como con la modelización de variables internas, seleccionamos la penalización que presente la menor Poisson Deviance, en este caso, correspondiente a un $\lambda^{opt} = 0.004914446$, o lo que es equivante, un $\log(\lambda^{opt}) = -5.31557625$ traduciendose este tunning en una *Poisson Deviance* de 0.1496503.

Figura 14.1.1. Validación Cruzada modelización *Ridge Regression*

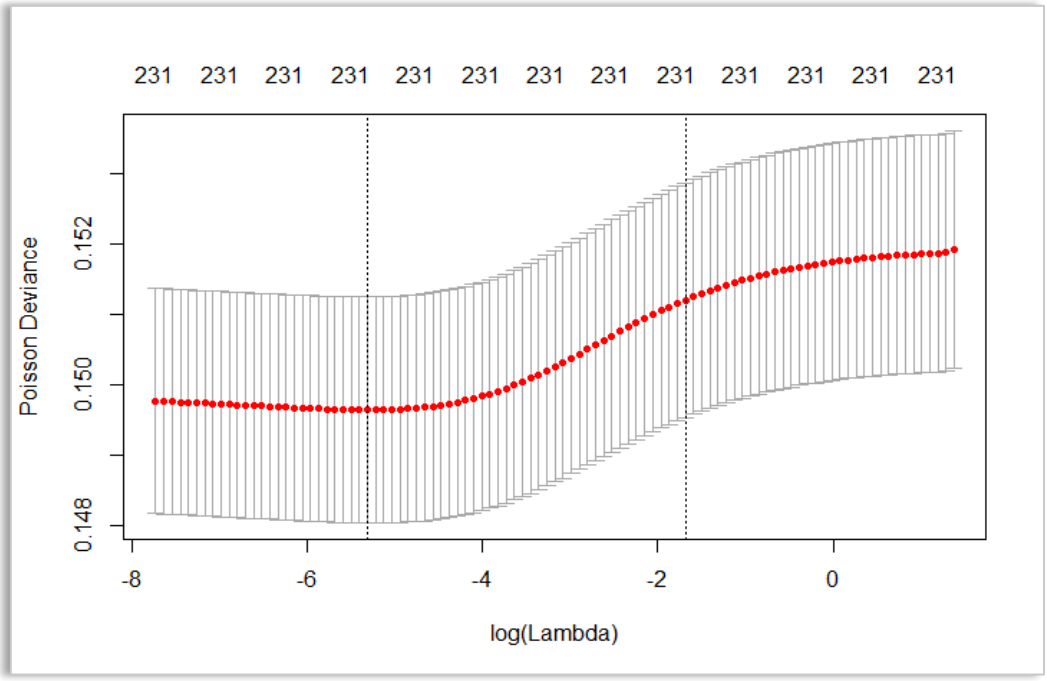


Gráfico elaborado por el autor.

Comparando esta λ^{opt} con la modelización de variables internas *Ridge Regression*, ésta es un *"poquitín"* más intensa, pero en cambio, la distorsión que se percibe sobre los coeficientes de las variables dummies externas es muy fuerte, causada por la baja exposición que presentan la mayoría de dummies al encontrarse muy tramificadas ante el interés personal de que sean variables muy discriminantes[24]. Sin embargo, las variables internas continúan presentando unas estimaciones similares a las presentadas en la modelización *Ridge Regression* del capítulo anterior.

La comparativa de resultados *Ridge Regression - GLM* se adjunta en el apéndice. En ella se puede ver lo comentado en el párrafo anterior.

14.2. Lasso

Continuando con el mismo esquema de lo expuesto durante el apartado anterior, en este caso la λ^{opt} es bastante más intensa que la que se empleaba para modelizar las

[24] este hecho nos va a causar un gran problema en la definición de las variables externas al incrementar consecuentemente la varianza de las variables.

variables internas. Esto significa que se necesita una mayor penalización a las variables para conseguir desechar variables que permitan minimizar la Poisson Deviance, situada en un valor de 0.1496652 (se consigue con un $\lambda^{opt} = 0.0001536116$, o lo que es lo mismo, con un $\log(\lambda) = -8.78108322$ → seleccionando aproximadamente 100 de las 232 variables introducidas en la modelización).

Figura 14.2.1. Validación Cruzada modelización *Lasso.*

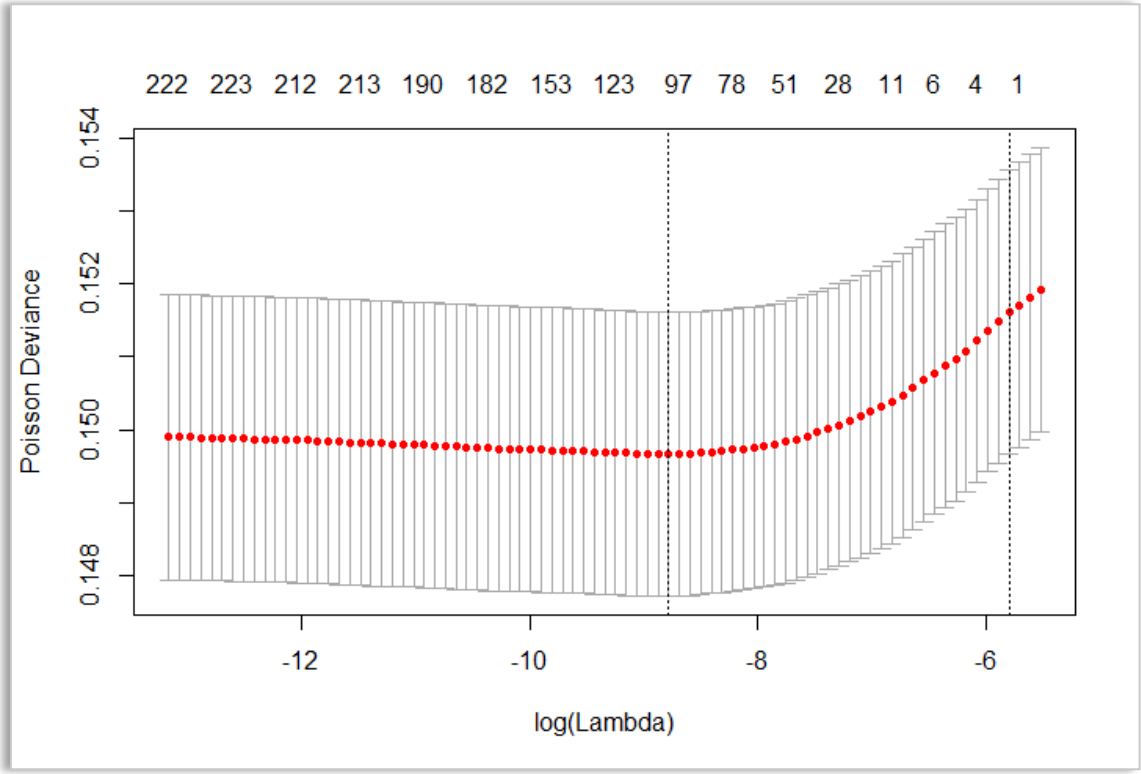


Gráfico elaborado por el autor.

Al igual que hemos realizado con la modelización Ridge Regression, los estimadores obtenidos a través de la modelización Lasso se van a adjuntar en el anexo del estudio, pero con estos resultados podemos resaltar la algunas variables externas que podrían ser significativas en la modelización GLM definitiva, como puede ser: **peso sector terciario, temperatura mínima, velocidad media del viento, racha de viento mayor a 91 km, días granizo.**

14.3. *Elastic Net*

Al igual que en el capítulo anterior, en la modelización *Elastic Net* debemos detectar antes de nada el α^{opt} que se definirá a través de la menor Poisson Deviance previamente un análisis de cada α para conocer cuál es la modelización que mejor se ajusta. Para ello, hemos ejecutado la macro con cada una de las posibles α, presentando los siguientes resultados:

Figura 14.3.1. Selección del α^{opt} para la modelización *Elastic Net*.

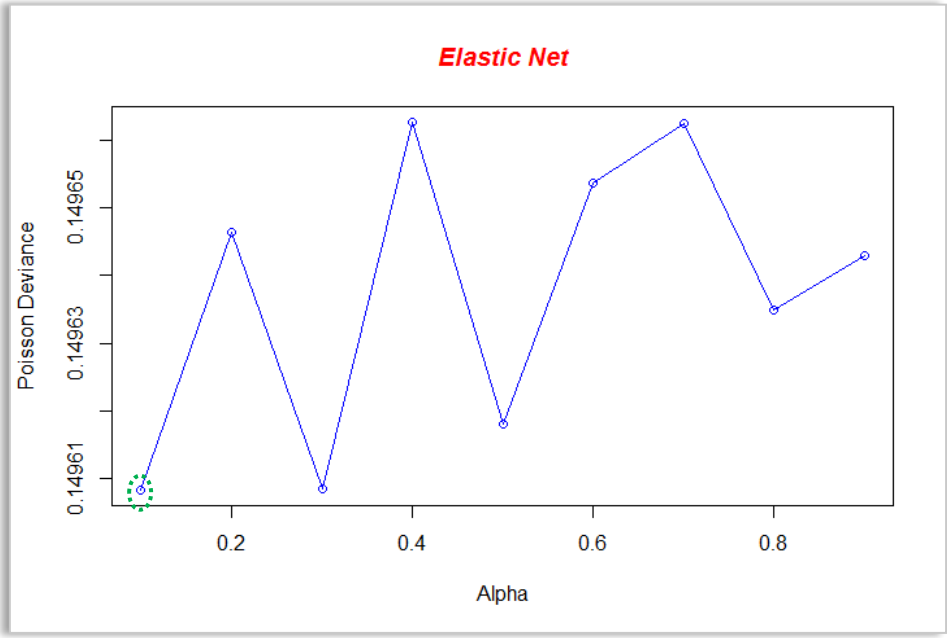


Grafico elaborado por el autor.

Como marca el gráfico anterior, el α^{opt} que menor Poisson Deviance es el 0.1. Para este α, la Poisson Deviance se situa en 0.1496082.

Como hemos dicho, la misma funcionalidad de la modelización Lasso y Elastic Net conlleva que ambas presenten un aspecto similar en su validación cruzada como puede comprobarse entre la figura 14.2.1 y la figura 14.3.2:

Figura 14.2.1. Validación Cruzada modelización *Elastic Net.*

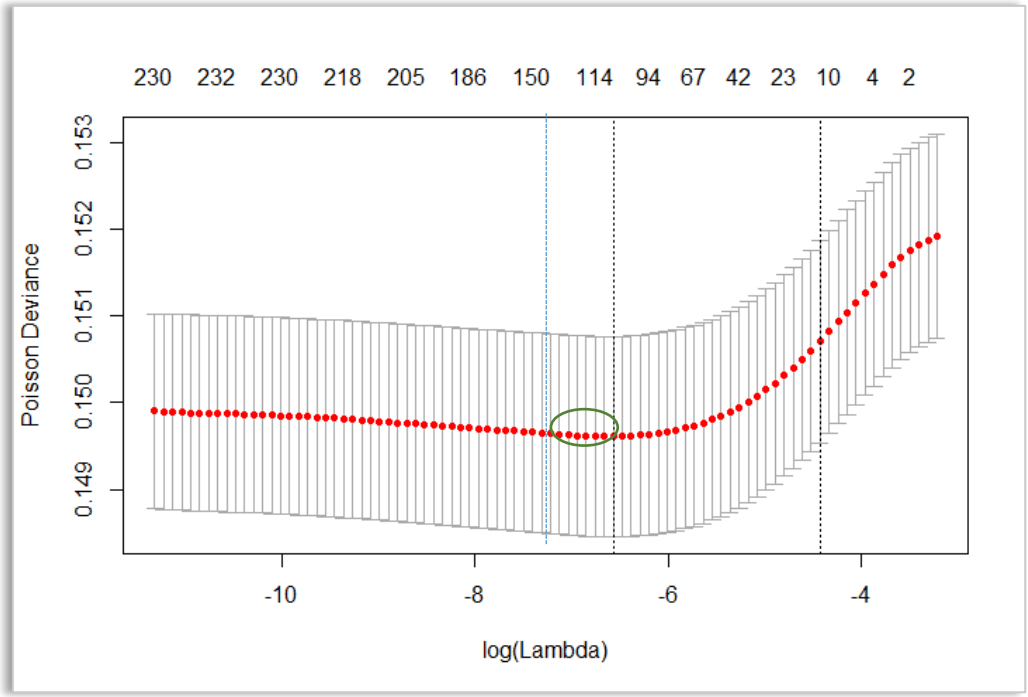


Grafico elaborado por el autor.

Como en el caso de la modelización *Lasso*, en un primer momento la selección de variables conlleva implícitamente una caída moderada de la Poisson hasta situarnos en un valor de 0.1496082, que corresponde a una $\lambda^{opt} = 0.001399652$, o lo que es lo mismo, un $\log \lambda^{opt} = -6.57153164$. Estos valores óptimos provocan que nos situemos en torno a 100 variables seleccionadas. Una mayor precisión en la selección de variables conlleva casi inmediatamente un fuerte incremento del Poisson Deviance.

Al realizar la selección de variables a través de los valores óptimos señalados en la validación cruzada, nos ha parecido que dicho *tunning* era muy estricto ante el problema de masa existente entre estas variables. Por ello, hemos decidido realizar el caso contrario a lo empleado en la modelización de variables internas exclusivamente, es decir, sacrificar mayores niveles de Poisson Deviance a cambio de conseguir un mayor número de variables durante la regularización (menor penalización), situándonos aproximadamente en la línea serpenteada de color azul (109 frente a 134 variables seleccionadas). Este cambio provoca que el tunning seleccionada para el análisis sea: $\lambda = 0.0008009$.

Al realizar la modelización *GLM,* hemos conseguido seleccionar tres de las 27 posibles variables externas para completar y mejorar la predicción de la frecuencia. En la columna de los estimadores Elastic Net se corrobora la selección automática de estas variables.

Figura 14.3.3. Coeficientes *Elastic Net* y modelización *GLM* definitiva[25].

C. Num	Variables	Dummy	Coef. Elastic	Coef. GLM
1	Intercepto		-4.2260	-4.3448 (0.0454***)
2	Peso Potencia	<=10.47	-0.0234	-0.0599 (0.0334**)
3	Peso Potencia	>13.44	0.0276	0.0483 (0.0271***)
4	Peso Vehículo	<=979.92	-0.1311	-0.1194 (0.0387***)
5	Peso Vehículo	>1619.76	0.1373	0.1551 (0.0331***)
6	Velocidad	<=167	-0.0222	-0.0546 (0.0315*)
7	Velocidad	(187;209]	-0.0372	-0.0386 (0.0256*)
8	Velocidad	>209	-0.0708	-0.0893 (0.0454*)
9	Ant. Carn. Ocas.	[0;7]	0.3564	0.3545 (0.0338***)
10	Ant. Carn. Ocas.	[8;38]	-0.1118	-0.1598 (0.0665***)
11	Ant. Carn. Ocas.	[39;57]	0.2999	-0.4139 (0.2777)
12	Forma de Pago	Semestral	0.1520	0.1583 (0.0218***)
13	Forma de Pago	Trimestral-Mensual	0.2012	0.1839 (0.0445***)
14	Motor	Gasolina	-0.1163	-0.0819 (0.0224***)
15	Motor	Otros/Desconocido	.	-0.1576 (0.2190)
16	Bonificación	Bonus Bajo	0.5399	0.5570 (0.0329***)
17	Bonificación	Bonus Medio	0.2961	0.3244 (0.0243***)
18	Puertas	3	-0.0361	-0.0495 (0.0243**)
19	Puertas	Indefinido	.	-15.6452 (13356.12)
20	Unidad Familiar	Sí	0.1035	0.0538

[25] Si se desea verificar la modelización completa *Elastic Net* para un $\lambda = 0.0008009$, ésta se adjuntará en el anexo

				(0.0211***)
21	Nº Auto Familia	0	0.3089	-0.0765
				(0.0284***)
22	Nº Auto Familia	Más de 2	0.0822	0.0901
				(0.0433**)
23	Nº Auto Familia	Indefinido	.	-15.6852
				(139.4564)
24	Cond. Otros Veh.	Sí	-0.1226	-0.0664
				(0.0227***)
25	Morosidad	A-B	-0.0384	-0.0267
				(0.0335**)
26	Morosidad	D-E	0.0233	0.0100
				(0.0239)
27	Morosidad	F	0.2870	0.2763
				(0.0335***)
28	Morosidad	Desconocido	0.1555	0.1488
				(0.0284***)
29	Grupo Marca	1	0.0559	0.0516
				(0.0199***)
30	Grupo Marca	3	-0.1343	-0.1364
				(0.0657**)
31	Grupo Marca	4	0.2164	0.1875
				(0.0570***)
32	Grupo Marca	5	0.2977	0.2967
				(0.0835***)
33	Estado Civil	Divorciado/Separado	0.0657	0.1322
				(0.0564**)
34	Estado Civil	Soltero	-0.0524	-0.0838
				(0.0302**)
35	Estado Civil	Otros	.	-0.1708
				(0.0775**)
36	Días Helada	84.01-119.04	0.0020	-0.0230
				(0.0344)
37	Días Helada	119.04-366.43	-0.0153	-0.0982
				(0.0306***)
38	Días Helada	366.43-622.77	-0.0029	-0.1960
				(0.0447***)
39	Días Helada	622.77-868.75	-0.0105	-0.1976
				(0.0409***)
40	Días Helada	868.75-high	-0.0040	-0.1310
				(0.0415***)
41	Días Helada	Indeterminado	-0.0427	-16.2037
				(2610.810)
42	Pes Sect Terc	Low-51.13	-0.0684	-0.1313
				(0.0303***)
43	Pes Sect Terc	51.13-67.59	-0.0212	-0.0397
				(0.0226*)
44	Pes Sect Terc	82.35-high	0.0097	0.0111
				(0.0322)
45	Pes Sect Terc	Indeterminado	.	.
46	Veloc Med Vient	Low-2.57	0.0338	0.1824

				(0.0488***)
47	Veloc Med Vient	2.57-2.59	.	0.2183
				(0.0497***)
48	Veloc Med Vient	2.59-2.71	0.0081	0.2051
				(0.0535***)
49	Veloc Med Vient	2.71-2.82	0.0598	0.1318
				(0.0468***)
50	Veloc Med Vient	2.82-2.90	.	0.0958
				(0.0351***)
51	Veloc Med Vient	2.91-3.05	0.0597	0.1712
				(0.0416***)
52	Veloc Med Vient	3.05-3.26	-0.0725	0.0703
				(0.0741)
53	Veloc Med Vient	3.26-3.51	-0.0110	0.2095
				(0.0432***)
54	Veloc Med Vient	3.51-high	0.0221	0.1699
				(0.0416***)
55	Veloc Med Vient	Indeterminado	.	.

Grafico elaborado por el autor.

Figura 14.3.4. Criterios para la Bondad de Ajuste del modelo *GLM*.

Criterios para la Bondad de Ajuste	
Criterio	Valor
Deviance	87973.67
Chi-Cuadrado de Pearson	544773.10
Verosimilitud log	-55572.83
AIC	111251.67
BIC	111841.11

Tabla elaborada por el autor.

Tres son las variables externas incorporadas a la modelización *GLM*, pudiendo dividirse a su vez en, dos variables meteorológicas: **Días Heladas** y **Velocidad Media del Viento**; y una sociodemográfica: **Peso del Sector Terciario**. Como indicábamos en el capítulo 9, la elevada tramificación de las variables y el alto grado de asociación entre las variables externas provocan la selección de pocas variables relevantes durante la modelización *GLM* a pesar de disponer de un gran número de variables externas en la base de datos. En cambio, parece que la incorporación de estas variables ayudan a predecir mejor la siniestralidad (al reducirse los valores del AIC: 111439.90 vs 111251.67) [26].

[26] El valor del BIC en este caso, no consigue reducir el valor que presentaba el modelo interno. Esto se debe a que se trata de una metodología que penaliza mucho la entrada de otra variable (ante el posible problema de sobre-parametrización del modelo), si no se trata de una verdaderamente significativa.

Tras mostrar el modelo de variables internas y externas conjuntamente, validaremos los residuos frente a los valores predichos por la regresión, siguiendo los mismos pasos que durante el capítulo 13[27].

Figura 14.3.5. Criterios para la Bondad de Ajuste del modelo *GLM*.

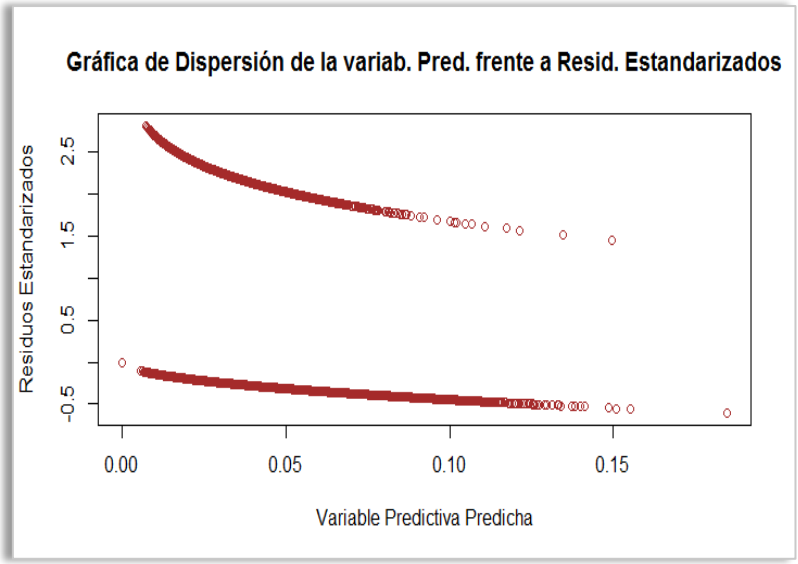


Gráfico elaborado por el autor.

Como era de suponer, los residuos se sitúan en dos rachas entorno a los valores de la variable predictiva, pudiendo validar los residuos y adelantando la inexistencia de outliers en la base de datos.

[27] Al igual que en el capítulo 13, el estudio de posibles datos atípicos a través de los Puntos Palanca y la Distancia de Cook será analizada en el anexo.

15.Conclusiones

Hace dos o tres décadas nadie podía imaginar la transformación radical que han sufrido las tarifas de autos en el sector español. Pero los cambios e innovaciones en los departamentos de *"Pricing"* y *"Gestión de Negocio"* no terminan. La guerra de precios y la captación de los clientes con un riesgo bajo permite el continuo desarrollo de nuevas técnicas y variables discriminantes.

Ante este futuro, es importante el conocimiento de nuevas técnicas estadísticas avanzadas que ayuden en el proceso de tarificación. El mundo del *Machine Learning* abre un abanico, hasta hoy en día prácticamente desconocido, con diferentes técnicas que permitirá sacar el máximo rendimiento a la información procedente del *Big Data.* Estas técnicas además facilitaran, en el futuro, el tratamiento de muchas variables procedentes de la información extraída por sistemas telemáticos como *"Pay as you Drive",* que seguramente cambie el panorama asegurador tradicional conocido hasta el momento.

El objetivo principal del estudio realizado ha sido demostrar si el empleo de técnicas como el *Ridge Regression*, *Lasso* o las *Elastic Net* permiten seleccionar variables para la posterior modelización *GLM.* Los resultados obtenidos durante este estudio han sido claros: ***"siempre que no exista un problema de masa en las variables[28], los modelos de selección de variables son complementos ideales en la modelización GLM; sin la necesidad de emplear técnicas convencionales e ineficientes como los modelos de selección Stepwise".*** Eso sí, el empleo de estas técnicas debe de ir acompañado del conocimiento de potentes softwares estadístico como puede ser el **SAS** o el **R-Studio**. Otra plataforma nueva e interesante para llevar a cabo estos estudios puede ser a través de **Microsoft Azure**, ya que permite ejecutar una mayor oferta de recursos estadísticos, optimizando el rendimiento a través de máquinas virtuales alojadas en la *nube*.

[28] Si existe un problema de masa la modelización *GLM* será también imprecisa al tener un problema de varianzas.

También podemos concluir que, la inclusión de variables exógenas en la tarificación a priori enriquece y mejora consistentemente las predicciones de las variables respuestas.

Destacar por otra parte, la importancia de un correcto análisis previo de las variables disponibles en la base de datos que permita conocerlas en profundidad y proceder posteriormente a su categorización; pudiéndose emplear para ello distintas técnicas estadísticas, entre las que nosotros destacamos los **Árboles de Decisión (CHAID)**, pertenecientes al colectivo de Aprendizaje Automático**.**

En definitiva, las aseguradoras deben estar preparadas para asumir todo tipo de cambios en sus procesos de tarificación, en un período de tiempo todavía indeterminado pero vislumbrado, en el que, la continua formación de los actuarios y la innovación será vital para la supervivencia de las empresas aseguradoras.

16.Apéndice

16.1. Categorización, Análisis Univariante y Bivariante

En este apartado, se podrá analizar las variables seleccionadas para la modelización GLM como variables relevantes en el estudio.

Peso-Potencia

Figura 16.1.1. Árbol de Decisión: Peso Potencia

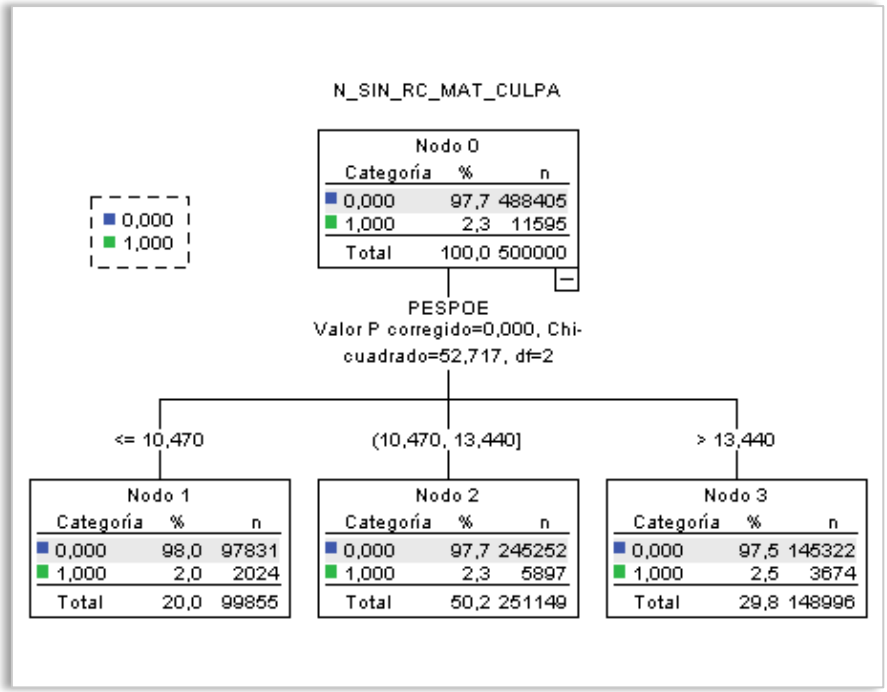


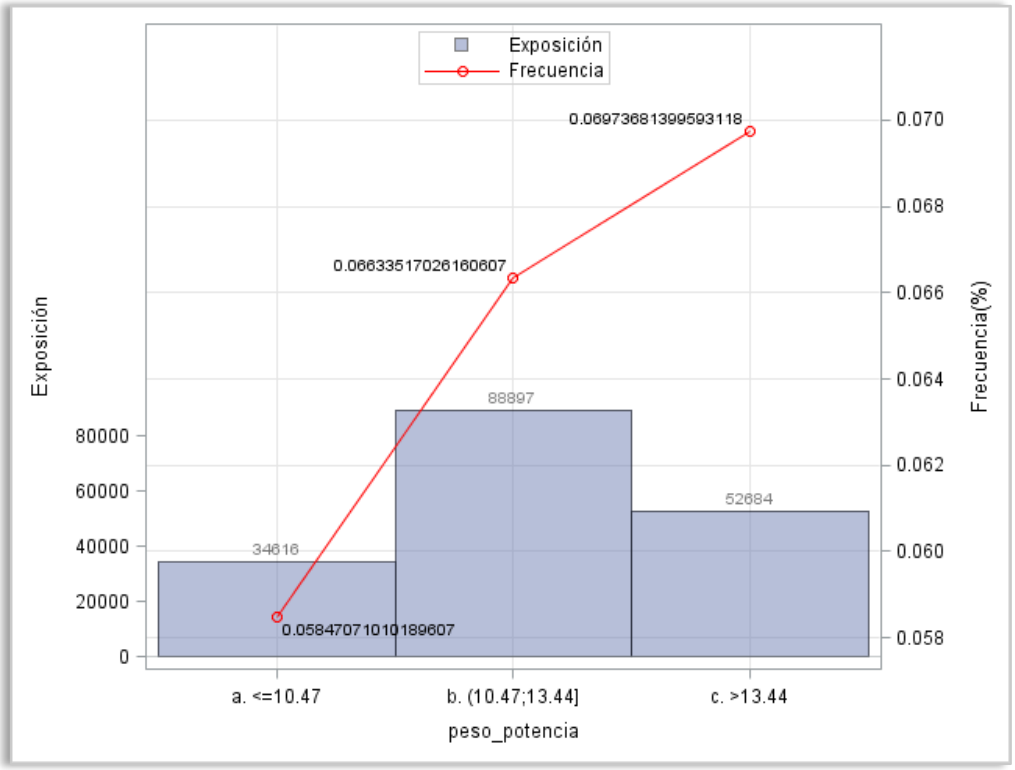
Gráfico elaborado por el autor.

Figura 16.1.2. Tabla Frecuencia: Peso Potencia

peso_potencia	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
a. <=10.47	99855	19.97	99855	19.97
b. (10.47;13.44]	251149	50.23	351004	70.20
c. >13.44	148996	29.80	500000	100.00

Tabla elaborada por el autor

Figura 16.1.3. Gráfico Bivariante: Peso Potencia



Gráfica elaborada por el autor

Peso Vehículo

Figura 16.1.4. Árbol de Decisión: Peso Vehículo

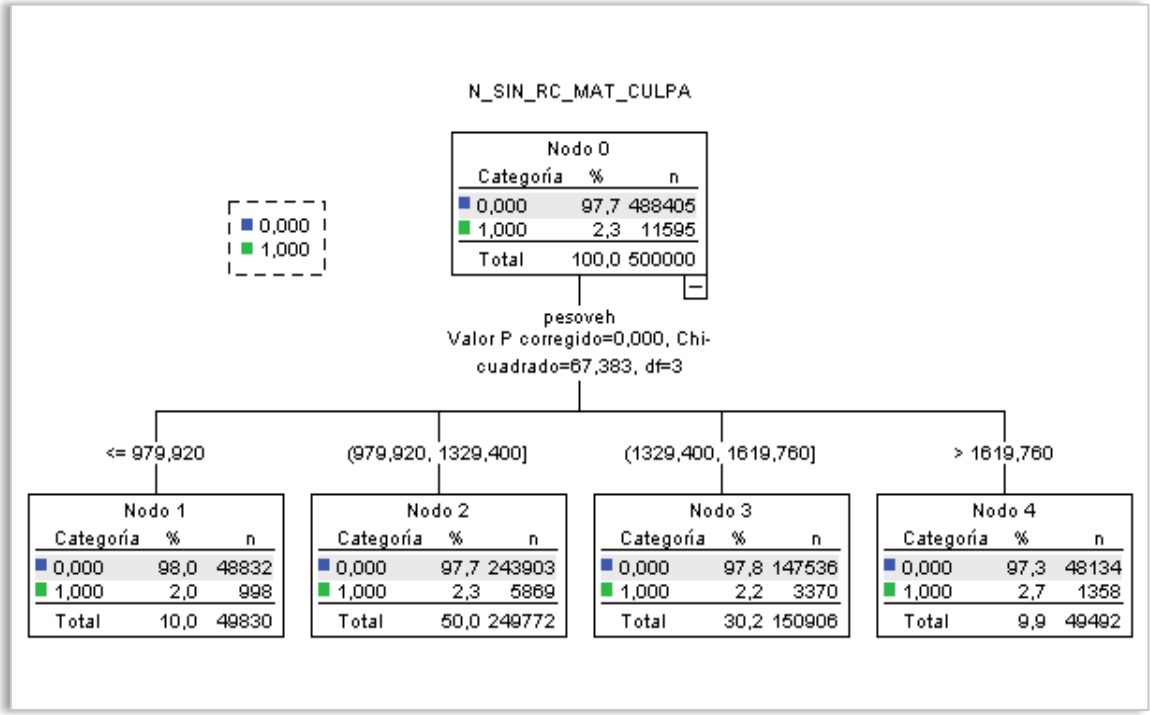


Gráfico elaborado por el autor.

Figura 16.1.5. Tabla Frecuencia: Peso Vehículo.

peso_vehiculo	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
a. <=979.92	49830	9.97	49830	9.97
b. (979.92;1619.76]	400678	80.14	450508	90.10
c. 1619.76	49492	9.90	500000	100.00

Tabla elaborada por el autor.

Figura 16.1.6. Gráfico Bivariante: Peso Vehículo

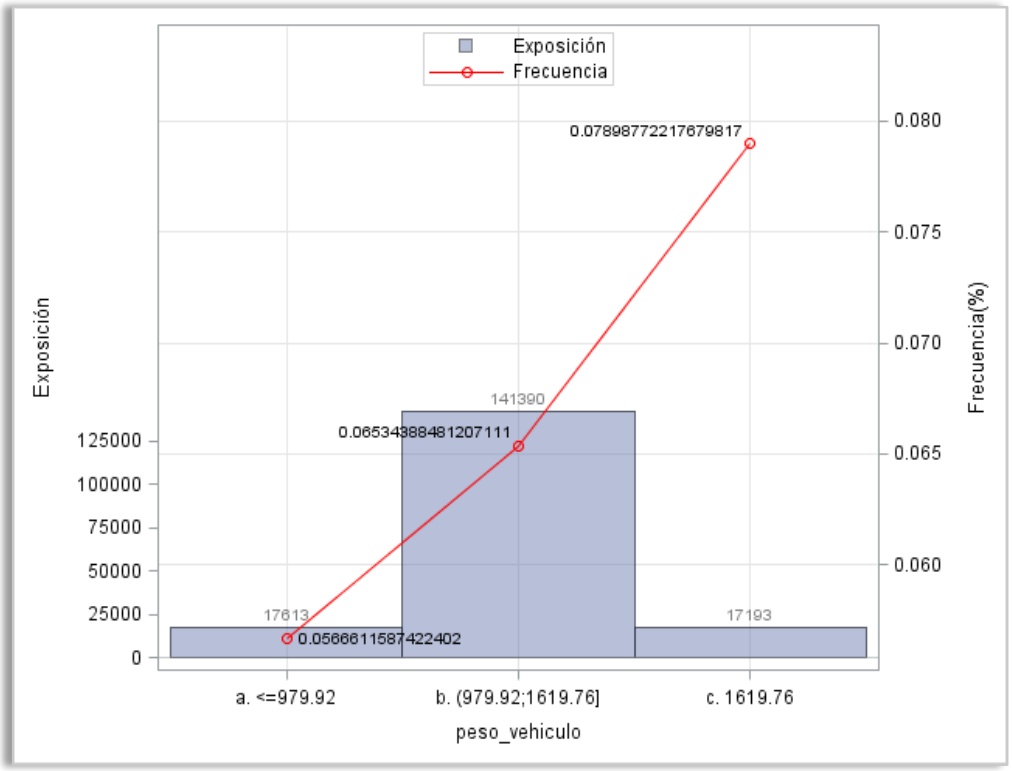


Gráfico elaborado por el autor.

Velocidad

Figura 16.1.7. Árbol de Decisión: Velocidad

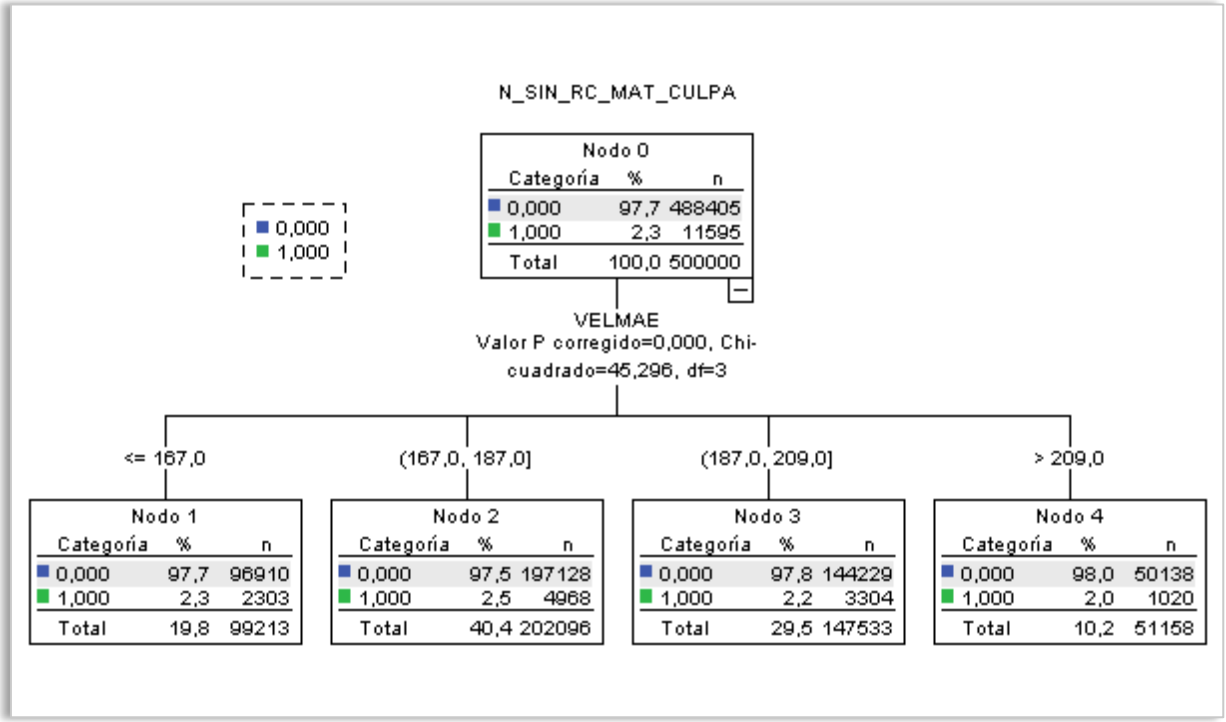


Gráfico elaborado por el autor.

Figura 16.1.8. Tabla Frecuencia: Velocidad

velocidad	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
a. <=167	99213	19.84	99213	19.84
b. (167;187]	202096	40.42	301309	60.26
c. (187;209]	147533	29.51	448842	89.77
d. >209	51158	10.23	500000	100.00

Tabla elaborada por el autor.

Figura 16.1.9. Gráfico Bivariante: Velocidad.

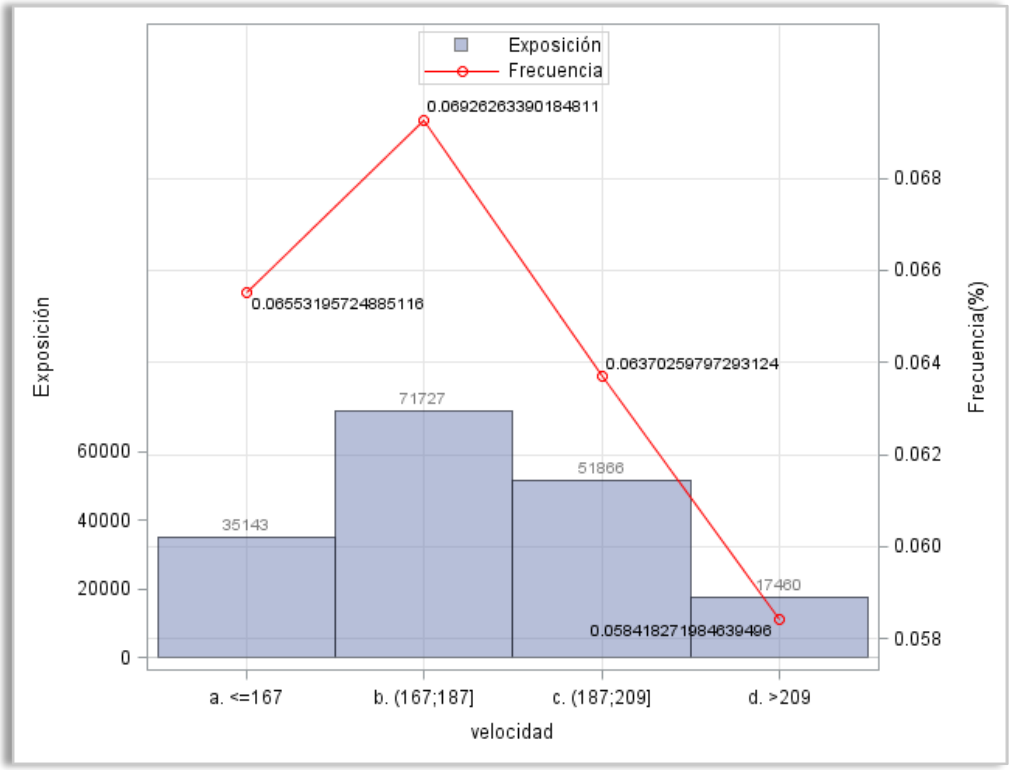


Gráfico elaborado por el autor.

Antigüedad Carnet Ocasional

Figura 16.1.10.Tabla Frecuencia: Antigüedad Carnet Ocasional

antiguedad_carnet_oca	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
a. 9999	453989	90.80	453989	90.80
b. [0;7]	32544	6.51	486533	97.31
c. [9;38]	13051	2.61	499584	99.92
d. [38;57]	416	0.08	500000	100.00

Tabla elaborada por el autor.

Figura 16.1.11. Gráfico Bivariante: Antigüedad Carnet Ocasional.

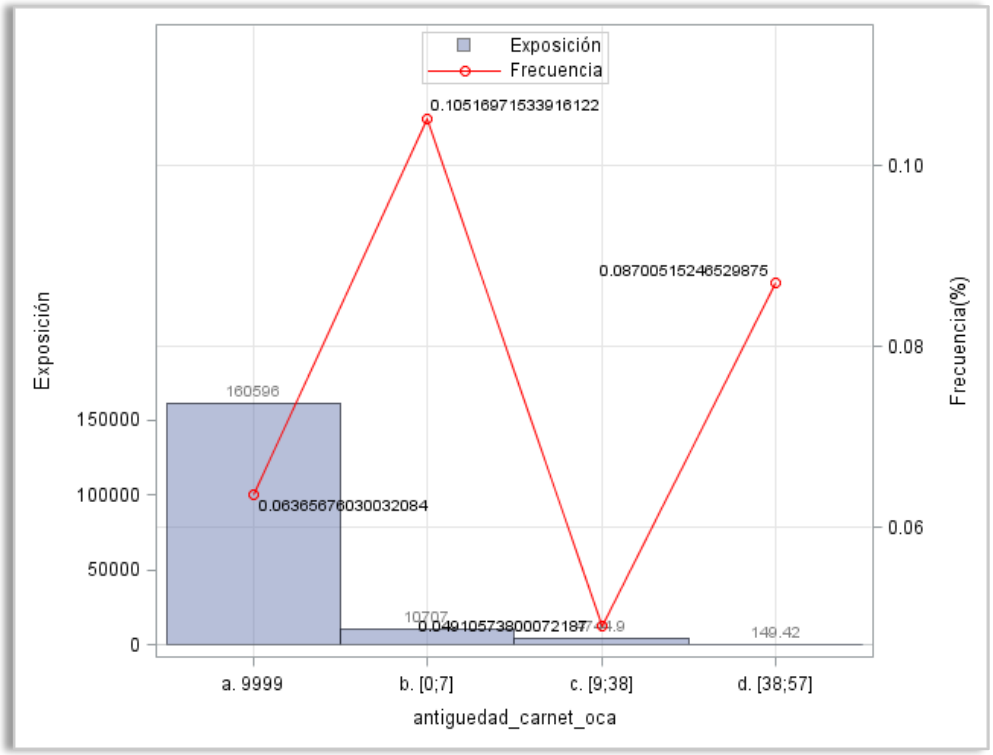


Gráfico elaborado por el autor.

Forma de Pago

Figura 16.1.12. Tabla Frecuencia: Forma de Pago

pago	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
1. Anual	365874	73.17	365874	73.17
2. Semestral	114506	22.90	480380	96.08
3. Trimestral-Mensual	19620	3.92	500000	100.00

Tabla elaborada por el autor.

Figura 16.1.13. Gráfico Bivariante: Forma de Pago.

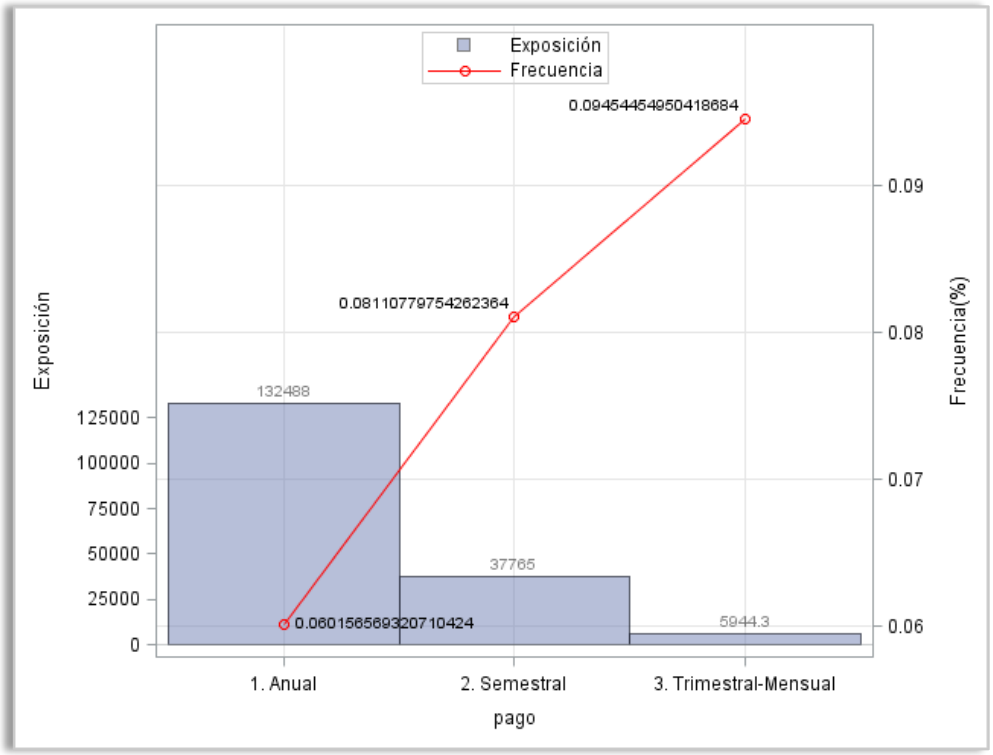


Gráfico elaborado por el autor.

Motor

Figura 16.1.14. Tabla Frecuencia: Tipo de Motor.

motor	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
1. Gasolina	178837	35.77	178837	35.77
2. Diesel	320115	64.02	498952	99.79
3. Otros/Desconocido	1048	0.21	500000	100.00

Tabla elaborada por el autor.

Figura 16.1.15. Gráfico Bivariante: Tipo de Motor.

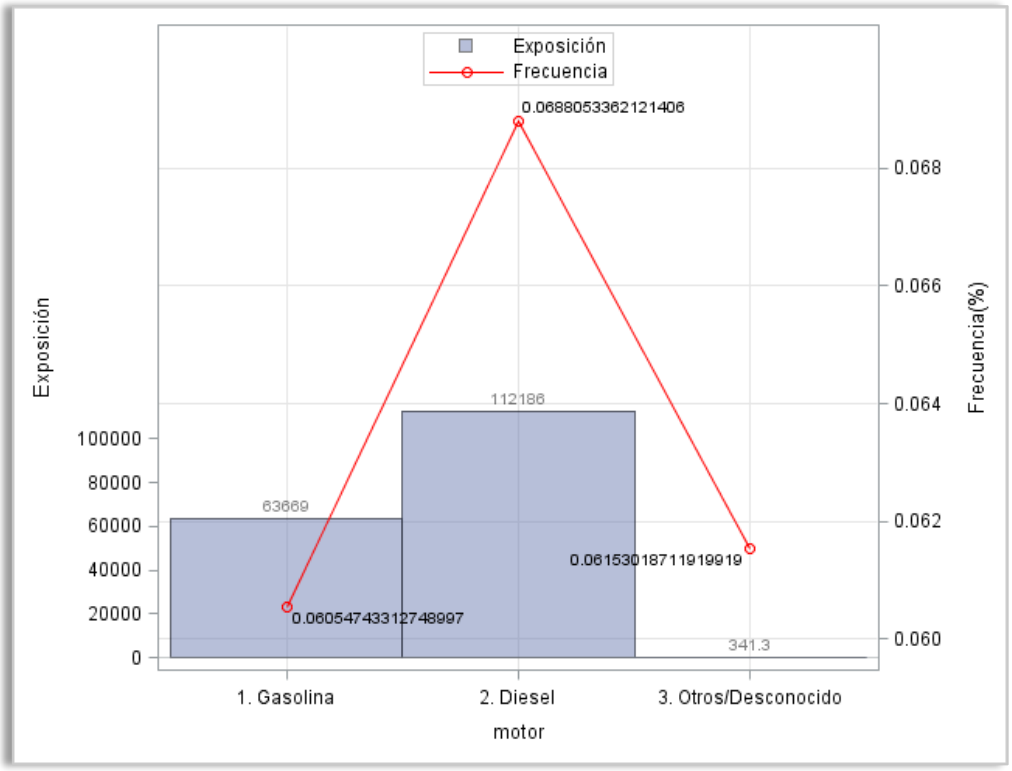


Gráfico elaborado por el autor.

Bonus

Figura 16.1.16. Tabla Frecuencia: Bonus

bonus	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
1. Bonus Bajo	34072	6.81	34072	6.81
2. Bonus Medio	81828	16.37	115900	23.18
3. Bonus Alto	384100	76.82	500000	100.00

Tabla elaborada por el autor.

Figura 16.1.17. Gráfico Bivariante: Bonus

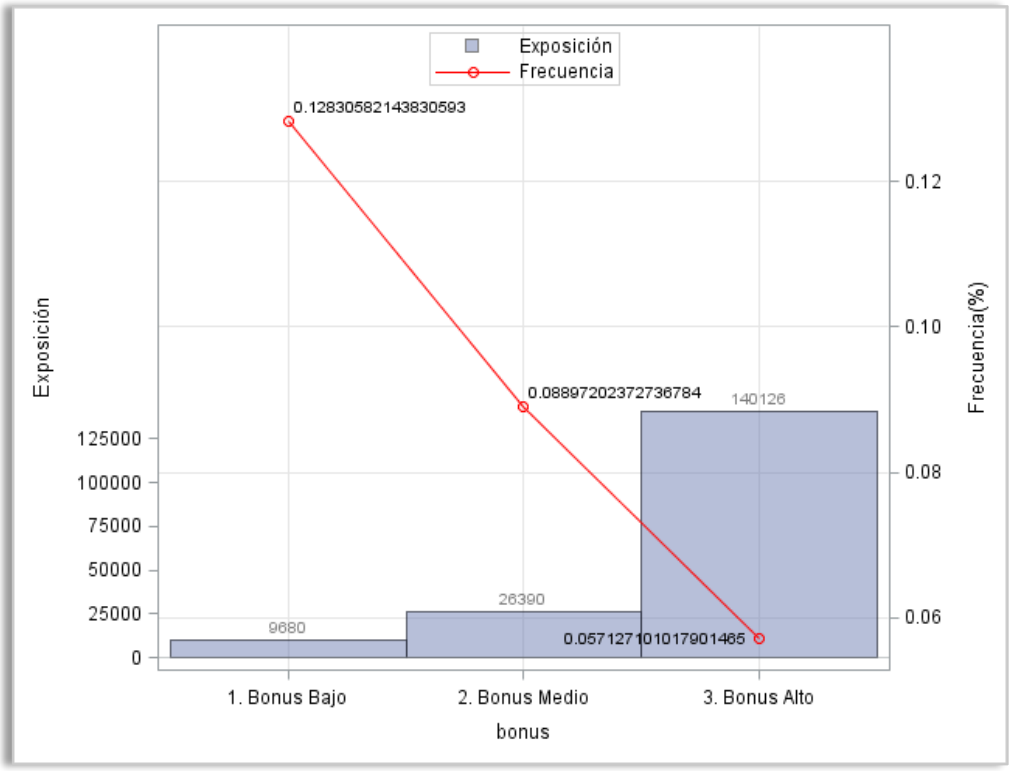


Gráfico elaborado por el autor.

Puertas

Figura 16.1.18. Tabla Frecuencia: Puertas

puertas	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
a. 3	109831	21.97	109831	21.97
b. 5	390167	78.03	499998	100.00
c. Indefinido	2	0.00	500000	100.00

Tabla elaborada por el autor.

Figura 16.1.19. Gráfico Bivariante: Puertas.

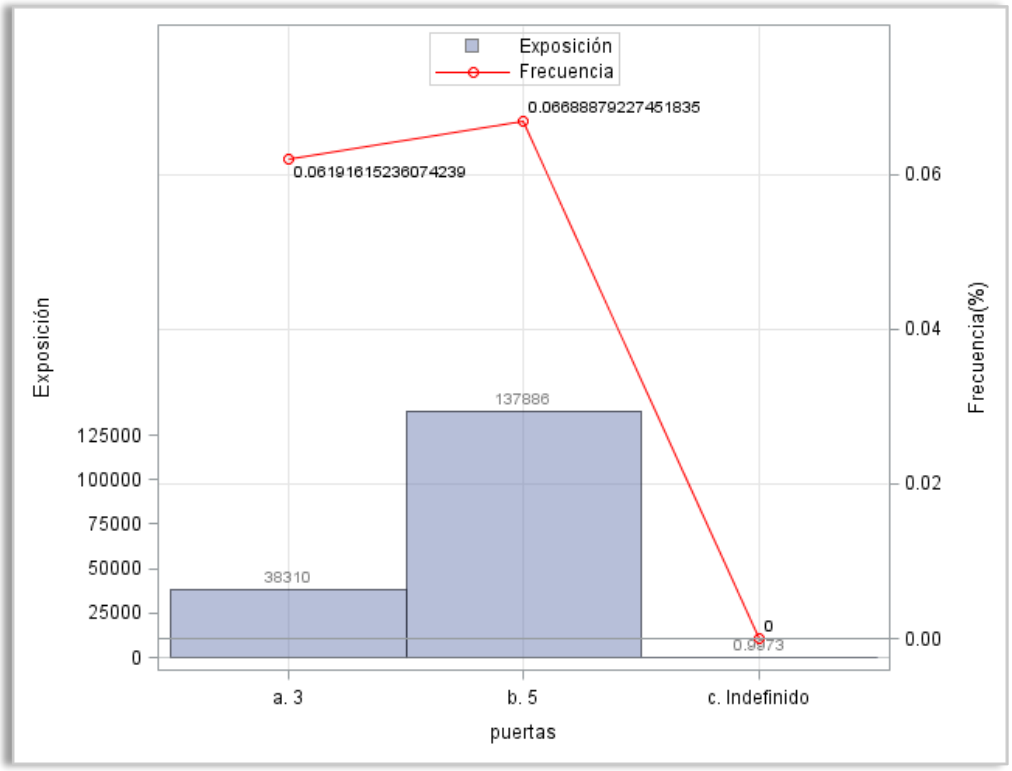


Gráfico elaborado por el autor.

Unidad Familiar

Figura 16.1.20. Tabla Frecuencia: Unidad Familiar

unidad_familiar	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
NO	322242	64.45	322242	64.45
SI	177758	35.55	500000	100.00

Tabla elaborada por el autor.

Figura 15.1.21. Gráfico Bivariante: Unidad Familiar

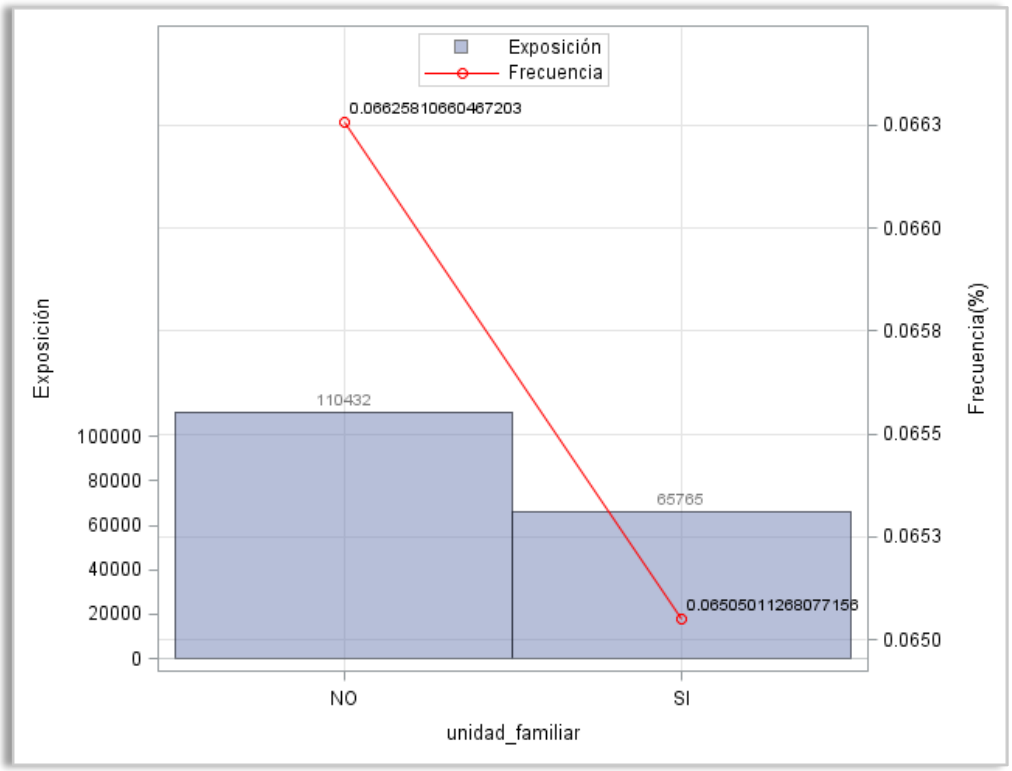


Gráfico elaborado por el autor

Número de Autos por Familia

Figura 16.1.22. Tabla Frecuencia: Número de Automóviles en la Familia

n_autos_familia	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
a. 0	76743	15.35	76743	15.35
b. 1-2	381367	76.27	458110	91.62
c. Más de 2	23758	4.75	481868	96.37
d. Indefinido	18132	3.63	500000	100.00

Tabla elaborada por el autor.

Figura 16.1.23. Gráfico Bivariante: Número de Automóviles en la Familia.

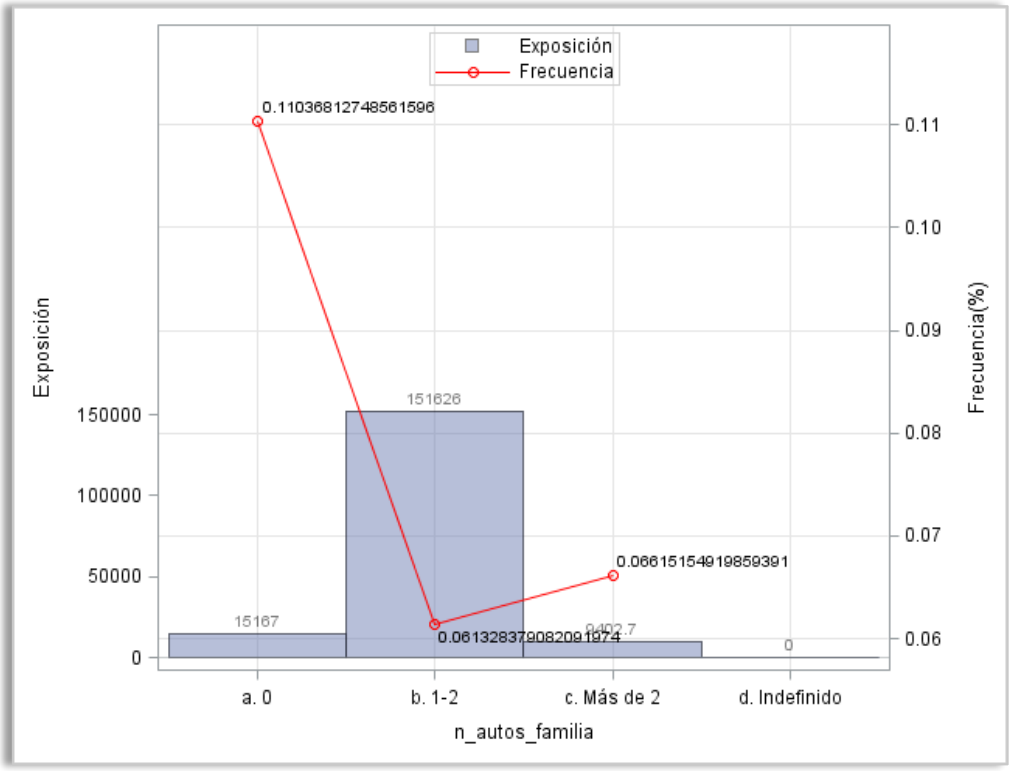


Gráfico elaborado por el autor.

Conduce Otros Vehículos

Figura 16.1.24. Tabla Frecuencia: Conduce Otros Vehículos

cond_otros_vehiculos	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
No	355017	71.00	355017	71.00
Si	144983	29.00	500000	100.00

Tabla elaborada por el autor.

Figura 16.1.25. Gráfico Bivariante: Conduce Otros Vehículos.

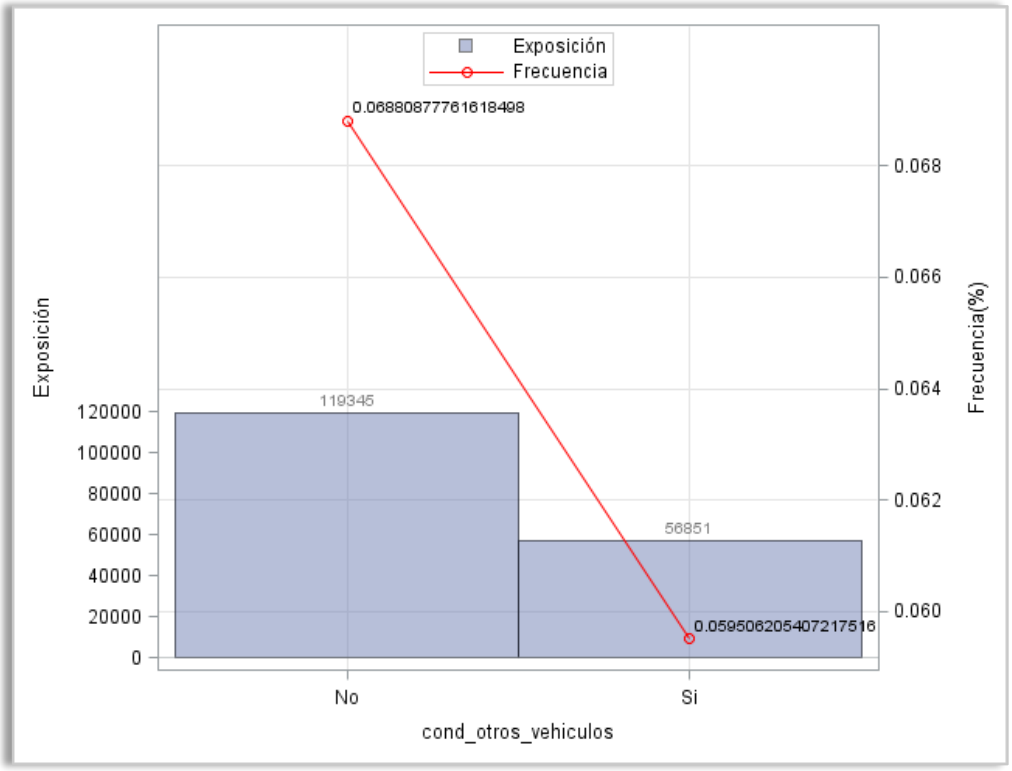


Gráfico elaborado por el autor.

Morosidad

Figura 16.1.26. Tabla Frecuencia: Morosidad.

morosidad	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
1. A-B	59696	11.94	59696	11.94
2. C	182758	36.55	242454	48.49
3. D-E	140177	28.04	382631	76.53
4. F	28865	5.77	411496	82.30
5. Desconocido	88504	17.70	500000	100.00

Tabla elaborada por el autor.

Gráfico 16.1.27. Gráfico Bivariante: Morosidad.

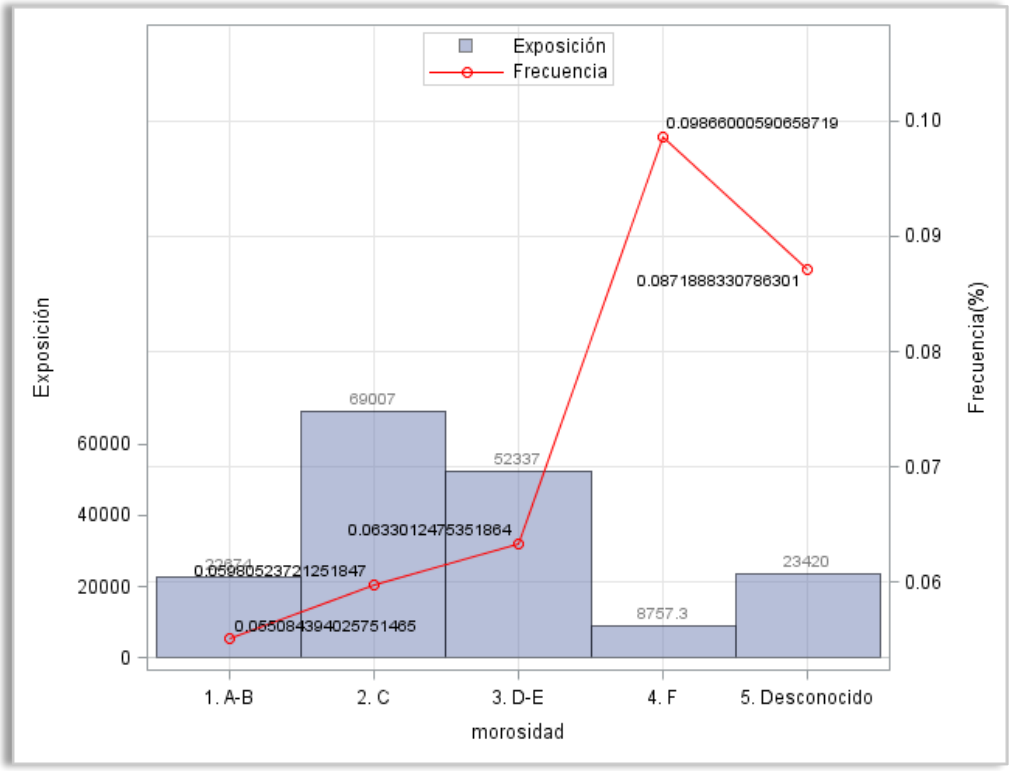


Gráfico elaborado por el autor.

Grupo Marca

Figura 16.1.28. Tabla Frecuencia: Grupo de Marca.

grupo_marca	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
a. Grupo 1	180837	36.17	180837	36.17
b. Grupo 2	287927	57.59	468764	93.75
c. Grupo 3	13450	2.69	482214	96.44
d. Grupo 4	13127	2.63	495341	99.07
e. Grupo 5	4659	0.93	500000	100.00

Tabla elaborada por el autor.

Gráfico 16.1.29. Gráfico Bivariante: Grupo de Marca.

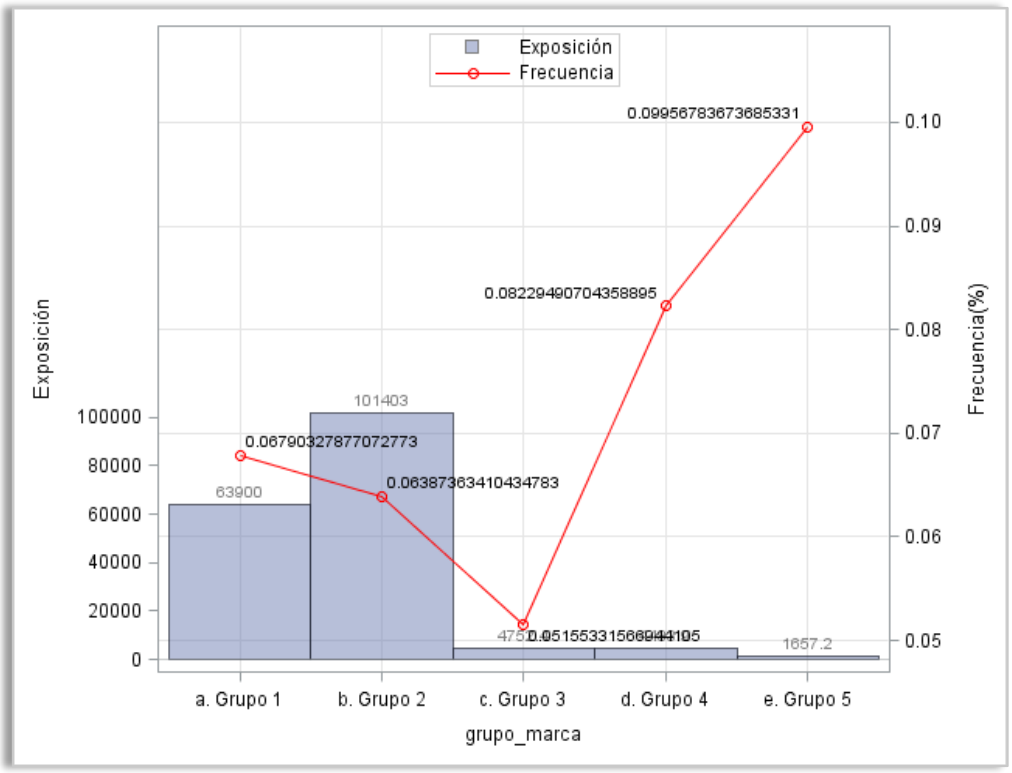


Gráfico elaborado por el autor.

Estado Civil

Figura 16.1.30. Tabla Frecuencia: Estado Civil.

estado_civil	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
a. Casado	420039	84.01	420039	84.01
b. Divorciado/Separado	12583	2.52	432622	86.52
c. Soltero	61083	12.22	493705	98.74
d. Otros	6295	1.26	500000	100.00

Tabla elaborada por el autor.

Gráfico 16.1.31. Gráfico Bivariante: Estado Civil.

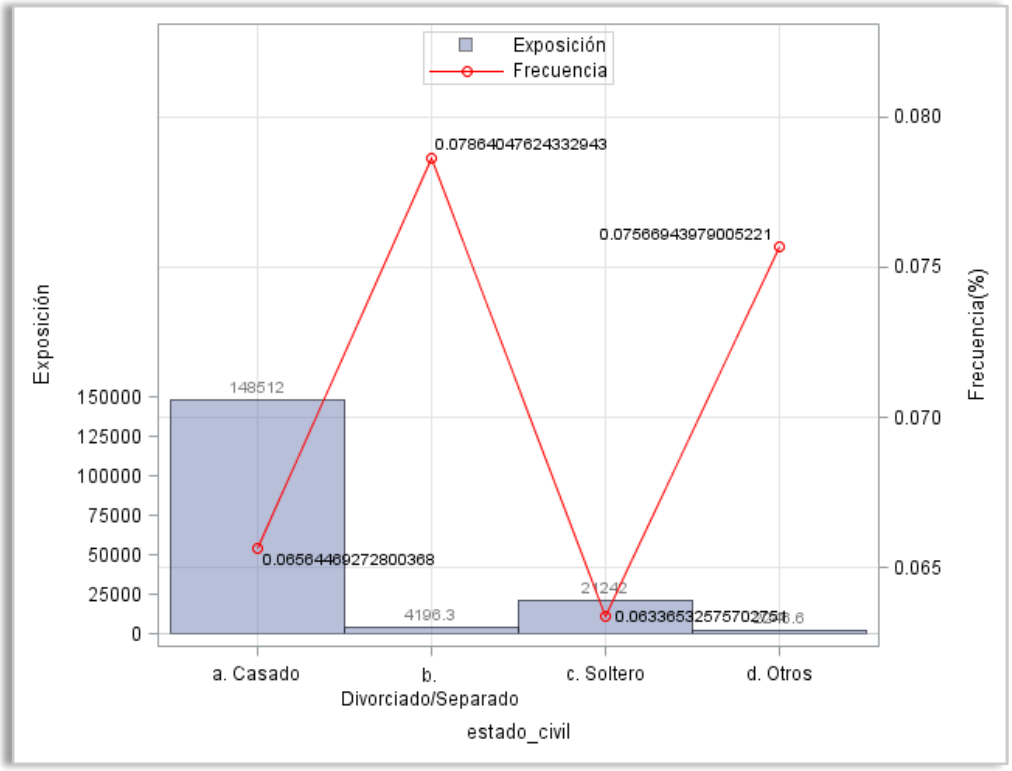


Gráfico elaborado por el autor.

Peso Sector Terciario

Figura 16.1.32. Tabla Frecuencia: Peso Sector Terciario.

peso_sector_terciario	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
01:low-51.13	79834	15.97	79834	15.97
02:51.13-67.59	180511	36.10	260345	52.07
03:67.59-82.35	186827	37.37	447172	89.43
04:82.35-high	52461	10.49	499633	99.93
05:Indeterminado	367	0.07	500000	100.00

Tabla elaborada por el autor.

Gráfico 16.1.33. Gráfico Bivariante: Peso Sector Terciario.

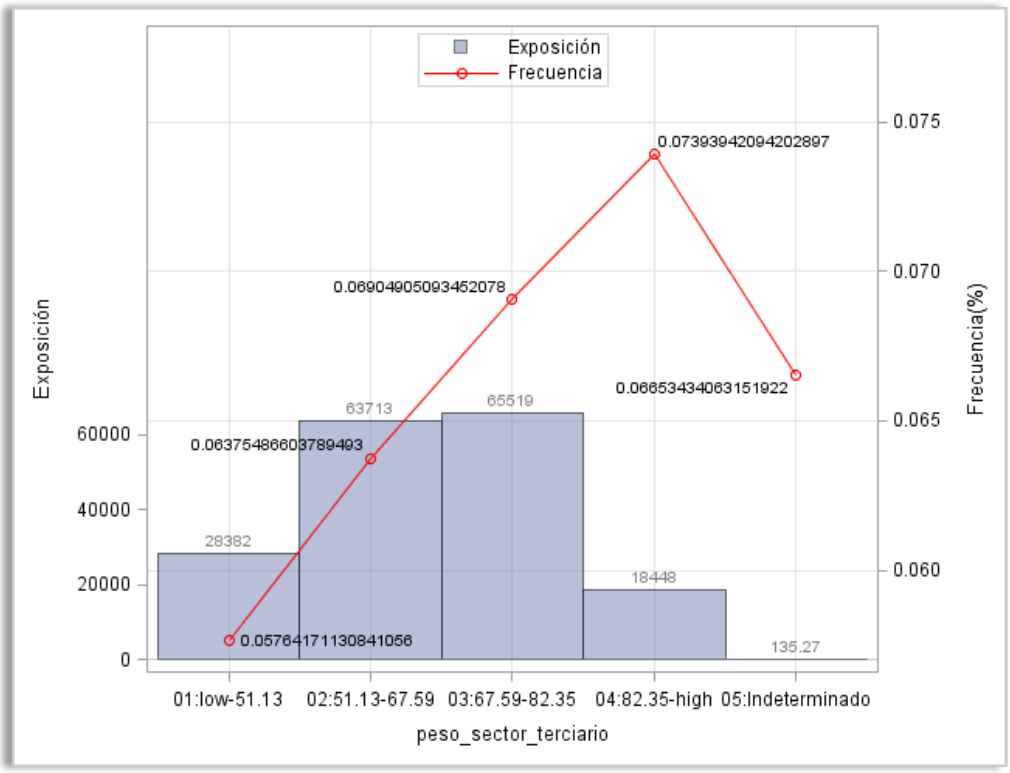


Gráfico elaborado por el autor.

Días Heladas

Figura 16.1.34. Tabla Frecuencia: Días Heladas.

dias_heladas	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
01:low-84.0115	185601	37.12	185601	37.12
02:84.0115-119.0414	109652	21.93	295253	59.05
03:119.0414-366.4339	84476	16.90	379729	75.95
04:366.4339-622.7705	35169	7.03	414898	82.98
05:622.7705-868.7585	42313	8.46	457211	91.44
06:868.7585-high	42373	8.47	499584	99.92
07:Indeterminado	416	0.08	500000	100.00

Tabla elaborada por el autor.

Gráfico 15.1.35. Gráfico Bivariante: Días Heladas.

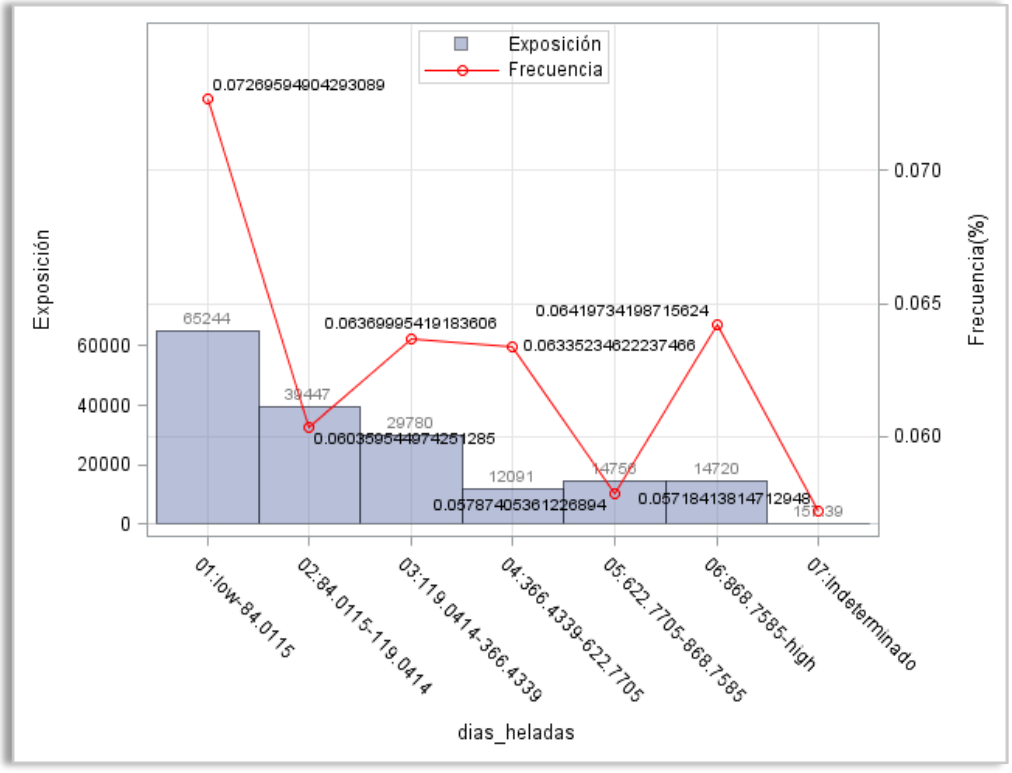


Gráfico elaborado por el autor.

Velocidad Media del Viento

Figura 16.1.36. Tabla Frecuencia: Velocidad Media del Viento.

velocid_media_viento	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
01:low-2.571127	34353	6.87	34353	6.87
02:2.571127-2.596619	36913	7.38	71266	14.25
03:2.596619-2.7195094	25028	5.01	96294	19.26
04:2.7195094-2.8228306	35144	7.03	131438	26.29
05:2.8228306-2.9032046	79982	16.00	211420	42.28
06:2.9032046-2.9159927	102114	20.42	313534	62.71
07:2.9159927-3.0506671	38945	7.79	352479	70.50
08:3.0506671-3.2695916	38584	7.72	391063	78.21
09:3.2695916-3.5119002	46532	9.31	437595	87.52
10:3.5119002-high	62038	12.41	499633	99.93
11:Indeterminado	367	0.07	500000	100.00

Tabla elaborada por el autor.

Gráfico 16.1.37. Gráfico Bivariante: Velocidad Media del Viento.

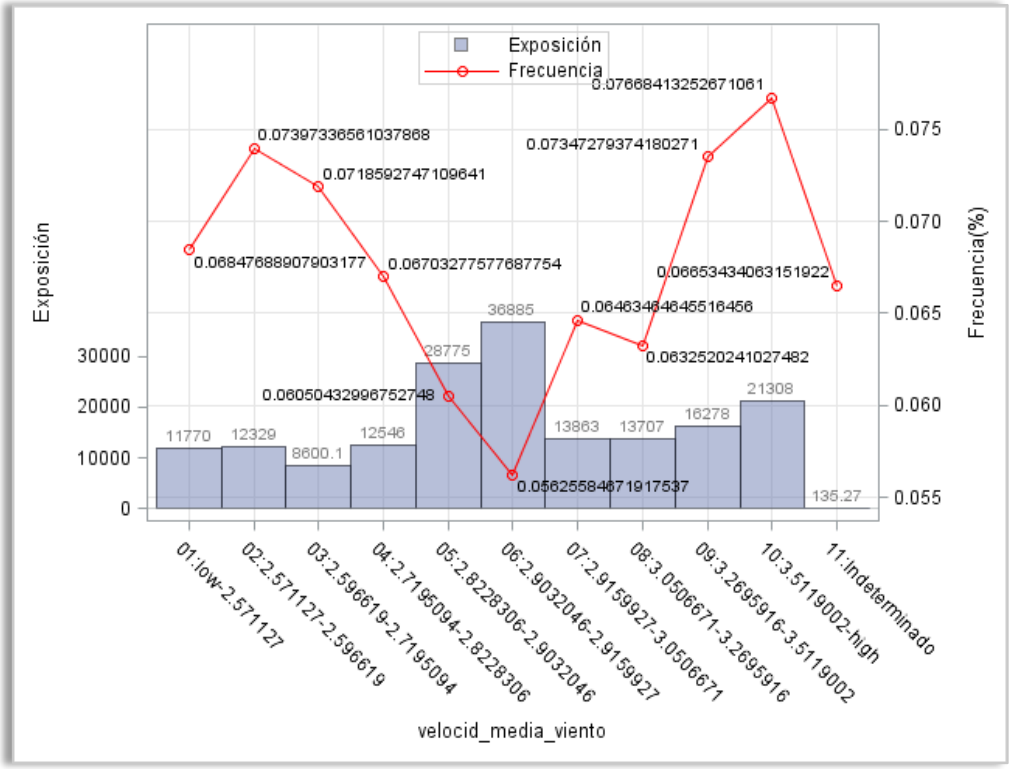


Gráfico elaborado por el autor.

16.2. Modelización Interna

Figura 16.2.1. Tramos Validación Cruzada *Elastic Net*

C. Num	Variables Seleccionadas	λ	Log(λ)
1	1	0.009966	-4.60858
2	2	0.009080	-4.70168
3	2	0.008274	-4.79464
4	3	0.007539	-4.88767
5	3	0.006869	-4.98074
6	5	0.006259	-5.07373
7	6	0.005703	-5.16676
8	6	0.005196	-5.25987
9	6	0.004734	-5.35298
10	8	0.004314	-5.44589
11	8	0.003931	-5.53886
12	10	0.003581	-5.63211
13	11	0.003263	-5.72511
14	11	0.002973	-5.81818
15	11	0.002709	-5.91118
16	13	0.002469	-6.00394
17	17	0.002249	-6.09727
18	18	0.002049	-6.19040
19	20	0.001867	-6.28342
20	22	0.001701	-6.37654
21	23	0.001550	-6.46950
22	27	0.001413	-6.56204
23	29	0.001287	-6.65544
24	32	0.001173	-6.74819
25	32	0.001069	-6.84103
26	35	0.000974	-6.93441
27	37	0.000887	-7.02744
28	44	0.000808	-7.12058
29	46	0.000737	-7.21360
30	49	0.000671	-7.30659
31	49	0.000612	-7.39960
32	51	0.000557	-7.49259
33	53	0.000508	-7.58562
34	53	0.000463	-7.67865
35	55	0.000422	-7.77169
36	58	0.000384	-7.86487
37	60	0.000350	-7.95786
38	61	0.000319	-8.05095
39	62	0.000291	-8.14391
40	63	0.000265	-8.23691
41	63	0.000241	-8.32988

42[29]	65	0.000220	-8.42279
43	65	0.000200	-8.51619
44	66	0.000182	-8.60931
45	66	0.000166	-8.70232
46	66	0.000152	-8.79492
47	66	0.000138	-8.88826
48	66	0.000126	-8.98082
49	67	0.000115	-9.07406
50	67	0.000104	-9.16728
51	68	0.000095	-9.26027
52	69	0.000087	-9.35329
53	69	0.000079	-9.44632
54	69	0.000072	-9.53940
55	70	0.000066	-9.63239
56	70	0.000060	-9.72551
57	70	0.000054	-9.81841
58	70	0.000050	-9.91152
59	71	0.000045	-10.00463
60	71	0.000041	-10.09756
61	71	0.000038	-10.19064
62	72	0.000034	-10.28358
63	72	0.000031	-10.37670
64	72	0.000028	-10.46983
65	72	0.000026	-10.56281
66	73	0.000024	-10.65596
67	73	0.000021	-10.74885
68	73	0.000020	-10.84202
69	73	0.000018	-10.93463
70	73	0.000016	-11.02803
71	73	0.000015	-11.12088
72	73	0.000013	-11.214303

Tabla elaborada por el autor.

[29] La fila marcada en rojo denota el λ^{opt} para un modelo Elastic Net con un $\alpha = 0.4$. Este λ selecciona 68 variables, necesitando un λ mayor con el que desechar un mayor número de variables intrascendentes en el modelo.

16.3. Modelización de Variables Externas

Tabla 15.3.1. Coeficientes *Ridge Regression* vs *GLM*.

C. Num	Variables	Dummy	Coef. RR	Coef. GLM	Dif. %
1	Intercepto		-4.2281	-4.4841	-5.71%
2	Peso Potencia	<=10.47	-0.0334	-0.0608	-45.07%
3	Peso Potencia	>13.44	0.0347	0.0470	-27.10%
4	Peso Vehículo	<=979.6	-0.1185	-0.1219	-2.79%
5	Peso Vehículo	>1649.56	0.1264	0.1591	-20.55%
6	Velocidad	<=167	-0.0313	-0.0533	-41.28%
7	Velocidad	(187;209]	-0.0344	-0.0393	-12.47%
8	Velocidad	>209	-0.0648	-0.0903	-28.24%
9	Ant. Carn. Ocas.	[0;7]	0.3322	0.3556	-6.58%
10	Ant. Carn. Ocas.	[8;38]	-0.1176	-0.1563	-24.76%
11	Ant. Carn. Ocas.	[39;57]	0.3369	0.4191	-19.61%
12	Forma de Pago	Semestral	0.1401	0.1618	-13.41%
13	Forma de Pago	Trimestral-Mensual	0.1963	0.1852	5.99%
14	Motor	Gasolina	-0.0993	-0.081	22.59%
15	Motor	Otros/Desconocido	0.0305	-0.1736	-117.57%
16	Bonificación	Bonus Bajo	0.4854	0.5572	-12.89%
17	Bonificación	Bonus Medio	0.2534	0.3250	-22.03%
18	Puertas	3	-0.0367	-0.0481	-23.70%
19	Puertas	Indefinido	-0.8858	-15.6496	-94.34%
20	Unidad Familiar	Sí	0.0806	0.0477	68.97%
21	Nº Auto Familia	0	0.2881	-0.0748	-485.16%
22	Nº Auto Familia	Más de 2	0.0797	0.0950	-16.11%
23	Nº Auto Familia	Indefinido	0	-15.6794	-100.00%
24	Cond. Otros Veh.	Sí	-0.1061	-0.0680	56.03%
25	Morosidad	A-B	-0.0502	-0.0012	4083.33%
26	Morosidad	D-E	0.0187	0.0249	-24.90%

27	Morosidad	F	0.2604	0.2918	-10.76%
28	Morosidad	Desconocido	0.1415	0.1538	-8.00%
29	Grupo Marca	1	0.0507	0.0495	2.42%
30	Grupo Marca	3	-0.1331	-0.1392	-4.38%
31	Grupo Marca	4	0.2030	0.1837	10.51%
32	Grupo Marca	5	0.2890	0.2919	-0.99%
33	Estado Civil	Divorciado/Separado	0.0737	0.1353	-45.53%
34	Estado Civil	Soltero	-0.0488	-0.084	-41.90%
35	Estado Civil	Otros	0.0247	0.1744	-85.84%
36	Tasa Paro	Low-6.27	-0.0635	-0.0454	39.87%
37	Tasa Paro	10.69-13.29	-0.0193	-0.0071	171.83%
38	Tasa Paro	13.29-17.57	-0.0155	-0.0249	-37.75%
39	Tasa Paro	17.57-22.86	0.0235	-0.0079	-397.47%
40	Tasa Paro	22.86-high	0.0037	-0.0089	-141.57%
41	Tasa Paro	Indeterminado	0.0062	0	0%
42	Ing. Medios	Low-1810.69	-0.0539	-0.0474	13.71%
43	Ing. Medios	2011.99-2183.58	-0.0271	-0.0094	188.30%
44	Ing. Medios	2183.58-2362.51	0.0106	0.0435	-75.63%
45	Ing. Medios	2362.51-high	-0.0057	0.0256	-122.27%
46	Ing. Medios	Indeterminado	0.0050	16.3634	-99.97%
47	Días Helada	84.01-119.04	0.0231	0.0339	-31.86%
48	Días Helada	119.04-366.43	-0.0227	-0.129	-82.40%
49	Días Helada	366.43-622.77	-0.0375	-0.3326	-88.73%
50	Días Helada	622.77-868.75	-0.0515	-0.3785	-86.39%
51	Días Helada	868.75-high	-0.0456	-0.3846	-88.14%
52	Días Helada	Indeterminado	-0.3480	-16.1394	-97.84%
53	Prec. Más 40 ml	Low-5.45	0.0352	0.1468	-76.02%
54	Prec. Más 40 ml	5.45-12.00	0.0366	0.1905	-80.79%
55	Prec. Más 40 ml	12.00-25.30	-0.0092	0.1139	-108.08%
56	Prec. Más 40 ml	25.30-34.52	0.0388	0.1499	-74.12%

57	Prec. Más 40 ml	34.52-44.87	-0.0086	0.045	-119.11%
58	Prec. Más 40 ml	52.28-high	-0.0055	0.0615	-108.94%
59	Prec. Más 40 ml	Indeterminado	0.0034	0	0%
60	Insol Media	Low-5.37	0.0616	-1.5146	-104.07%
61	Insol Media	5.37-6.38	-0.0155	-1.5912	-99.03%
62	Insol Media	7.27-7.40	-0.0045	0.1069	-104.21%
63	Insol Media	7.40-7.59	-0.0467	0.2001	-123.34%
64	Insol Media	7.59-7.80	0.0118	0.2876	-95.90%
65	Insol Media	7.80-7.88	0.0612	0.3056	-79.97%
66	Insol Media	7.88-high	0.0001	0.2219	-99.95%
67	Insol Media	Indeterminado	0.0033	0	0%
68	Insolación	Low-50.46	0.0520	1.4762	-96.48%
69	Insolación	58.13-59.15	0.0082	-0.033	-124.85%
70	Insolación	59.15-62.80	0.0007	-0.2013	-100.35%
71	Insolación	62.80-63.52	-0.0370	-0.2202	-83.20%
72	Insolación	63.52-high	0.0366	-0.1493	-124.51%
73	Insolación	Indeterminado	0.0051	0	0%
74	D Prec sup 10 ml	Low-281.63	0.0112	0.0014	700.00%
75	D Prec sup 10 ml	281.63-298.72	0.0241	0.0073	230.14%
76	D Prec sup 10 ml	298.72-371	0.0214	0.0285	-24.91%
77	D Prec sup 10 ml	699.05-high	0.0662	0.0971	-31.82%
78	D Prec sup 10 ml	Indeterminado	0.0095	0	0%
79	Días Lluvia	Low-1524	0.0258	0.0294	-12.24%
80	Días Lluvia	1725-1725.43	0.0321	0.0682	-52.93%
81	Días Lluvia	1725.43-1901.27	-0.0044	-0.0521	-91.55%
82	Días Lluvia	2322.26-3326.32	0.0496	-0.0168	-395.24%
83	Días Lluvia	3326.32-high	-0.0143	0.0059	-342.37%
84	Días Lluvia	Indeterminado	0.0161	0	0%
85	Peso Sect Terc	Low-51.13	-0.0673	-0.1151	-41.53%
86	Peso Sect Terc	51.13-67.59	-0.0223	-0.034	-34.41%

87	Peso Sect Terc	82.35-high	0.0162	-0.0139	-216.55%
88	Peso Sect Terc	Indeterminado	0.0228	0	0%
89	T max men 5º día	Low-3.45	-0.0010	-0.0564	-98.23%
90	T max men 5º día	3.45-9	0.0429	0.034	26.18%
91	T max men 5º día	9-15.59	0.0217	0.0013	1569.23%
92	T max men 5º día	15.59-27.63	0.0149	-0.0531	-128.06%
93	T max men 5º día	66.65-77.99	-0.0077	0.0237	-132.49%
94	T max men 5º día	77.99-318.51	-0.0099	-0.0642	-84.58%
95	T max men 5º día	318.51-high	0.0341	0.3123	-89.08%
96	T max men 5º día	Indeterminado	0.0279	0	0%
97	Días Precip	Low-60.58	-0.0092	0.1307	-107.04%
98	Días Precip	60.58-71.15	0.0066	0.097	-93.20%
99	Días Precip	82.16-91.42	0.0085	0.015	-43.33%
100	Días Precip	91.42-114.67	-0.0489	-0.1186	-58.77%
101	Días Precip	114.67-high	0.0331	0.0016	1968.75%
102	Días Precip	Indeterminado	0.0300	0	0%
103	Temp Inf Max	Low-11.60	0.0175	0.1337	-86.91%
104	Temp Inf Max	11.60-12.20	-0.0081	0.0652	-112.42%
105	Temp Inf Max	12.20-12.75	0.0318	0.062	-48.71%
106	Temp Inf Max	12.75-13.90	0.0049	0.0155	-68.39%
107	Temp Inf Max	14.89-15.75	-0.0055	0.0377	-114.59%
108	Temp Inf Max	15.75-16.87	0.0259	0.1372	-81.12%
109	Temp Inf Max	16.87-17.70	0.0156	0.1428	-89.08%
110	Temp Inf Max	17.70-high	0.0223	0.2347	-90.50%
111	Temp Inf Max	Indeterminado	0.0289	0	0%
112	Temp Mínim	Low-0.55	0.0140	-0.0464	-130.17%
113	Temp Mínim	0.55-2.36	0.0122	-0.2607	-104.68%
114	Temp Mínim	2.36-3.98	-0.0541	-0.0933	-42.02%
115	Temp Mínim	3.98-4.85	-0.0661	-0.0257	157.20%
116	Temp Mínim	4.85-5.56	0.0595	-0.0526	-213.12%

117	Temp Mínim	6.43-7.12	-0.0300	0.0017	-1864.71%
118	Temp Mínim	7.12-8.39	0.0274	0.0555	-50.63%
119	Temp Mínim	8.39-high	0.0039	0.0746	-94.77%
120	Temp Mínim	Indeterminado	0.0256	0	0%
121	Temp Med Max	Low-17.38	0.0433	0.0212	104.25%
122	Temp Med Max	17.38-17.99	0.0328	0.0378	-13.23%
123	Temp Med Max	17.99-19.03	-0.0266	-0.1562	-82.97%
124	Temp Med Max	19.03-19.51	0.0551	0.0792	-30.43%
125	Temp Med Max	20.50-21.76	0.0049	-0.0468	-110.47%
126	Temp Med Max	21.76-22.38	0.0288	-0.0598	-148.16%
127	Temp Med Max	22.38-22.99	0.0066	-0.2304	-102.86%
128	Temp Med Max	22.99-high	0.0134	-0.1406	-109.53%
129	Temp Med Max	Indeterminado	0.0210	0	0%
130	Temp Med Min	Low-6.43	-0.0488	0.2056	-123.74%
131	Temp Med Min	6.43-7.57	0.0681	0.362	-81.19%
132	Temp Med Min	7.57-8.04	-0.0239	0.3579	-106.68%
133	Temp Med Min	8.04-9.32	-0.0635	0.0591	-207.45%
134	Temp Med Min	9.32-10.07	0.0174	0.0124	40.32%
135	Temp Med Min	10.07-10.59	0.0335	0.1536	-78.19%
136	Temp Med Min	11.30-11.92	0.0051	-0.0228	-122.37%
137	Temp Med Min	11.92-12.89	0.0025	-0.1148	-102.18%
138	Temp Med Min	12.89-high	-0.0105	-0.1922	-94.54%
139	Temp Med Min	Indeterminado	0.0166	0	0%
140	Vel. Med Vient	Low-2.57	0.0619	0.1644	-62.35%
141	Vel. Med Vient	2.57-2.59	0.0281	0.0433	-35.10%
142	Vel. Med Vient	2.59-2.71	0.0446	0.1947	-77.09%
143	Vel. Med Vient	2.71-2.82	0.0645	0.1662	-61.19%
144	Vel. Med Vient	2.82-2.90	0.0098	0.0663	-85.22%
145	Vel. Med Vient	2.91-3.05	0.0553	0.0976	-43.34%
146	Vel. Med Vient	3.05-3.26	-0.0771	-0.0767	0.52%

147	Vel. Med Vient	3.26-3.51	-0.0249	-0.0496	-49.80%
148	Vel. Med Vient	3.51-high	0.0186	-0.0148	-225.68%
149	Vel. Med Vient	Indeterminado	0.0128	0	0%
150	P. Atm. Niv. Mar	Low-1016.44	-0.0096	-0.0009	966.67%
151	P. Atm. Niv. Mar	1016.54-1016.78	0.0134	0.0412	-67.48%
152	P. Atm. Niv. Mar	1016.78-1016.85	-0.0091	0.0263	-134.60%
153	P. Atm. Niv. Mar	1016.85-1017.00	0.0187	0.0922	-79.72%
154	P. Atm. Niv. Mar	1017.00-1017.29	-0.0924	-0.0843	9.61%
155	P. Atm. Niv. Mar	1017.29-1017.50	0.0100	0.0829	-87.94%
156	P. Atm. Niv. Mar	1017.50-1017.86	0.0870	0.042	107.14%
157	P. Atm. Niv. Mar	1017.86-1018.18	0.0455	0.0549	-17.12%
158	P. Atm. Niv. Mar	1018.18-high	0.0083	0.106	-92.17%
159	P. Atm. Niv. Mar	Indeterminado	0.0326	0	0%
160	Cuenca Nival	Si	-0.1984	-0.0516	284.50%
161	Cuenca Nival	Indeterminado	0.0086	0	0%
162	Zona Inundable	Si	0.0291	0.0414	-29.71%
163	Zona Inundable	Indeterminado	0.0079	0	0%
164	Temp Media	Low-12.31	0.0145	-0.0811	-117.88%
165	Temp Media	12.31-13.42	-0.0582	-0.144	-59.58%
166	Temp Media	13.42-14.66	0.0905	-0.1721	-152.59%
167	Temp Media	14.66-15.16	0.0695	-0.0654	-206.27%
168	Temp Media	15.16-15.35	-0.0105	-0.0944	-88.88%
169	Temp Media	15.84-17.24	0.0099	0.0706	-85.98%
170	Temp Media	17.24-18.32	0.0081	-0.0914	-108.86%
171	Temp Media	18.32-18.69	0.0327	-0.0156	-309.62%
172	Temp Media	18.69-high	0.0079	-0.1735	-104.55%
173	Temp Media	Indeterminado	0.0074	0	0%
174	Precip total	Low-378.06	-0.0621	-0.2132	-70.87%
175	Precip total	378.06-409.62	-0.0747	-0.1906	-60.81%
176	Precip total	409.62-420.07	0.0112	-0.1198	-109.35%

177	Precip total	420.07-438.73	-0.0099	-0.0744	-86.69%
178	Precip total	438.73-488.78	-0.0047	-0.1306	-96.40%
179	Precip total	488.78-566.86	0.0281	-0.0346	-181.21%
180	Precip total	566.86-647.35	-0.0067	-0.0146	-54.11%
181	Precip total	709.95-1076.72	0.0424	0.1537	-72.41%
182	Precip total	1076.72-high	0.0886	0.4229	-79.05%
183	Precip total	Indeterminado	0.0070	0	0%
184	Prec. Max.	Low-11.97	0.0372	-0.0473	-178.65%
185	Prec. Max.	11.97-12.68	-0.0223	-0.0777	-71.30%
186	Prec. Max.	12.68-13.30	0.0581	-0.0175	-432.00%
187	Prec. Max.	13.30-13.99	0.0138	-0.0441	-131.29%
188	Prec. Max.	13.99-16.44	0.0222	-0.0754	-129.44%
189	Prec. Max.	16.44-18.24	-0.0457	-0.1661	-72.49%
190	Prec. Max.	18.24-20.25	0.0233	-0.0861	-127.06%
191	Prec. Max.	20.25-22.74	-0.0526	-0.1412	-62.75%
192	Prec. Max.	24.31-high	0.0508	0.0009	5544.44%
193	Prec. Max.	Indeterminado	0.0068	0	0%
194	Congelación	Low-3.01	0.0221	0.1199	-81.57%
195	Congelación	3.01-8.12	-0.0005	0.0196	-102.55%
196	Congelación	8.12-14.86	0.0294	0.0945	-68.89%
197	Congelación	14.86-25.98	0.0609	0.1454	-58.12%
198	Congelación	25.98-48.37	0.0113	0.0354	-68.08%
199	Congelación	62.89-71.25	-0.0394	-0.0345	14.20%
200	Congelación	71.25-161.00	-0.0193	0.0516	-137.40%
201	Congelación	161.00-305.67	0.0427	0.1847	-76.88%
202	Congelación	305.67-high	0.0305	-0.0818	-137.29%
203	Congelación	Indeterminado	0.0069	0	0%
204	Rac Vient + 55km	Low- 235.52	-0.0026	-0.0824	-96.84%
205	Rac Vient + 55km	235.52-366.16	-0.0008	-0.0377	-97.88%
206	Rac Vient + 55km	366.16-431.03	0.0613	0.0234	161.97%

207	Rac Vient + 55km	431.03-506.94	-0.0380	-0.0854	-55.50%
208	Rac Vient + 55km	506.94-588.71	0.0495	-0.0082	-703.66%
209	Rac Vient + 55km	588.71-704.50	-0.0004	0.0248	-101.61%
210	Rac Vient + 55km	704.50-809.52	0.0131	0.0359	-63.51%
211	Rac Vient + 55km	880.83-1140.01	-0.0235	0.0404	-158.17%
212	Rac Vient + 55km	1140.01-high	0.0087	0.0621	-85.99%
213	Rac Vient + 55km	Indeterminado	0.0070	0	0%
214	Rac Vient + 91km	Low-1.72	0.0120	0.1707	-92.97%
215	Rac Vient + 91km	1.72-3.50	-0.0502	0.0438	-214.61%
216	Rac Vient + 91km	3.50-5	0.0666	0.162	-58.89%
217	Rac Vient + 91km	5-8	-0.0098	0.0426	-123.00%
218	Rac Vient + 91km	8-10.12	0.0843	0.1578	-46.58%
219	Rac Vient + 91km	10.12-14.85	0.0045	0.0025	80.00%
220	Rac Vient + 91km	17.95-25.99	0.0681	0.1064	-36.00%
221	Rac Vient + 91km	25.99-40.99	-0.0450	0.067	-167.16%
222	Rac Vient + 91km	40.99-high	0.0056	0.1192	-95.30%
223	Rac Vient + 91km	Indeterminado	-0.0068	0	0%
225	Días Gran	Low-8.26	0.0442	0.0215	105.58%
226	Días Gran	8.26-12.73	0.0092	-0.0026	-453.85%
227	Días Gran	12.73-15.99	-0.0046	-0.0025	84.00%
228	Días Gran	15.99-22.29	0.0685	0.0261	162.45%
229	Días Gran	30.00-36.99	0.0310	-0.0027	-1248.15%
230	Días Gran	36.99-43.99	0.0412	0.063	-34.60%
231	Días Gran	43.99-72.62	-0.0479	-0.0235	103.83%
232	Días Gran	72.62-94.64	-0.1091	-0.2441	-55.31%
233	Días Gran	94.64-high	-0.0824	-0.3877	-78.75%
234	Días Gran	Indeterminado	0.0067	0	0%
235	Rur_urb	1.95-7.99	-0.0549	-0.0258	112.79%
236	Rur_urb	7.99-25.42	-0.0096	0.0211	-145.50%
237	Rur_urb	25.42-64.17	-0.0051	-0.0095	-46.32%

238	Rur_urb	64.17-123.64	0.0271	0.0381	-28.87%
239	Rur_urb	123.64-291.33	-0.1165	-0.0654	78.13%
240	Rur_urb	291.33-487.48	0.0607	0.2152	-71.79%
241	Rur_urb	487.48-798.07	0.0272	0.1809	-84.96%
242	Rur_urb	798.07-1116.70	0.0286	0.2535	-88.72%
243	Rur_urb	1116.70-high	0.0040	0.2244	-98.22%
244	Rur_urb	Indeterminado	0.0066	0	0%

Tabla elaborada por el autor.

Tabla 15.3.2. Coeficientes *Lasso* vs *GLM*.

C. Num	Variables	Dummy	Coef. Lasso	Coef. RR	Coef. GLM	Dif. %
1	Intercepto		-4.2177	-4.2281	-4.4841	-5.71%
2	Peso Potencia	<=10.47	-0.0232	-0.0334	-0.0608	-45.07%
3	Peso Potencia	>13.44	0.0150	0.0347	0.0470	-27.10%
4	Peso Vehículo	<=979.6	-0.1269	-0.1185	-0.1219	-2.79%
5	Peso Vehículo	>1649.56	0.1350	0.1264	0.1591	-20.55%
6	Velocidad	<=167	.	-0.0313	-0.0533	-41.28%
7	Velocidad	(187;209]	-0.0266	-0.0344	-0.0393	-12.47%
8	Velocidad	>209	-0.0536	-0.0648	-0.0903	-28.24%
9	Ant. Carn. Ocas.	[0;7]	0.3522	0.3322	0.3556	-6.58%
10	Ant. Carn. Ocas.	[8;38]	-0.0870	-0.1176	-0.1563	-24.76%
11	Ant. Carn. Ocas.	[39;57]	0.1827	0.3369	0.4191	-19.61%
12	Forma de Pago	Semestral	0.1474	0.1401	0.1618	-13.41%
13	Forma de Pago	Trimestral-Mensual	0.1892	0.1963	0.1852	5.99%
14	Motor	Gasolina	-0.1147	-0.0993	-0.081	22.59%
15	Motor	Otros/Desconocido	.	0.0305	-0.1736	-117.57%
16	Bonificación	Bonus Bajo	0.5437	0.4854	0.5572	-12.89%
17	Bonificación	Bonus Medio	0.2964	0.2534	0.3250	-22.03%
18	Puertas	3	-0.0318	-0.0367	-0.0481	-23.70%
19	Puertas	Indefinido	.	-0.8858	-15.6496	-94.34%
20	Unidad Familiar	Sí	0.0977	0.0806	0.0477	68.97%
21	Nº Auto Familia	0	0.3072	0.2881	-0.0748	-485.16%
22	Nº Auto Familia	Más de 2	0.0627	0.0797	0.0950	-16.11%
23	Nº Auto Familia	Indefinido	.	0	-15.6794	-100.00%
24	Cond. Otros Veh.	Sí	-0.1136	-0.1061	-0.0680	56.03%
25	Morosidad	A-B	-0.0377	-0.0502	-0.0012	4083.33%
26	Morosidad	D-E	0.0129	0.0187	0.0249	-24.90%
27	Morosidad	F	0.2764	0.2604	0.2918	-10.76%
28	Morosidad	Desconocido	0.1474	0.1415	0.1538	-8.00%

29	Grupo Marca	1	0.0495	0.0507	0.0495	2.42%
30	Grupo Marca	3	-0.1335	-0.1331	-0.1392	-4.38%
31	Grupo Marca	4	0.1994	0.2030	0.1837	10.51%
32	Grupo Marca	5	0.2807	0.2890	0.2919	-0.99%
33	Estado Civil	Divorciado/Separado	0.0441	0.0737	0.1353	-45.53%
34	Estado Civil	Soltero	-0.0422	-0.0488	-0.084	-41.90%
35	Estado Civil	Otros	.	0.0247	0.1744	-85.84%
36	Tasa Paro	Low-6.27	-0.0500	-0.0635	-0.0454	39.87%
37	Tasa Paro	10.69-13.29	.	-0.0193	-0.0071	171.83%
38	Tasa Paro	13.29-17.57	.	-0.0155	-0.0249	-37.75%
39	Tasa Paro	17.57-22.86	0.0187	0.0235	-0.0079	-397.47%
40	Tasa Paro	22.86-high	.	0.0037	-0.0089	-141.57%
41	Tasa Paro	Indeterminado	.	0.0062	0	0%
42	Ing. Medios	Low-1810.69	-0.0371	-0.0539	-0.0474	13.71%
43	Ing. Medios	2011.99-2183.58	-0.0172	-0.0271	-0.0094	188.30%
44	Ing. Medios	2183.58-2362.51	.	0.0106	0.0435	-75.63%
45	Ing. Medios	2362.51-high	.	-0.0057	0.0256	-122.27%
46	Ing. Medios	Indeterminado	.	0.0050	16.3634	-99.97%
47	Días Helada	84.01-119.04	.	0.0231	0.0339	-31.86%
48	Días Helada	119.04-366.43	-0.0019	-0.0227	-0.129	-82.40%
49	Días Helada	366.43-622.77	.	-0.0375	-0.3326	-88.73%
50	Días Helada	622.77-868.75	.	-0.0515	-0.3785	-86.39%
51	Días Helada	868.75-high	.	-0.0456	-0.3846	-88.14%
52	Días Helada	Indeterminado	.	-0.3480	-16.1394	-97.84%
53	Prec. Más 40 ml	Low-5.45	.	0.0352	0.1468	-76.02%
54	Prec. Más 40 ml	5.45-12.00	.	0.0366	0.1905	-80.79%
55	Prec. Más 40 ml	12.00-25.30	.	-0.0092	0.1139	-108.08%
56	Prec. Más 40 ml	25.30-34.52	0.0190	0.0388	0.1499	-74.12%
57	Prec. Más 40 ml	34.52-44.87	.	-0.0086	0.045	-119.11%
58	Prec. Más 40 ml	52.28-high	.	-0.0055	0.0615	-108.94%

59	Prec. Más 40 ml	Indeterminado	.	0.0034	0	0%
60	Insol Media	Low-5.37	0.0484	0.0616	-1.5146	-104.07%
61	Insol Media	5.37-6.38	.	-0.0155	-1.5912	-99.03%
62	Insol Media	7.27-7.40	.	-0.0045	0.1069	-104.21%
63	Insol Media	7.40-7.59	.	-0.0467	0.2001	-123.34%
64	Insol Media	7.59-7.80	.	0.0118	0.2876	-95.90%
65	Insol Media	7.80-7.88	0.0199	0.0612	0.3056	-79.97%
66	Insol Media	7.88-high	.	0.0001	0.2219	-99.95%
67	Insol Media	Indeterminado	.	0.0033	0	0%
68	Insolación	Low-50.46	0.0899	0.0520	1.4762	-96.48%
69	Insolación	58.13-59.15	.	0.0082	-0.033	-124.85%
70	Insolación	59.15-62.80	.	0.0007	-0.2013	-100.35%
71	Insolación	62.80-63.52	.	-0.0370	-0.2202	-83.20%
72	Insolación	63.52-high	0.0164	0.0366	-0.1493	-124.51%
73	Insolación	Indeterminado	.	0.0051	0	0%
74	D Prec sup 10 ml	Low-281.63	.	0.0112	0.0014	700.00%
75	D Prec sup 10 ml	281.63-298.72	.	0.0241	0.0073	230.14%
76	D Prec sup 10 ml	298.72-371	.	0.0214	0.0285	-24.91%
77	D Prec sup 10 ml	699.05-high	.	0.0662	0.0971	-31.82%
78	D Prec sup 10 ml	Indeterminado	.	0.0095	0	0%
79	Días Lluvia	Low-1524	0.0105	0.0258	0.0294	-12.24%
80	Días Lluvia	1725-1725.43	0.0154	0.0321	0.0682	-52.93%
81	Días Lluvia	1725.43-1901.27	.	-0.0044	-0.0521	-91.55%
82	Días Lluvia	2322.26-3326.32	0.0025	0.0496	-0.0168	-395.24%
83	Días Lluvia	3326.32-high	.	-0.0143	0.0059	-342.37%
84	Días Lluvia	Indeterminado	.	0.0161	0	0%
85	Peso Sect Terc	Low-51.13	-0.0542	-0.0673	-0.1151	-41.53%
86	Peso Sect Terc	51.13-67.59	-0.0119	-0.0223	-0.034	-34.41%
87	Peso Sect Terc	82.35-high	0.0085	0.0162	-0.0139	-216.55%
88	Peso Sect Terc	Indeterminado	.	0.0228	0	0%

89	T max men 5º día	Low-3.45	.	-0.0010	-0.0564	-98.23%
90	T max men 5º día	3.45-9	0.0206	0.0429	0.034	26.18%
91	T max men 5º día	9-15.59	.	0.0217	0.0013	1569.23%
92	T max men 5º día	15.59-27.63	.	0.0149	-0.0531	-128.06%
93	T max men 5º día	66.65-77.99	.	-0.0077	0.0237	-132.49%
94	T max men 5º día	77.99-318.51	.	-0.0099	-0.0642	-84.58%
95	T max men 5º día	318.51-high	.	0.0341	0.3123	-89.08%
96	T max men 5º día	Indeterminado	.	0.0279	0	0%
97	Días Precip	Low-60.58	.	-0.0092	0.1307	-107.04%
98	Días Precip	60.58-71.15	.	0.0066	0.097	-93.20%
99	Días Precip	82.16-91.42	.	0.0085	0.015	-43.33%
100	Días Precip	91.42-114.67	-0.0348	-0.0489	-0.1186	-58.77%
101	Días Precip	114.67-high	.	0.0331	0.0016	1968.75%
102	Días Precip	Indeterminado	.	0.0300	0	0%
103	Temp Inf Max	Low-11.60	.	0.0175	0.1337	-86.91%
104	Temp Inf Max	11.60-12.20	.	-0.0081	0.0652	-112.42%
105	Temp Inf Max	12.20-12.75	0.0106	0.0318	0.062	-48.71%
106	Temp Inf Max	12.75-13.90	.	0.0049	0.0155	-68.39%
107	Temp Inf Max	14.89-15.75	.	-0.0055	0.0377	-114.59%
108	Temp Inf Max	15.75-16.87	.	0.0259	0.1372	-81.12%
109	Temp Inf Max	16.87-17.70	.	0.0156	0.1428	-89.08%
110	Temp Inf Max	17.70-high	.	0.0223	0.2347	-90.50%
111	Temp Inf Max	Indeterminado	.	0.0289	0	0%
112	Temp Mínim	Low-0.55	.	0.0140	-0.0464	-130.17%
113	Temp Mínim	0.55-2.36	.	0.0122	-0.2607	-104.68%
114	Temp Mínim	2.36-3.98	-0.0133	-0.0541	-0.0933	-42.02%
115	Temp Mínim	3.98-4.85	-0.0407	-0.0661	-0.0257	157.20%
116	Temp Mínim	4.85-5.56	0.0683	0.0595	-0.0526	-213.12%
117	Temp Mínim	6.43-7.12	-0.0342	-0.0300	0.0017	-1864.71%
118	Temp Mínim	7.12-8.39	0.0239	0.0274	0.0555	-50.63%

119	Temp Mínim	8.39-high	.	0.0039	0.0746	-94.77%
120	Temp Mínim	Indeterminado	.	0.0256	0	0%
121	Temp Med Max	Low-17.38	0.0063	0.0433	0.0212	104.25%
122	Temp Med Max	17.38-17.99	.	0.0328	0.0378	-13.23%
123	Temp Med Max	17.99-19.03	.	-0.0266	-0.1562	-82.97%
124	Temp Med Max	19.03-19.51	0.0269	0.0551	0.0792	-30.43%
125	Temp Med Max	20.50-21.76	.	0.0049	-0.0468	-110.47%
126	Temp Med Max	21.76-22.38	0.0213	0.0288	-0.0598	-148.16%
127	Temp Med Max	22.38-22.99	.	0.0066	-0.2304	-102.86%
128	Temp Med Max	22.99-high	.	0.0134	-0.1406	-109.53%
129	Temp Med Max	Indeterminado	.	0.0210	0	0%
130	Temp Med Min	Low-6.43	.	-0.0488	0.2056	-123.74%
131	Temp Med Min	6.43-7.57	0.0919	0.0681	0.362	-81.19%
132	Temp Med Min	7.57-8.04	.	-0.0239	0.3579	-106.68%
133	Temp Med Min	8.04-9.32	-0.0599	-0.0635	0.0591	-207.45%
134	Temp Med Min	9.32-10.07	.	0.0174	0.0124	40.32%
135	Temp Med Min	10.07-10.59	.	0.0335	0.1536	-78.19%
136	Temp Med Min	11.30-11.92	.	0.0051	-0.0228	-122.37%
137	Temp Med Min	11.92-12.89	.	0.0025	-0.1148	-102.18%
138	Temp Med Min	12.89-high	.	-0.0105	-0.1922	-94.54%
139	Temp Med Min	Indeterminado	.	0.0166	0	0%
140	Vel. Med Vient	Low-2.57	0.0054	0.0619	0.1644	-62.35%
141	Vel. Med Vient	2.57-2.59	.	0.0281	0.0433	-35.10%
142	Vel. Med Vient	2.59-2.71	.	0.0446	0.1947	-77.09%
143	Vel. Med Vient	2.71-2.82	0.0238	0.0645	0.1662	-61.19%
144	Vel. Med Vient	2.82-2.90	.	0.0098	0.0663	-85.22%
145	Vel. Med Vient	2.91-3.05	0.0476	0.0553	0.0976	-43.34%
146	Vel. Med Vient	3.05-3.26	-0.0289	-0.0771	-0.0767	0.52%
147	Vel. Med Vient	3.26-3.51	.	-0.0249	-0.0496	-49.80%
148	Vel. Med Vient	3.51-high	0.0303	0.0186	-0.0148	-225.68%

149	Vel. Med Vient	Indeterminado	.	0.0128	0	0%
150	P. Atm. Niv. Mar	Low-1016.44	.	-0.0096	-0.0009	966.67%
151	P. Atm. Niv. Mar	1016.54-1016.78	.	0.0134	0.0412	-67.48%
152	P. Atm. Niv. Mar	1016.78-1016.85	.	-0.0091	0.0263	-134.60%
153	P. Atm. Niv. Mar	1016.85-1017.00	.	0.0187	0.0922	-79.72%
154	P. Atm. Niv. Mar	1017.00-1017.29	-0.0407	-0.0924	-0.0843	9.61%
155	P. Atm. Niv. Mar	1017.29-1017.50	.	0.0100	0.0829	-87.94%
156	P. Atm. Niv. Mar	1017.50-1017.86	0.1127	0.0870	0.042	107.14%
157	P. Atm. Niv. Mar	1017.86-1018.18	0.0709	0.0455	0.0549	-17.12%
158	P. Atm. Niv. Mar	1018.18-high	.	0.0083	0.106	-92.17%
159	P. Atm. Niv. Mar	Indeterminado	.	0.0326	0	0%
160	Cuenca Nival	Si	-0.1395	-0.1984	-0.0516	284.50%
161	Cuenca Nival	Indeterminado	.	0.0086	0	0%
162	Zona Inundable	Si	0.0113	0.0291	0.0414	-29.71%
163	Zona Inundable	Indeterminado	.	0.0079	0	0%
164	Temp Media	Low-12.31	.	0.0145	-0.0811	-117.88%
165	Temp Media	12.31-13.42	-0.0763	-0.0582	-0.144	-59.58%
166	Temp Media	13.42-14.66	-0.0958	0.0905	-0.1721	-152.59%
167	Temp Media	14.66-15.16	0.0403	0.0695	-0.0654	-206.27%
168	Temp Media	15.16-15.35	.	-0.0105	-0.0944	-88.88%
169	Temp Media	15.84-17.24	.	0.0099	0.0706	-85.98%
170	Temp Media	17.24-18.32	.	0.0081	-0.0914	-108.86%
171	Temp Media	18.32-18.69	.	0.0327	-0.0156	-309.62%
172	Temp Media	18.69-high	.	0.0079	-0.1735	-104.55%
173	Temp Media	Indeterminado	.	0.0074	0	0%
174	Precip total	Low-378.06	.	-0.0621	-0.2132	-70.87%
175	Precip total	378.06-409.62	-0.0465	-0.0747	-0.1906	-60.81%
176	Precip total	409.62-420.07	.	0.0112	-0.1198	-109.35%
177	Precip total	420.07-438.73	.	-0.0099	-0.0744	-86.69%
178	Precip total	438.73-488.78	.	-0.0047	-0.1306	-96.40%

179	Precip total	488.78-566.86	0.0258	0.0281	-0.0346	-181.21%
180	Precip total	566.86-647.35	.	-0.0067	-0.0146	-54.11%
181	Precip total	709.95-1076.72	.	0.0424	0.1537	-72.41%
182	Precip total	1076.72-high	0.1066	0.0886	0.4229	-79.05%
183	Precip total	Indeterminado	.	0.0070	0	0%
184	Prec. Max.	Low-11.97	0.0281	0.0372	-0.0473	-178.65%
185	Prec. Max.	11.97-12.68	.	-0.0223	-0.0777	-71.30%
186	Prec. Max.	12.68-13.30	0.0326	0.0581	-0.0175	-432.00%
187	Prec. Max.	13.30-13.99	.	0.0138	-0.0441	-131.29%
188	Prec. Max.	13.99-16.44	.	0.0222	-0.0754	-129.44%
189	Prec. Max.	16.44-18.24	.	-0.0457	-0.1661	-72.49%
190	Prec. Max.	18.24-20.25	0.0132	0.0233	-0.0861	-127.06%
191	Prec. Max.	20.25-22.74	-0.0435	-0.0526	-0.1412	-62.75%
192	Prec. Max.	24.31-high	.	0.0508	0.0009	5544.44%
193	Prec. Max.	Indeterminado	.	0.0068	0	0%
194	Congelación	Low-3.01	0.0289	0.0221	0.1199	-81.57%
195	Congelación	3.01-8.12	.	-0.0005	0.0196	-102.55%
196	Congelación	8.12-14.86	0.0156	0.0294	0.0945	-68.89%
197	Congelación	14.86-25.98	0.0522	0.0609	0.1454	-58.12%
198	Congelación	25.98-48.37	.	0.0113	0.0354	-68.08%
199	Congelación	62.89-71.25	.	-0.0394	-0.0345	14.20%
200	Congelación	71.25-161.00	-0.0050	-0.0193	0.0516	-137.40%
201	Congelación	161.00-305.67	.	0.0427	0.1847	-76.88%
202	Congelación	305.67-high	.	0.0305	-0.0818	-137.29%
203	Congelación	Indeterminado	.	0.0069	0	0%
204	Rac Vient + 55km	Low- 235.52	.	-0.0026	-0.0824	-96.84%
205	Rac Vient + 55km	235.52-366.16	.	-0.0008	-0.0377	-97.88%
206	Rac Vient + 55km	366.16-431.03	0.0363	0.0613	0.0234	161.97%
207	Rac Vient + 55km	431.03-506.94	-0.0002	-0.0380	-0.0854	-55.50%
208	Rac Vient + 55km	506.94-588.71	0.0528	0.0495	-0.0082	-703.66%

209	Rac Vient + 55km	588.71-704.50	.	-0.0004	0.0248	-101.61%
210	Rac Vient + 55km	704.50-809.52	.	0.0131	0.0359	-63.51%
211	Rac Vient + 55km	880.83-1140.01	.	-0.0235	0.0404	-158.17%
212	Rac Vient + 55km	1140.01-high	.	0.0087	0.0621	-85.99%
213	Rac Vient + 55km	Indeterminado	.	0.0070	0	0%
214	Rac Vient + 91km	Low-1.72	0.0010	0.0120	0.1707	-92.97%
215	Rac Vient + 91km	1.72-3.50	-0.0087	-0.0502	0.0438	-214.61%
216	Rac Vient + 91km	3.50-5	0.0631	0.0666	0.162	-58.89%
217	Rac Vient + 91km	5-8	.	-0.0098	0.0426	-123.00%
218	Rac Vient + 91km	8-10.12	0.0980	0.0843	0.1578	-46.58%
219	Rac Vient + 91km	10.12-14.85	.	0.0045	0.0025	80.00%
220	Rac Vient + 91km	17.95-25.99	0.0435	0.0681	0.1064	-36.00%
221	Rac Vient + 91km	25.99-40.99	-0.0359	-0.0450	0.067	-167.16%
222	Rac Vient + 91km	40.99-high	.	0.0056	0.1192	-95.30%
223	Rac Vient + 91km	Indeterminado	.	-0.0068	0	0%
225	Días Gran	Low-8.26	0.0312	0.0442	0.0215	105.58%
226	Días Gran	8.26-12.73	.	0.0092	-0.0026	-453.85%
227	Días Gran	12.73-15.99	.	-0.0046	-0.0025	84.00%
228	Días Gran	15.99-22.29	0.0484	0.0685	0.0261	162.45%
229	Días Gran	30.00-36.99	0.0065	0.0310	-0.0027	-1248.15%
230	Días Gran	36.99-43.99	0.0111	0.0412	0.063	-34.60%
231	Días Gran	43.99-72.62	.	-0.0479	-0.0235	103.83%
232	Días Gran	72.62-94.64	-0.0256	-0.1091	-0.2441	-55.31%
233	Días Gran	94.64-high	.	-0.0824	-0.3877	-78.75%
234	Días Gran	Indeterminado	.	0.0067	0	0%
235	Rur_urb	1.95-7.99	-0.0103	-0.0549	-0.0258	112.79%
236	Rur_urb	7.99-25.42	.	-0.0096	0.0211	-145.50%
237	Rur_urb	25.42-64.17	.	-0.0051	-0.0095	-46.32%
238	Rur_urb	64.17-123.64	0.0045	0.0271	0.0381	-28.87%
239	Rur_urb	123.64-291.33	-0.1417	-0.1165	-0.0654	78.13%

240	Rur_urb	291.33-487.48	.	0.0607	0.2152	-71.79%
241	Rur_urb	487.48-798.07	.	0.0272	0.1809	-84.96%
242	Rur_urb	798.07-1116.70	.	0.0286	0.2535	-88.72%
243	Rur_urb	1116.70-high	.	0.0040	0.2244	-98.22%
244	Rur_urb	Indeterminado	.	0.0066	0	0%

Tabla elaborada por el autor.

Figura 15.2.1. Tramos Validación Cruzada *Elastic Net*, $\alpha = 0.1$

C. Num	Var. Seleccionad.	λ	Log(λ)
1	1	0.0398600	-3.22238196
2	2	0.0363200	-3.31538673
3	2	0.0330900	-3.40852416
4	3	0.0301500	-3.50157036
5	3	0.0274800	-3.59429681
6	4	0.0250300	-3.68768017
7	5	0.0228100	-3.78055624
8	5	0.0207800	-3.87376429
9	5	0.0189400	-3.96647919
10	7	0.0172600	-4.05936359
11	8	0.0157200	-4.15282149
12	9	0.0143300	-4.24540004
13	11	0.0130500	-4.33896715
14	11	0.0118900	-4.43205757
15	13	0.0108400	-4.52451228
16	16	0.0098740	-4.61785024
17	19	0.0089970	-4.71086409
18	24	0.0081980	-4.80386506
19	28	0.0074700	-4.89686028
20	30	0.0068060	-4.9899507
21	34	0.0062010	-5.08304471
22	39	0.0056500	-5.17609973
23	43	0.0051480	-5.26914699
24	51	0.0046910	-5.3621095
25	56	0.0042740	-5.45520512
26	59	0.0038950	-5.5480616
27	66	0.0035490	-5.64108941
28	68	0.0032330	-5.73434478
29	73	0.0029460	-5.82730696
30	79	0.0026840	-5.92044706
31	86	0.0024460	-6.01330124
32	93	0.0022290	-6.10620222
33	95	0.0020310	-6.199227
34	98	0.0018500	-6.29256964
35[30]	104	0.0016860	-6.38539642
36	108	0.0015360	-6.47857364
37	110	0.0014000	-6.57128304
38	110	0.0012750	-6.6648091
39	115	0.0011620	-6.75761262
40	116	0.0010590	-6.85043021
41	124	0.0009647	-6.94369339
42	130	0.0008790	-7.03672566
43	135	0.0008009	-7.12977446

[30] La fila coloreada de amarillo detecta el λ^{opt} de la modelización externa para la técnica *Elastic Net*. Como mencionamos, nosotros necesitábamos aumentar el número de variables por lo que escogimos un segundo λ a cambio de sacrificar *Poisson Deviance*.

44	141	0.0007298	-7.22274003
45	145	0.0006649	-7.3158739
46	151	0.0006059	-7.4087956
47	158	0.0005521	-7.50178137
48	160	0.0005030	-7.59492039
49	164	0.0004583	-7.68798657
50	165	0.0004176	-7.78098652
51	166	0.0003805	-7.87402438
52	181	0.0003467	-7.96705071
53	187	0.0003159	-8.06008485
54	190	0.0002878	-8.15324476
55	192	0.0002623	-8.24602167
56	192	0.0002390	-8.33904701
57	195	0.0002177	-8.43239259
58	199	0.0001984	-8.52522536
59	201	0.0001808	-8.61811911
60	206	0.0001647	-8.71138492
61	206	0.0001501	-8.80420882
62	209	0.0001367	-8.89772181
63	209	0.0001246	-8.99040195
64	214	0.0001135	-9.08370772
65	214	0.0001034	-9.1769056
66	219	0.0000943	-9.26955973
67	219	0.0000859	-9.36255958
68	220	0.0000783	-9.45560173
69	219	0.0000713	-9.54861423
70	220	0.0000650	-9.64158493
71	220	0.0000592	-9.73475795
72	231	0.0000539	-9.82763824
73	230	0.0000491	-9.92083719
74	231	0.0000448	-10.0137489
75	229	0.0000408	-10.1068285
76	230	0.0000372	-10.1997396
77	230	0.0000339	-10.2929809
78	229	0.0000309	-10.3860497
79	234	0.0000281	-10.4790295
80	234	0.0000256	-10.5721373
81	233	0.0000234	-10.6649136
82	233	0.0000213	-10.7582129
83	234	0.0000194	-10.851269
84	234	0.0000177	-10.9442084
85	234	0.0000161	-11.0373126
86	231	0.0000147	-11.1303879
87	231	0.0000134	-11.2232454
88	231	0.0000122	-11.3165367
89	231	0.0000111	-11.4094668
90	231	0.0000101	-11.5019855
91	231	0.0000092	-11.5953293
92	231	0.0000084	-11.6883509
93	231	0.0000076	-11.7814587

94	231	0.0000070	-11.8744694
95	231	0.0000063	-11.9675283
96	237	0.0000058	-12.060588
97	237	0.0000053	-12.1534802
98	236	0.0000048	-12.2466863
99	235	0.0000044	-12.339604
100	236	0.0000040	-12.4327223

Tabla elaborada por el autor.

16.4. Análisis de Datos Atípicos

Como sabemos, es muy común que en las estimaciones de modelos existan datos atípicos que distorsionen la predicción. Para detectar estos datos emplearemos dos modelos estadísticos muy frecuentes en estos casos: el **Efecto Palanca** *(Leverage)* y la **Distancia de Cook**.

➢ Puntos Palanca

Los Puntos Palanca *(Leverage Point)* son observaciones que potencialmente pueden tener una gran capacidad de influencia sobre la regresión, siendo su expresión matemática:

$$\boldsymbol{v_{ii} = x'_i(X'X)^{-1}x_i = \frac{1}{n}(1 + \underbrace{(\tilde{x}_i - \bar{x})'S_x^{-1}(\tilde{x}_i - \bar{x})}_{\textbf{Distancia de Mahalanobis}})}$$

Donde:

$\boldsymbol{\tilde{x}'_i = (x_{1i}, x_{2i}, \dots, x_{ki})}$

$\boldsymbol{x'_i = (1\tilde{x}'_i)}$

$\boldsymbol{\bar{x}}$**:** Vector de medias de las *k* variables explicativas.

$\boldsymbol{S_x}$: Matriz de covarianzas.

Si sabemos que los puntos de palanca son una medida de distancia entre $\tilde{x}_i$ y $\bar{x}$ (centro de gravedad), dicha medida puede simplificarse debido a que la medida solo depende de *k* y *n*.

$$\boldsymbol{\bar{v} = \frac{1}{n}\sum_{i=1}^{n} v_{ii} = \frac{tr(H)}{n} = \frac{k+1}{n}}$$

Por lo que, si estamos estudiando si dichas observaciones son puntos palanca, esta debe ser superior al doble de la distancia existente entre $\tilde{x}_i$ y $\bar{x}$.

Variables Internas	Variables Externas
$v_{ii} > 2\overline{v} = 2(k+1)/n$ $v_{ii} > 2 * (49+1)/500000$ $v_{ii} = \frac{100}{500000} > 0.0002$	$v_{ii} > 2\overline{v} = 2(k+1)/n$ $v_{ii} > 2 * (72+1)/500000$ $v_{ii} = \frac{146}{500000} > 0.000292$

Todo Leverage superior a este dato será un posible punto de influencia. Si filtramos la base de datos por el resultado del *leverage* obtendremos que 17188 de las 500000 de observaciones son potencialmente influyentes para el modelo interno y 13377 de las 500000 de observaciones son potencialmente influyentes para el modelo de variables externas.

Pero en sí, estos datos no me dice nada sin conocer los verdaderos puntos influyentes detectados a través de la Distancia de Cook.

➢ Puntos Influyentes

Una variable será influyente si al eliminarla, la estimación cambia mucho, pudiendo modificar los parámetros estimados y consigo los resultados predichos inicialmente. Para medir esos datos empleamos la ***Distancia de Cook*** que matemáticamente se expresará como:

$$D(i) = \frac{r_i^2}{k+1}\frac{v_{ii}}{1-v_{ii}}$$

Donde:

$$r_i = \frac{e_i}{s_R\sqrt{1-v_{ii}}}, i = 1, 2, \dots, n$$

Diciéndose que una variable será efectivamente influyente si:

Modelo Variables Internas	Modelo Variables Externas
$D(i) > F^{\alpha}_{k+1,n-k-1}$ $F^{0.05}_{50,499950} = 1.35$	$D(i) > F^{\alpha}_{k+1,n-k-1}$ $F^{0.05}_{50,499927} = 1.35$

Si volvemos a filtrar las bases de datos para verificar si alguna observación es influyente, vemos que para ninguno de las dos modelizaciones existe ninguna observación que supere el 1.35 para la variable Distancia de Cook.

Gráficamente, las observaciones se distribuyen de la siguiente manera:

16.4.1. Modelización Variables Internas

Figura 17.4.1. Gráfica Puntos Palanca

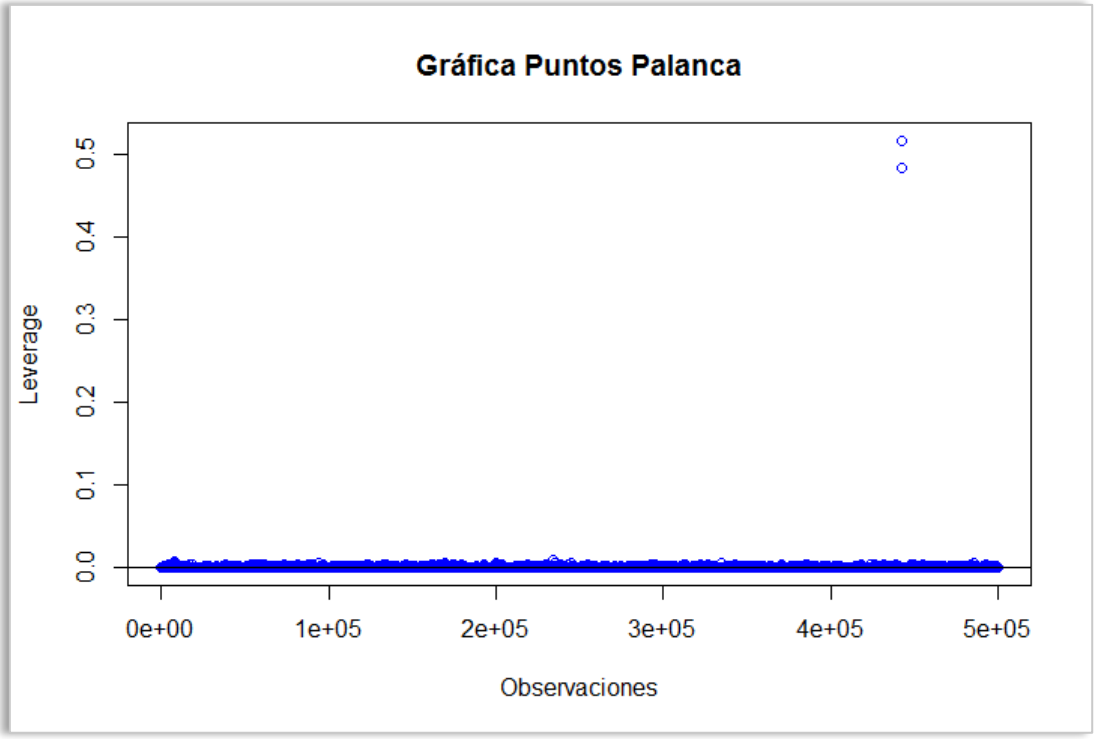


Gráfico elaborado por el autor.

Figura 17.4.2. Gráfica Puntos Influyentes.

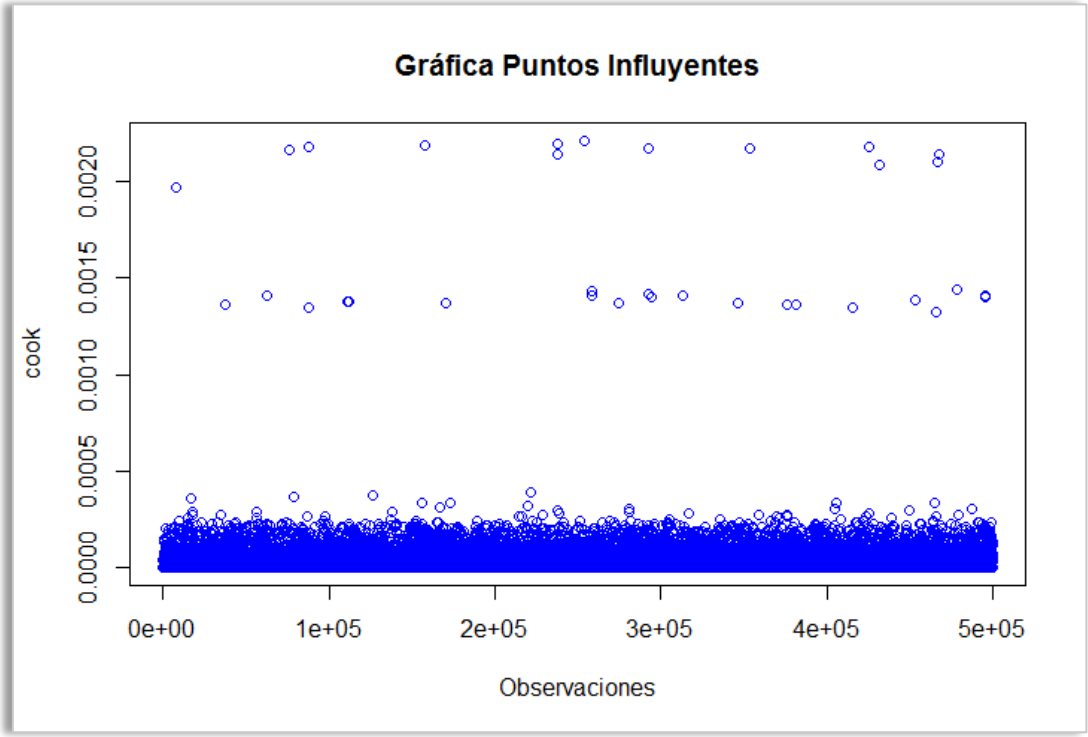


Gráfico elaborado por el autor.

16.4.2. Modelización Variables Internas y Externas

Figura 17.4.1. Gráfica Puntos Palanca

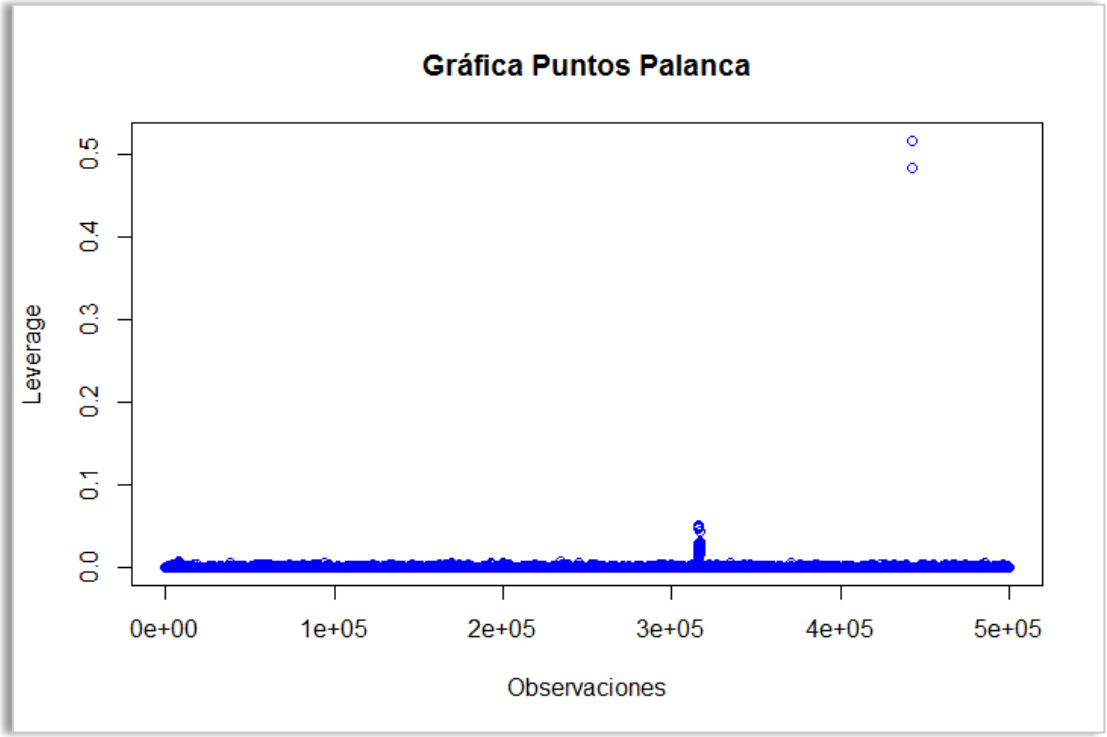


Gráfico elaborado por el autor.

Figura 17.4.2. Gráfica Puntos Influyentes.

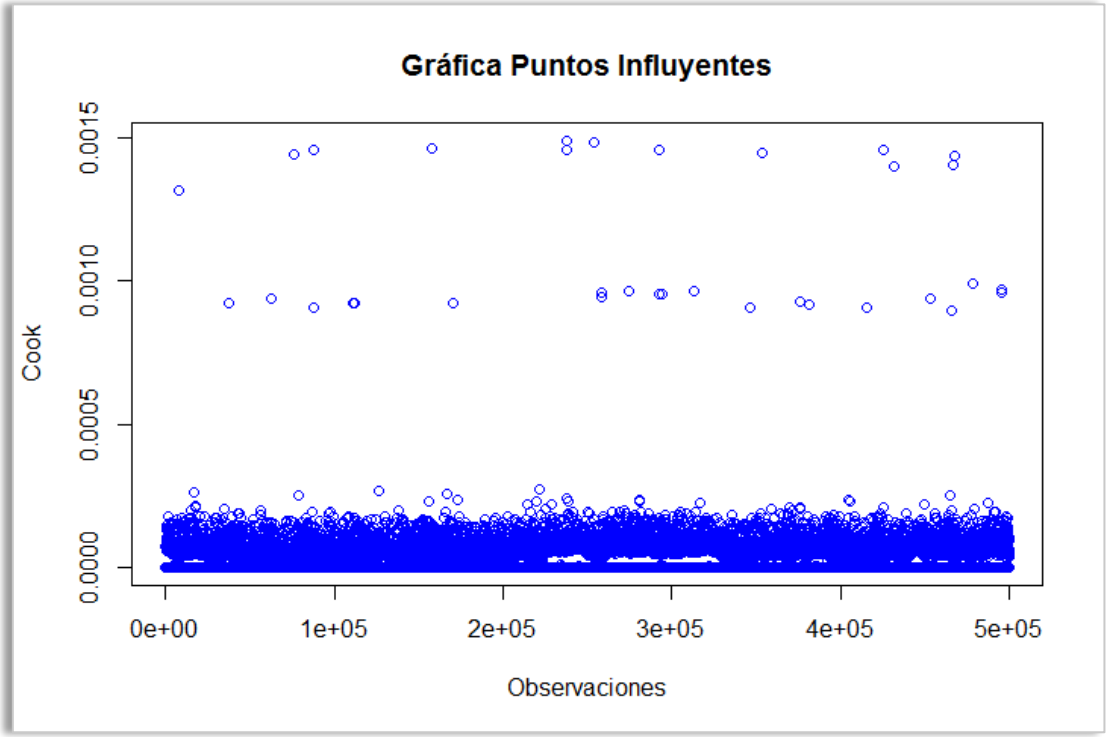


Gráfico elaborado por el autor.

17.Bibliografía[31]

17.1. Libros, Papers y Legislación

AGRESTI, A. 2015. *Foundations of Linear and Generalized Models*. New Yersey: Wiley. ISBN: 978-1-118-73003-4.

ÁLVAREZ JAREÑO, J.A; MUÑIZ RODRÍGUEZ, PRUDENCIO. 2010. *Reparametrización de los principales Distribuciones de Probabilidad en el Estudio del Número de Siniestros debido a las Anomalías Muestrales en las Carteras del Seguro de Responsabilidad Civil del Automóvil. Determinación del Índice de Dispersión*. Anales 2010, pp. 1-24.

BALBÁS DE LA CORTE, A. 2015. *Tarificación no Vida.* Madrid. Universidad Carlos III.

BOJ DEL VAL, E; CLARAMUNT BIELSA, M.M; FORTIANA GREGORI, J. 2001. *Herramientas Estadísticas para el Estudio de Perfiles de Riesgo*. Anales 2001, p.p. 59-89.

BOUSOÑO, C; HERAS, A.; TOLMOS, P. 2008. *Factores de Riesgo y Cálculo de Primas mediante Técnicas de Aprendizaje*. Majadahonda. Madrid. Fundación Mapfre.

CARO CARRETERO, R. Segmentación y Predicción en los Modelos de Tarificación. *Segundo Congreso Internacional de Matemáticas en la Ingenieria y la Arquitectura.*

CASTRO, S. 2011. *Análisis de Datos en Grandes Dimensiones. Estimación y Selección de Variables en Regresión*. Instituto de Estadística y Departamento de Métodos Cuantitativos. Universidad de la República.

CASTRO, S. 2013. *Estimación y Selección de Variables en Grandes Dimensiones… Regresión Ridge, Lasso, Elastic Net, SCAD*.

DASTIS OLAZ, J. 2015. *Modelos GAM aplicados al Seguro de Hogar.* RODRIGUEZ PARDÓ, J. M; SIMÓN DEL POTRO, J.(dir). Trabajo Fin de Máster. Universidad Carlos III de Madrid.

DOBSON, A. 2002. An Introduction to Generalized Linear Models.

[31] Referenciamos siguiendo la normativa ISO 690.

EMBID HERRANZ, I. 2011. *Canales de Distribución en Seguros: Efectividad Comercial y Eficiencia Operativa.* MARTÍN DÁVILA, M; ZORRILLA FERNÁNDEZ, V. (dir). Trabajo Fin de Máster. Universidad Rey Juan Carlos.

ESPAÑA. 2001. Reglamento Artículado del Sistema CICOS. *Asociación Empresarial del Seguro.*

EUROPA. 2012. Directrices sobre la Aplicación de la Directiva 2004/133/CE del Consejo a los Seguros, a la luz de la Setencia del Tribunal de Justicia de la Unión Europea en el Asunto C-239/09 (Test-Achats). *Comisión Europea. Diario de la Unión Europea*.

FRIEDMAN, J; HASTIE, T; SIMON, N; TIBSHIRANI, R. 2016. *Lasso and Elastic-Net Regularized Generalized Linear Models: Package 'glmnet'*. Package R.

GLM-Introducción. Máster en Ciencias Actuariales. Universidad de Valencia.

HASTIE, T; ROBERT, T; FRIEDMAN, H. J. 2008. *The Elements of Statistical Learning: Data Mining Inference and Prediction.* Second Edition. Standford (California): Springer.

HASTIE, T; QUIAN, J. 2014. *Glmnet Vignette*.Standford.

HERAS, A. 2015. *Introducción a la Tarificación en Seguros no Vida*. Madrid. Universidad Complutense de Madrid.

HERRANZ VALERA, J. 2015. Análisis de Supervivencia: Alta Dimensionalidad. *VII Jornadas de Usuarios de R. Salamanca*.

LÓPEZ-GONZÁLEZ, E; RUIZ-SOLER, M. 2011. Análisis de datos con el Modelo Líneal Generalizado. Una aplicación con R. *Revista Española de Pedagogía,* (248), pp. 59-80.

MARTÍN CABELLO, J. A. 2015. *Análisis e Inclusión de Variables Exógenas en la Tarificación de Autos mediante Modelización GLM.* RODRIGUEZ PARDÓ, J. M; SIMÓN DEL POTRO, J.(dir). Trabajo Fin de Máster. Universidad Carlos III de Madrid.

MELGAR HIRALDO, M; GUERRERO CASAS, F. 2005. Los Siniestros en el Seguro del Automóvil: un Análisis Econométrico Aplicado. *Estudios de Economía Aplicada*, **23**(1), pp.355-375. ISSN: 1133-3197.

NADAL, R. 2014. La eficiencia en precios y costes, como elementos clave para mantener la competitividad en el seguro de automóvil. *Instituto de Actuarios Españoles,* (34), pp. 20-23.

OHLSSON, E; JOHANSSON, B. 2010. *Non-Life Insurance Pricing with Generalized Linear Models*.

PASCAL VILLACAMPA, M. 2005-2006. Proceso Tarificación en el Seguro de Automóvil. Trabajo Fin de Máster. Universitat de Barcelona.

PLAZA CAMPOS, L. 2015. *Sistemas de Geolocalización (GIS) en el Pricing GLM del Seguro Multirriesgo del Hogar.* RODRIGUEZ PARDÓ, J. M; SIMÓN DEL POTRO, J.(dir). Trabajo Fin de Máster. Universidad Carlos III de Madrid.

QUISHPE TASIGUANO, I. D. 2015. *Factores de Riesgo de Siniestralidad y Cálculo de Primas de los Vehículos Asegurados en el Ecuador mediante Modelos Líneales Generalizados.* GUACHAMÍN GUERRA, M. E; MEDINA VALLEJO, J. C. (dir). Trabajo Fin de Grado. Quito.

PRESNELL, B. 2000. *An Introduction to Categorical Data Analysis Using R.*

Servicio de Estadísticas y Estudios del Sector Seguros en España (ICEA). [Sitio Web]. 2016. Madrid. [Consulta: abril de 2016]. Disponible en: http://www.icea.es/es-ES/Paginas/home.aspx

SUNDBERG, R. 2006. Shrinkage Regression. *Encyclopedia of Environmetrics*, **4**, pp.1994-1998.

Tecnologías de la Información y Redes para las Entidades Aseguradoras (TIREA). [Sitio Web]. Disponible en: http://www.tirea.es/Entidades-Aseguradoras/Autos.aspx

TORTELLA, G [et al.]. 2014. *Historia del Seguro en España*. Fundación Mapfre.

VEGAS MONTANER, A. 2014. Seguro del Automóvil en España: de la Edad Media a la Edad Moderna. *Instituto de Actuarios Españoles*, (34), pp7-12.

WOOLDRIDGE, J. M. 2010. *Introducción a la Econometría: Un Enfoque Moderno*. 4º Edición. Cengage Learning.

ZOU, H; HASTIE, T. 2005. Regularization and Variable Selection via the Elastic Net. *J. R. Statist. Soc. B*, (67), pp. 301-320.

17.2. Software

SAS Customer Support Knowledge and Conmunity

R-Studio