UNIVERSIDAD CARLOS III DE MADRID

ESCUELA DE POSTGRADO

"Modelos GAM aplicados al seguro de hogar"

Trabajo fin de máster presentado por el alumno

Javier Dastis Olaz

Para optar al título de

MÁSTER EN CIENCIAS ACTUARIALES Y FINANCIERAS

Bajo la dirección de

José Miguel Rodríguez Pardo

Jesús Simón del Potro

MADRID - ESPAÑA

JUNIO 2015

Agradecimientos

Deseo expresar mi más sincero agradecimiento a los profesores y directores de este proyecto José Miguel Rodríguez Pardo y Jesús Simón del Potro, por su dedicación y su constante apoyo, que han sido fundamentales a lo largo de este camino.

Gracias a Miguel Usábel, uno de los mejores profesores de los que he podido disfrutar en mi vida académica, por haberme inculcado las ganas de mejorar y de llegar aún más lejos.

A mis padres, abuelos y tata, que a lo largo de estos años me han apoyado y dado ánimos para seguir adelante.

También quiero dedicar este proyecto a mi hermana, Itziar, que siempre estará a mi lado protegiendome como cuando eramos niños. Nunca te olvidaré, estarás presente en mi vida y en mi memoria.

Y a Coco, uno de los pilares de mi vida. Con sólo una sonrisa haces que olvide todos los problemas y me ayudas a superar cualquier obstaculo que se me presente.

Por último, quiero dar las gracias a mis amigos, en especial a uno de ellos, Fernando, que ha sido como un hermano para mí y con él he vivido momentos inolvidables.

Gracias a todos.

ÍNDICE GENERAL

Introducción
Modelos GAM aplicados al seguro de hogar

El presente trabajo se centrará en el estudio de la frecuencia siniestral en el ramo de hogar.

El ramo de hogar se caracteriza por tener bajos índices de frecuencia en contraposición a otros ramos como el de autos.

El primer paso que llevaremos a cabo será el análisis de la variable clave en nuestro estudio: el número de siniestros. Este estudio será un paso fundamental, ya que, posteriormente, al modelizar dicha variable aleatoria se requerirá que cumpla una serie de hipótesis, entre ellas que se ajuste a una distribución de la familia exponencial. En nuestro caso, veremos que la variable aleatoria número de siniestros no ajusta a una distribución Poisson, pero si a una binomial negativa. Este hecho sucede a menudo cuando se estudia el número de siniestros de una cartera, ya que la heterogeneidad es algo implícito. La distribución binomial negativa no es miembro de la familia exponencial, sin embargo, dicha distribución no deja de ser una distribución Poisson sobredispersa en la que dicha sobredispersión viene dada por la distribución gamma que, además, es máximo entrópica.

Posteriormente se realizará un análisis de los factores de riesgo que puedan explicar la frecuencia siniestral en el ramo de hogar. En concreto, hemos seleccionado: la superficie, el capital del continente, el capital del contenido, la modalidad de pago, el año de la cartera de estudio, el tipo-uso-ubicación del inmueble y, por último, la antigüedad de la póliza. A continuación, modelizaremos la variable aleatoria del número de siniestros con un GLM de conteo en el que podremos comprobar como nuestro modelo nos devolverá mejores predicciones en aquellos clúster lo suficientemente poblados. Esto es algo obvio, ya que el GLM trabaja en medias.

Finalmente, asumiremos la hipótesis de que la parte no explicada del GLM vendrá dada toda ella por la localización (código postal) del riesgo. Para ello, modelizaremos la deviance residual estandarizada con un GAM en el que la variable explicativa será el código postal. Los splines se generarán en función de las geocordenadas (latitud y longitud) del código postal. Los resultados se representarán en un mapa físico utilizando el software MapPoint.

Primera parte
Tarificación seguros no vida

La existencia del ser humano siempre ha estado ligada con la incertidumbre. El ser humano, desde el momento en que nace, está expuesto a la ocurrencia de sucesos que pueden ocasionar consecuencias negativas.

Este es uno de los principales motivos de la creación del negocio asegurador. El seguro es un instrumento que permite paliar las consecuencias económicas negativas, medibles en términos monetarios, que acarree la realización del suceso.

Existen dos ramos principales en el negocio asegurador: vida y no vida. Las principales diferencias entre ambos son:

I. Los seguros de no vida se caracterizan por ser seguros a corto plazo (normalmente un año), por lo que el tipo de interés no es demasiado relevante.
II. En los seguros de no vida, las indemnizaciones están en función del daño derivado del siniestro, por lo que puede producirse situaciones de infraseguro y sobreseguro al no coincidir suma asegurada con valor asegurado.
III. En seguros de vida la suma asegurada está fijada a la hora de suscribir el contrato (no tiene ningún componente aleatorio), mientras que en no vida, viene determinada por la cuantía del daño definida por una variable aleatoria: la severidad.
IV. El principal factor para determinar la cuantía de las primas en seguros de vida es la edad del asegurado, mientras que en no vida se dan una mayor serie de factores de riesgo, por ejemplo en seguro de auto: clase de vehículo, zona de conducción, potencia, cilindrada…Esta tendencia está cambiando en las entidades aseguradoras, ya que en vez de tarificar con las clásicas tablas de mortalidad, empiezan a utilizar Modelos Lineales Generalizados (GLM) identificando distintos factores de riesgo que afecten a la mortalidad de sus asegurados: peso, talla, profesión…
V. Los seguros de vida presentan una mayor estabilidad que los de no vida al tener menos fluctuaciones en torno a sus valores medios.
VI. Desde el punto de vista técnico, los seguros de no vida, son mucho más complejos que los seguros de vida.
VII. El componente aleatorio es diferente para ambos ramos. En vida, la variable aleatoria básica es la probabilidad de fallecimiento/supervivencia, mientras que en no vida se trabaja con las variables aleatorias de frecuencia y severidad.

1
Modelos de siniestralidad: frecuencia y severidad

El coste agregado es uno de los procesos estocásticos más importantes en seguros. El coste agregado de una cartera es la suma, en cantidades aleatorias, de todas las severidades ocurridas en un intervalo definido de tiempo. Los dos componentes que forman la distribución del coste agregado de una cartera de no vida son la frecuencia y la severidad. Estas variables aleatorias deben ser modeladas por separado y posteriormente combinadas para representar la distribución del coste agregado de una cartera. La frecuencia se modelará utilizando distribuciones de probabilidad discretas de valores enteros no negativos, mientras que la severidad se modelará ajustando distribuciones de probabilidad de carácter continuo:

$$y = \sum_{i=0}^{N_t} S_i$$

Siendo:

- y el coste agregado de la cartera.
- N_t la variable aleatoria del número de siniestros (frecuencia).
- S_i las severidades de la cartera.

A continuación, se describirán las distribuciones de probabilidad más utilizadas para modelar cada una de estas variables aleatorias.

1.1 Distribuciones para modelar la frecuencia

En la teoría clásica actuarial, para modelar la frecuencia en una cartera de pólizas, se utilizan distribuciones de probabilidad discretas que tomen valores enteros no negativos. Las más utilizadas son:

1.1.1 Distribución binomial

Una variable aleatoria ξ seguirá una distribución B(n, p) cuando:

$$\xi = \xi_1 + \xi_2 + \xi_3 \ldots . \xi_n$$

Siendo todas las variables aleatorias anteriores independientes y con la misma distribución B(1, p).

La distribución binomial cuenta el número de éxitos, dentro de una secuencia de n ensayos de Bernoulli independientes con una probabilidad fija p de ocurrencia de éxito. Un suceso Bernoulli se caracteriza por ser dicotómico, es decir, puede tomar sólo dos valores codificados, 1 con probabilidad p y 0 con probabilidad q.

La función de probabilidad de una distribución binomial, deberá proporcionar la probabilidad de que en n repeticiones se consigan x número de éxitos:

$$P(\xi = x) = \binom{n}{x} p^x \, q^{n-x}$$

Fuente: elaboración propia

La esperanza matemática y la varianza serán:

$$E(\xi) = E(\xi_1 + \xi_2 + \ldots + \xi_n) = E(\xi_1) + E(\xi_2) + \ldots + E(\xi_n) =$$

$$= p + p + \ldots + p = np$$

$$V(\xi) = V(\xi_1 + \xi_2 + \ldots + \xi_n) = V(\xi_1) + V(\xi_2) + \ldots + V(\xi_n) =$$

$$= pq + pq + \ldots + pq = npq$$

Y su función característica:

$$\varphi_\xi(t) = \varphi_{\xi_1+\xi_2+\ldots+\xi_n}(t) = \varphi_{\xi_1}(t)\varphi_{\xi_2}(t)\ldots\varphi_{\xi_n}(t) =$$

$$= \prod_{j=1}^{n} \varphi_{\xi_j}(t) = \prod_{j=1}^{n} (q + pe^{it}) = (q + pe^{it})^n$$

Ya que al ser variables independientes, la función característica de su suma será el producto de las funciones características de cada una de ellas.

Además todas tienen la misma función característica por ser variables aleatorias Bernoulli:

$$\varphi(t) = E\left(e^{it\xi}\right) = \sum_{j=1}^{2} e^{itx_j} p_j = e^{it0}q + e^{it1}p = q + pe^{it}$$

En una distribución binomial, si el número de ensayos n es suficientemente grande (normalmente mayor de 30) y p=q, la distribución binomial podrá aproximarse a una normal gracias al Teorema Central del Límite. Además si n tiende a infinito y p es tal que el producto entre ambos tiende a λ, entonces la distribución binomial podrá ajustarse a una distribución de Poisson de parámetro λ.

Por último la distribución binomial cumple una propiedad muy importante: la propiedad aditiva o reproductiva:

$$y = \sum_{i=1}^{n} x_i \sim B\left(\sum_{i=1}^{n} n_i, p\right)$$

Es decir, dadas n variables binomiales independientes, su suma es también una variable binomial de parámetros $n_1 + n_2 + \ldots + n_n$ y p.

1.1.2 Distribución geométrica

La distribución geométrica se representa por $\xi \sim G(p)$.

Esta distribución representa el número de ensayos necesarios hasta la obtención del primer éxito.

Su función de probabilidad queda definida de la siguiente forma:

$$P(\xi = x) = q^x p$$

Y su función de distribución:

$$F(x) = P(\xi \leq x) = \sum_{h=0}^{x} q^h p = p \sum_{h=0}^{x} q^h = p \frac{q^x\, q - 1}{q - 1} =$$

$$= p \frac{q^{x+1} - 1}{q - 1} = p \frac{1 - q^{x+1}}{q - 1} = 1 - q^{x+1}$$

La función característica de la distribución geométrica tiene la siguiente forma:

$$\varphi(t) = E\left(e^{it\xi}\right) = \sum_{h=0}^{\infty} e^{itx}\, p(\xi = x) = \sum_{x=0}^{\infty} e^{itx}\, q^x p =$$

$$= p \sum_{x=0}^{\infty} (q e^{it})^x = p \frac{1}{1 - q e^{it}} = p(1 - q e^{it})^{-1}$$

La esperanza matemática de una distribución geométrica viene dada por:

$$E(\xi) = \frac{q}{p}$$

Y la varianza será:

$$V(\xi) = \frac{q}{p^2}$$

La distribución geométrica no presenta la propiedad aditiva, pero tiene otra propiedad muy importante denominada falta de memoria.

$$P(\xi \geq s + t / \xi \geq t) = \frac{P(\xi > s + t\,;\, \xi > t)}{P(\xi > t)} = \frac{P(\xi > s + t)}{P(\xi > t)} = \frac{P(\xi > s + t - 1)}{P(\xi > t)} =$$

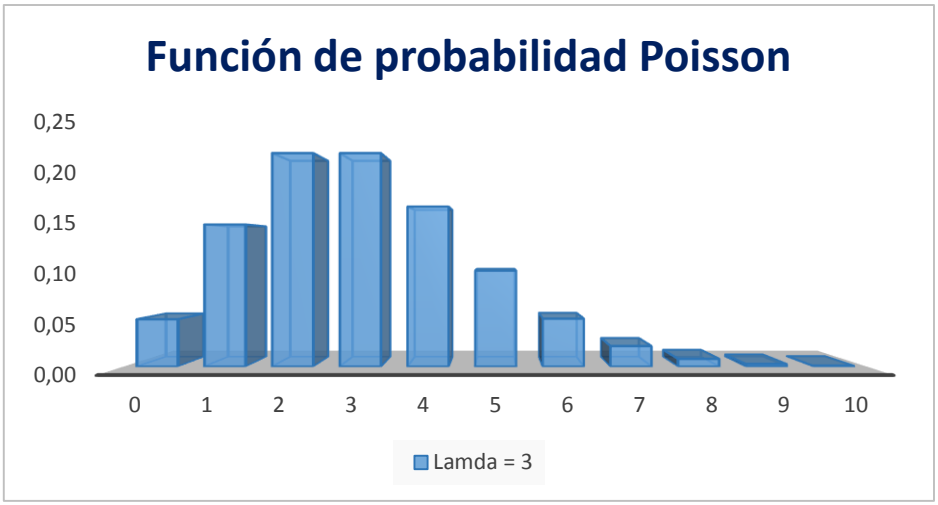
$$= \frac{(1 - p)^{s+t-1}}{(1 - p)^t} = (1 - p)^{s-1} = P(\xi > s - 1) = P(\xi \geq s)$$

1.1.3 Distribución de Poisson

En el campo actuarial, la distribución de Poisson es una de las más utilizadas para modelar la frecuencia en una cartera de pólizas. Se trata de una distribución que expresa a partir de una frecuencia de ocurrencia media λ, la probabilidad de que ocurra un determinado número de eventos durante un cierto periodo de tiempo. A esta distribución también se la conoce como la distribución de los sucesos raros, ya que se especializa en la ocurrencia de sucesos con probabilidades muy bajas.

Su función de probabilidad es:

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



Fuente: elaboración propia

Para demostrar que se trata de una verdadera distribución de probabilidad:

$$\sum_{x=0}^{\infty} P(\xi = x) = \sum_{x=0}^{\infty} \frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\left(1 + \frac{\lambda}{1!} + \frac{\lambda^2}{2!} + \frac{\lambda^3}{3!} + \cdots\right) =$$

$$= e^{-\lambda}e^{\lambda} = 1$$

Su función característica será:

$$\varphi(t) = E\left(e^{it\xi}\right) = \sum_{x=0}^{\infty} e^{itx}\, P(\xi = x) = \sum_{x=0}^{\infty} e^{itx}\frac{e^{-\lambda}\lambda^x}{x!} = e^{-\lambda}\sum_{x=0}^{\infty}\frac{(\lambda e^{it})^x}{x!} =$$

$$= e^{-\lambda}e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}$$

La esperanza de la distribución de Poisson vendrá dada por:

$$E(\xi) = \sum_x xP(\xi = x) = \sum_x x\ \frac{e^{-\lambda}\lambda^x}{x!} = \sum_x x\ \frac{e^{-\lambda}\lambda^x}{(x-1)!\,x} = \sum_x \frac{e^{-\lambda}\lambda^x}{(x-1)!} =$$

$$= \lambda \sum_x \frac{e^{-\lambda}\lambda^{x-1}}{(x-1)!} = \lambda \sum_z \frac{e^{-\lambda}\lambda^z}{z!} = \lambda$$

La varianza, en cambio, queda explicada de la siguiente forma:

$$V(\xi) = E(x^2) - E(x)^2$$

$$E\big(x(x-1)\big) = E(x^2) - E(x)$$

Luego,

$$V(\xi) = E\big(x(x-1)\big) + E(x) - E(x)^2$$

$$E\big(x(x-1)\big) = \sum_x x(x-1)\frac{e^{-\lambda}\lambda^x}{x!} = \sum_x x(x-1)\frac{e^{-\lambda}\lambda^x}{(x-2)!\,(x-1)!\,x} =$$

$$= \sum_x \frac{e^{-\lambda}\lambda^x}{(x-2)!} = \lambda^2 \sum_x \frac{e^{-\lambda}\lambda^{x-2}}{(x-2)!} = \lambda^2 \sum_z \frac{e^{-\lambda}\lambda^z}{z!} = \lambda^2$$

Y, sustituyendo, obtenemos la varianza:

$$V(\xi) = E\big(x(x-1)\big) + E(x) - E(x)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda$$

La distribución de Poisson cumple una propiedad muy importante de aditividad, ya que la suma de variables aleatorias independientes de Poisson es otra variable aleatoria de Poisson cuyo parámetro es la suma de los parámetros de las originales, es decir:

Si $x_i \sim P(\lambda_i)$ y, todas ellas son v.a independientes, entonces $y = \sum_i x_i \sim P(\sum_i \lambda_i)$

Además, si suponemos que para cada valor t>0, que representa el tiempo, el número de sucesos de un fenómeno aleatorio se distribuye conforme a una Poisson de parámetro λt, entonces los tiempos transcurridos entre dos sucesos consecutivos se distribuyen con una exponencial.

Por último, tal y como se comentó anteriormente, la distribución de Poisson es el caso límite de la binomial cuando n tiende a infinito y p a cero:

$$\lim_{\substack{n\to\infty \\ p\to 0}} p(\xi = x) = \lim_{\substack{n\to\infty \\ p\to 0}} \binom{n}{x} p^x q^{n-x} = \lim_{\substack{n\to\infty \\ p\to 0}} \frac{n!}{x!\,(n-x)!} p^x q^{n-x} =$$

$$= \lim_{n\to\infty} \frac{n!}{x!\,(n-x)!} \left(\frac{\lambda}{n}\right)^x \left(1 - \frac{\lambda}{n}\right)^{n-x} =$$

$$= \frac{\lambda^x}{x!} \lim_{n\to\infty} \frac{n(n-1)\dots(n-(x-1))}{n^x} \frac{\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n}{\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^x} = \frac{\lambda^x}{x!} \lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n =$$

$$= \frac{\lambda^x}{x!} \lim_{n\to\infty} \left(\left(1+\frac{1}{\frac{n}{-\lambda}}\right)^{\frac{n}{-\lambda}}\right)^{-\lambda} = \frac{\lambda^x}{x!}\, e^{-\lambda}$$

Sin embargo, en el campo actuarial, asumir una λ constante no es muy realista, ya que puede suceder que este parámetro sea variable en función del momento del tiempo, ya que en algunos tipos de seguro, justo después de un siniestro las posibilidades de que otro suceda son más altas que después de que haya pasado un cierto lapso de tiempo o enfriamiento. Este suceso es conocido como contagio. Por lo que λ pasaría a estar en función del tiempo.

Por ejemplo,

$$\lambda(z) = e^{-2z}$$

$$\lambda(z) = \frac{1}{1+z}$$

Ambos modelos son funciones decrecientes, por lo que los dos representarían un proceso de contagio, sin embargo la naturaleza del contagio sería distinta, ya que como es fácil suponer, el modelo exponencial decrecerá más rápidamente por lo que el contagio durará menos en el tiempo.

La función de probabilidad asociada a este nuevo modelo de Poisson será del tipo:

$$P(\xi_t = x) = \frac{(\int_0^t \lambda(z)dz)^x e^{-\int_0^t \lambda(z)dz}}{x!}$$

A pesar de todo, no parece muy coherente que todos los asegurados de una cartera tengan la misma λ, por lo que algunos modelos tratan de inducir la aleatoriedad en este parámetro.

En este modelo, λ será constante en el tiempo pero ahora será una variable aleatoria que seguirá una función de distribución H conocida como función de estructura.

$$\lambda \rightarrow H(\lambda)$$

Por lo que para obtener la nueva función de probabilidad en este modelo, deberemos resolver la siguiente mixtura:

$$P(\xi_t = x) = \int_0^\infty \frac{(\lambda t)^x e^{-(\lambda t)}}{x!} h(\lambda)d\lambda$$

Uno de los casos más conocidos en la práctica, es cuando la función de estructura sigue una distribución Gamma:

$$x|\lambda \sim P(\lambda) \text{ Y } \lambda \sim G(r, \theta = \tfrac{(1-p)}{p})$$

$$f(x) = P(\xi = x) = \int_0^\infty f_{x|\lambda}(x) f_\lambda(\lambda) d\lambda = \int_0^\infty \frac{e^{-\lambda}\lambda^x}{x!} \frac{\lambda^{r-1} e^{-\frac{\lambda}{\theta}}}{\theta^r \Gamma(r)} d\lambda =$$

$$= \frac{(1-p)^{-r} p^r}{x!\,\Gamma(r)} \int_0^\infty e^{-\lambda/(1-p)}\, \lambda^{r+x-1} d\lambda = \frac{(1-p)^{-r} p^r}{x!\,\Gamma(r)} (1-p)^{r+x}\, \Gamma(r+x) =$$

$$\frac{\Gamma(r+x)}{x!\,\Gamma(r)} p^r (1-p)^x = \binom{x+r-1}{x} p^r (1-p)^x$$

Que es la función de probabilidad de una binomial negativa.

Con,

$$\Gamma(r+x) = \int_0^\infty t^{r+x-1}\, e^{-t} dt = \int_0^\infty \left(\frac{\lambda}{1-p}\right)^{r+x-1} e^{-\frac{\lambda}{1-p}} \frac{d\lambda}{1-p} =$$

$$= \left(\frac{1}{1-p}\right)^{r+x} \int_0^\infty \lambda^{r+x-1}\, e^{-\lambda/(1-p)} d\lambda$$

1.1.4 Distribución binomial negativa

La distribución binomial negativa es, junto con la de Poisson, la más utilizada en la práctica para modelar la variable aleatoria del número de siniestros.

Si suponemos que se repite de forma independiente un ensayo de Bernoulli con probabilidad p de éxito constante en todas las pruebas, a la variable aleatoria BN(r,p) que mide el número de éxitos hasta alcanzar el r-ésimo fallo se le llama variable binomial negativa.

Su función de probabilidad tiene la siguiente forma:

$$P(\xi = x) = \binom{x+r-1}{x} p^r (1-p)^x$$

Y su función de distribución:

$$F(x) = P(\xi \le x) = \begin{cases} 0, & x < 0 \\ \sum_{K=0}^{x} \binom{k+r-1}{k} q^k p^r, & k \le x < k+1 \end{cases}$$

La función característica de la distribución binomial negativa es:

$$\varphi(t) = E\left(e^{it\xi}\right) = \sum_{k=0}^{\infty} e^{itk}\binom{-r}{k}(-q)^k p^r = p^r \sum_{k=0}^{\infty}\binom{-r}{k}(-qe^{it})^k = p^r(1-qe^{it})^{-r}$$

La esperanza matemática de esta variable viene determinada por:

$$E(\xi) = r\frac{p}{(1-p)}$$

Y su varianza:

$$V(\xi) = r\frac{p}{(1-p)^2}$$

Si el parámetro r de una distribución binomial negativa tomase el valor de 1, obtendríamos una variable que se ajustaría a una distribución geométrica donde se producirían k fracasos hasta lograr un éxito.

Por último, es importante resaltar que la distribución binomial negativa cumple la propiedad aditiva, ya que si las variables aleatorias $\xi_1, \xi_2, \ldots, \xi_n$ tienen distribución binomial negativa su suma seguirá otra distribución binomial negativa de la forma:

$$BN\left(\sum_{j=1}^{n} r_j\,, p\right)$$

1.2 Distribuciones para modelar la severidad

La severidad se refiere a las pérdidas monetarias derivadas de la realización de un suceso. En la práctica, para modelar esta variable aleatoria, se utilizan distribuciones continuas que tomen valores positivos. Usualmente, esta variable aleatoria puede modelarse utilizando mixturas de distribuciones, ya que muchas compañías trabajan por módulos (por ejemplo en el ramo de autos, el módulo de un taller) y estos valores corresponderían a puntos de masa, modelando el resto de siniestros de forma continua.

Las distribuciones de probabilidad más comúnmente usadas en la práctica son:

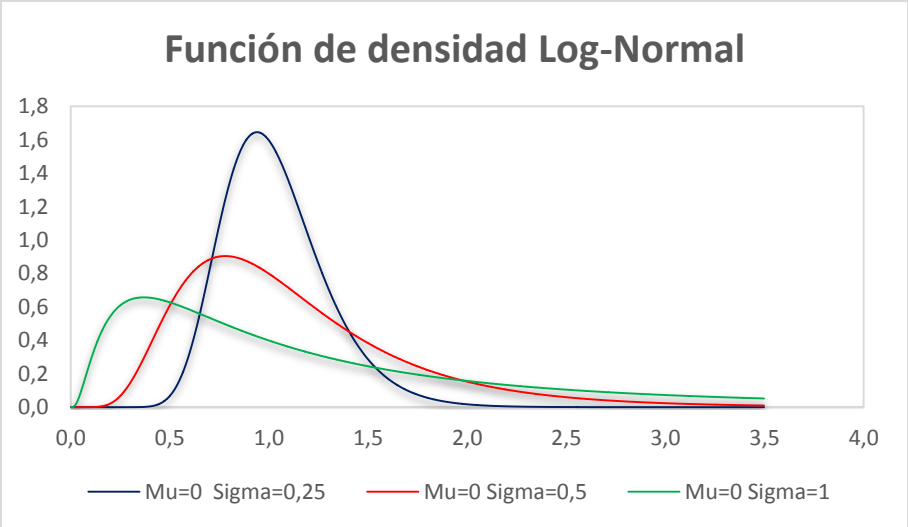
1.2.1 Distribución log-Normal

La distribución log-normal es usada habitualmente para modelar siniestros de cola larga o gruesa, es decir, donde se den valores muestrales muy alejados de la media o coeficientes de asimetría positivos muy elevados.

Una distribución log-normal es una distribución de probabilidad de una variable aleatoria cuyo logaritmo está normalmente distribuido, o lo que es lo mismo si ξ es una variable aleatoria que sigue una distribución normal, entonces e^{ξ} sigue una log-normal.

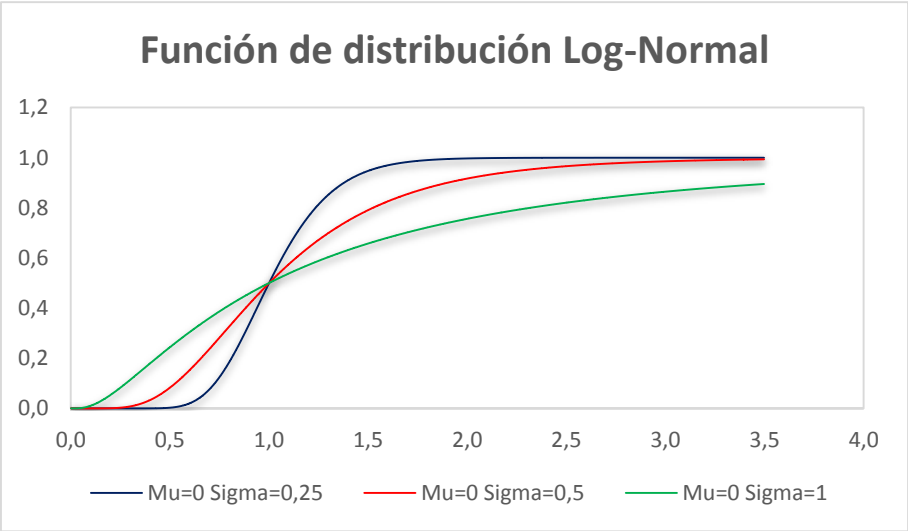
La función de densidad de probabilidad de una variable aleatoria que se distribuya log-normalmente será:

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



Fuente: elaboración propia

Y cuya función de distribución tendrá la siguiente forma:



Fuente: elaboración propia

La esperanza matemática de una variable distribuida lognormalmente será:

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

Y la varianza estará definida de la siguiente forma:

$$V(\xi) = (e^{\sigma^2} - 1)e^{2\pi+\sigma^2}$$

1.2.2 Distribución gamma

La distribución gamma es muy utilizada habitualmente en la práctica para modelar siniestros de cola corta o exponencial, es decir, aquellos que no presentan valores muestrales muy alejados de la media. Esta distribución suele ajustarse bien a riesgos de carácter no catastrófico.

Antes de caracterizar esta distribución, definiré la función gamma: se trata de una función que extiende el concepto de factorial a los números complejos.

$$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}\, e^{-x} dx, para\ \alpha > 0$$

Además si α=n y n es un entero positivo, tenemos:

$$\Gamma(n) = (n-1)!$$

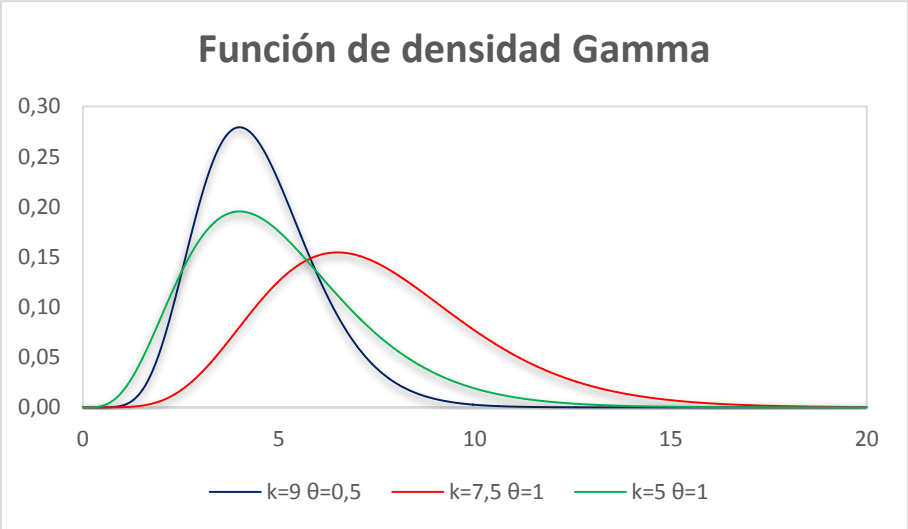
La función de densidad de una variable aleatoria distribuida conforme a una gamma es biparamétrica, dependiendo de los parámetros forma (k) y escala (θ). El parámetro de forma sitúa la máxima intensidad de probabilidad y el parámetro de escala determina la

forma o alcance de la asimetría positiva desplazando la densidad de probabilidad en la cola derecha. La forma de la función de densidad es:

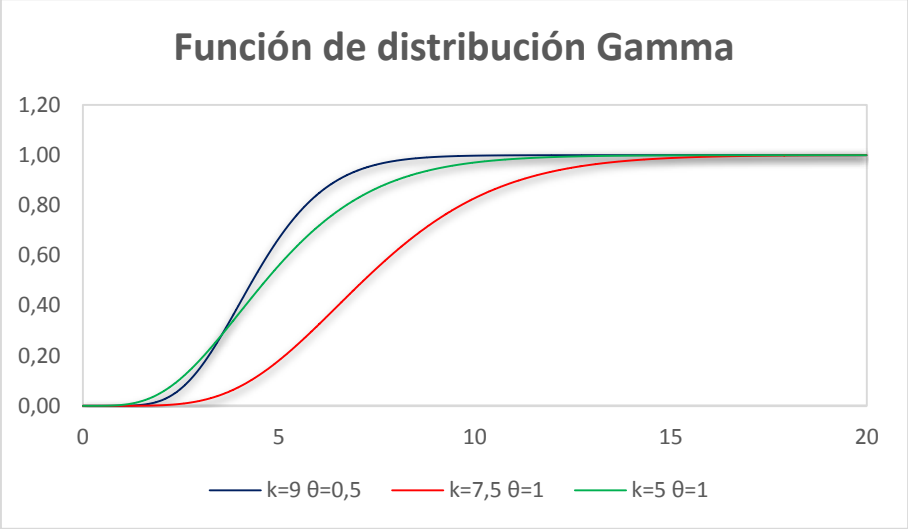
$$f(x,k,\theta) = \frac{x^{k-1}e^{-\frac{x}{\theta}}}{\theta^k \Gamma(k)}$$

Alternativamente, la densidad de la distribución gamma puede ser representada con el parámetro de forma (α) y el inverso de la escala (β):

$$f(x,\alpha,\beta) = \frac{\beta^\alpha x^{\alpha-1}e^{-x\beta}}{\Gamma(\alpha)}$$



Fuente: elaboración propia



Fuente: elaboración propia

La función característica de la variable aleatoria gamma vendrá definida por:

$$\varphi(t) = E\left(e^{it\xi}\right) = \int_0^\infty e^{itx} f(x)dx = \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty e^{itx} x^{\alpha-1} e^{-\beta x} dx = \left(1 - \frac{it}{\beta}\right)^{-\alpha}$$

La esperanza matemática y la varianza de una variable aleatoria distribuida con una gamma son respectivamente:

$$E(\xi) = \frac{\alpha}{\beta}$$

$$V(\xi) = \frac{\alpha}{\beta^2}$$

La distribución gamma cumple la propiedad aditiva o reproductiva. Si n variables aleatorias independientes ξ_j se distribuyen conforme a una gamma G(α_j,β) y definimos la variable $\mu = \xi_1 + \xi_2 + \cdots + \xi_n$, cuya función característica es el producto de las funciones características de cada variable, la función de densidad de la variable μ será también una gamma:

$$\varphi_\mu(t) = \prod_{j=1}^{n} \varphi_{\xi_j}(t) = \prod_{j=1}^{n} \left(1 - \frac{it}{\beta}\right)^{-\alpha_j} = \left(1 - \frac{it}{\beta}\right)^{-\sum_{j=1}^{n} \alpha_j}$$

Por lo que la variable $\mu \sim G(\sum_{j=1}^{n} \alpha_j, \beta)$

Por último como casos particulares de la distribución gamma tenemos:

- Cuando el parámetro α es igual a 1, tendremos la distribución exponencial.
- Si además α es un entero, obtendremos la distribución erlang que describe la duración del tiempo transcurrido hasta que aparecen x sucesos que siguen una distribución de Poisson.
- Si $\alpha = \frac{n}{2}$ y $\beta = \frac{1}{2}$ obtendremos la distribución chi-cuadrado.

1.2.3 Distribución de Pareto

La distribución de Pareto es una de las más usadas a la hora de trabajar con riesgos catastróficos ya que es una distribución de cola larga y, refleja por tanto, siniestralidades muestrales más alejadas de la media.

Su función de densidad es del tipo:

$$f_x(x) = \begin{cases} \frac{\alpha\theta^\alpha}{x^{\alpha+1}}, & x \geq \theta \\ 0, & en\ otro\ caso \end{cases}$$

El parámetro θ es un indicador de posición y el parámetro α es un indicador de dispersión, cuanto mayor es se obtienen densidades más concentradas en las proximidades del mínimo.

La esperanza matemática de una variable aleatoria que sigue una distribución de Pareto es:

$$E(\xi) = \int_0^\infty xf(x)dx = \int_0^\infty x\frac{\alpha\theta^\alpha}{x^{\alpha+1}}dx = \alpha\theta^\alpha\int_0^\infty x^{-\alpha}\,dx = \frac{\alpha}{\alpha-1}\theta$$

Para que la esperanza exista, α tiene que ser mayor que 1.

La varianza quedaría definida por:

$$E(\xi^2) = \int_0^\infty x^2\frac{\alpha\theta^\alpha}{x^{\alpha+1}}dx = \alpha\theta^\alpha\int_0^\infty x^{-\alpha+1}dx = \frac{\alpha}{\alpha-2}\theta^2\ \forall\,\alpha > 2$$

$$V(\xi) = \frac{\alpha}{\alpha-2}\theta^2 - (\frac{\alpha}{\alpha-1}\theta)^2 = \frac{\alpha\theta^2}{(\alpha-1)^2(\alpha-2)}$$

1.2.4 Distribución de Weibull

La distribución de Weibull tiene su aplicación en la modelización de procesos que involucren riesgo, por eso es muy utilizada para modelar pérdidas. Es una distribución paramétrica que depende de un parámetro de forma k y de un parámetro de tasa λ.

Esta distribución asume que la probabilidad de ocurrencia de un suceso cambia con el paso del tiempo, de manera que:

- Si k=1 la probabilidad es constante (exponencial).
- Si k>1 la probabilidad es creciente.
- Si k<1 la probabilidad es decreciente.

La función de densidad de probabilidad de una variable aleatoria que sigue una distribución de Weibull es de la forma:

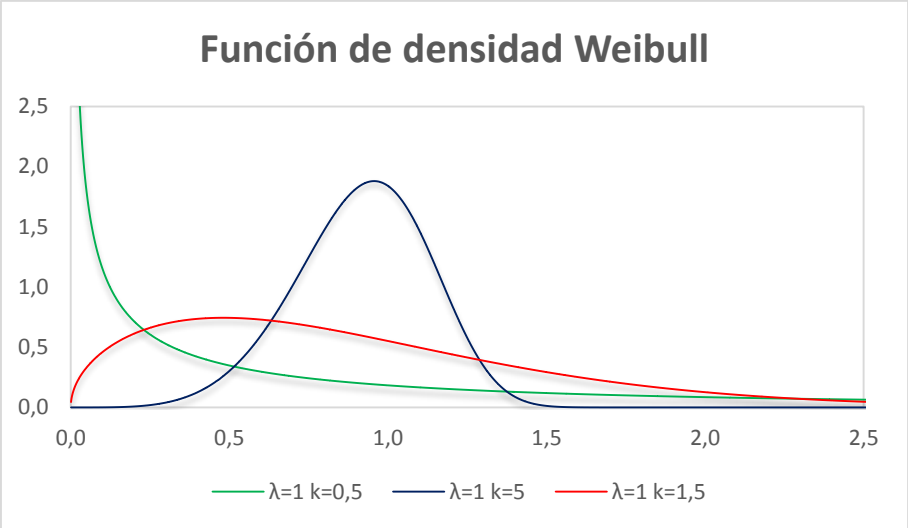
$$f(x,\lambda,k) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}e^{-(x/\lambda)^k}\ \forall\,x \geq 0$$

Su función de distribución viene dada por:

$$F(x,\lambda,k) = 1 - e^{(x/\lambda)^k}\ \forall\,x \geq 0$$

Siendo su hazard rate:

$$h(x,k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}$$



Fuente: elaboración propia

La esperanza matemática y varianza de una variable aleatoria que sigue una distribución de Weibull son respectivamente:

$$E(\xi) = \lambda\Gamma\left(1+\frac{1}{k}\right)$$

$$V(\xi) = \lambda^2\left(\Gamma\left(1+\frac{2}{k}\right) - \Gamma^2\left(1+\frac{1}{k}\right)\right)$$

La distribución de Weibull mantiene una relación con otras distribuciones:

- Si k=1, entonces la distribución de Weibull coincidirá con una distribución exponencial de parámetro 1/λ.
- La distribución de Weibull también puede caracterizarse a través de una distribución uniforme: si una variable aleatoria ξ sigue una distribución uniforme, entonces $k(-\ln(\xi))^{1/\xi}$ seguirá una distribución de Weibull de parámetros λ y k.

2
Medidas de riesgo

Es sabido que una compañía aseguradora puede incurrir en grandes pérdidas ya que el riesgo es su materia prima. Por ello, las medidas de riesgo han ido cobrando una especial relevancia en los últimos años.

Las medidas de riesgo representan el riesgo en términos de potenciales pérdidas económicas. Estas medidas están basadas en la distribución de pérdidas de una cartera por lo que será básico tener bien caracterizada esta distribución.

Las primas que cobra una entidad aseguradora están muy relacionadas con las pérdidas potenciales, por ello, una correcta medición de riesgo cobra gran importancia a la hora de determinar el precio del mismo.

Como se explicó en el tema anterior, la variable aleatoria *y* representa el coste agregado de una cartera. Esta variable es una conjunción de las variables aleatorias frecuencia y severidad, por lo que la manera básica de calcular una prima será aplicar el principio de prima del valor esperado:

$$\pi(y) = (1 + k_1)E(y)$$

Siendo,

- k_1 el recargo de seguridad que se cobrará al cliente.
- $E(y) = E(N)E(S)$

Sin embargo, tal y como se ha mencionado anteriormente, las entidades aseguradoras suelen tarificar usando no sólo las perdidas esperadas, sino que suelen aplicar un recargo por las pérdidas no esperadas de la siguiente manera:

$$\pi(y) = (1 + k_1)E(y) + k_2\rho(-y)$$

Donde,

- k_2 es el recargo por riesgo.
- $\rho(-y)$ será la medida de riesgo que aplique la entidad aseguradora.

2.1 Propiedades de las medidas de riesgo

Las medidas de riesgo tienen una serie de propiedades que se expondrán a continuación. Si *y* es una variable aleatoria que representa el coste agregado de una cartera, entonces una medida de riesgo $\rho(y)$ podrá ser:

- Invariante por traslaciones: si a y le sumamos una constante k, entonces el riesgo decrecerá en k. Formalmente,
$$\rho(y + k) = \rho(y) - k$$

- Homogénea: ρ es independiente de la escala:
$$\rho(ky) = k\rho(y)$$
Siempre que $k \geq 0$.

- Invariante por ley: $\rho(y_1) = \rho(y_2)$ siempre que y_1 e y_2 tengan la misma distribución.

- Decreciente: a mayor riqueza le corresponderá menor riesgo, ya que si la probabilidad de $y_1 \geq y_2$ es 1, entonces $\rho(y_1) \leq \rho(y_2)$.

- Aditiva para riesgos comonótonos: dos riesgos serán comonótonos cuando la probabilidad $(\Delta y_1)(\Delta y_2) \geq 0$ sea 1, es decir, van en la misma dirección. Esta propiedad indica que el riesgo de una suma de los riesgos es igual a la suma de los riesgos siempre y cuando no haya efecto de diversificación:
$$\rho(y_1 + y_2) = \rho(y_1) + \rho(y_2)$$

- Subaditiva: con independencia de que dos riesgos sean comonótonos, se cumplirá que:
$$\rho(y_1 + y_2) \leq \rho(y_1) + \rho(y_2)$$

- Dominante de la esperanza matemática:
$$\rho(y) \geq -E(y)$$

- Una medida de riesgo será coherente cuando cumpla las siguientes propiedades: invariante por traslaciones, homogeneidad, decreciente y subaditividad.

2.2 Medidas de riesgo más utilizadas en tarificación

A continuación se realizará una breve descripción de las medidas de riesgo más utilizadas en la práctica actuarial y, más concretamente, a la hora de tarificar. Estas medidas son:

2.2.1 Value at Risk (VaR)

El VaR es una de las medidas de riesgo más utilizada en la práctica, principalmente debido a las directivas de Basilea III y Solvencia II.

Podríamos definir el VaR como la pérdida máxima que podría sufrir una entidad aseguradora en un determinado horizonte temporal y bajo un determinado nivel de confianza, es decir, el valor correspondiente al α-ésimo cuantil (q_α) de la función de distribución de pérdidas. Formalmente, si F es la función de distribución del coste agregado de una cartera y, entonces:

$$VaR_{1-\alpha}(y) = -Inf\{x : F(x) > \alpha\}$$

Además si la función de distribución es continua e invertible, entonces:

$$VaR_{1-\alpha}(y) = -F^{-1}(\alpha)$$

Como mencioné con anterioridad será clave tener bien caracterizada la distribución del coste agregado de la cartera.

El VaR es una medida de riesgo universal, debido principalmente a su fácil interpretación, ya que resume en un solo número todas las posibles fuentes de riesgo, sin embargo, no está exento de problemas. El más importante de todos ellos es la falta de subaditividad, ya que en ocasiones tiende a no diversificar riesgos en algunos problemas de optimización. Esta falta de subaditividad está presente en todas las distribuciones que no sean de tipo elíptico como la normal, lo que puede llevar a resultados contradictorios. Por ello se dice que el VaR no es una medida coherente de riesgo.

2.2.2 Average Value at Risk (CVaR)

El CVaR es una medida de riesgo que cuantifica las pérdidas que pueden producirse en las colas de las distribuciones, se trata de la pérdida esperada en aquellos casos en que las pérdidas excedan el valor del VaR, por lo que el valor del VaR nunca será superior al del CVaR. Es decir, se trata del promedio del VaR a niveles de confianza mayores:

$$CVaR_{1-\alpha}(y) = \frac{1}{\alpha}\int_0^\alpha VaR_{1-t}(y)dt$$

El CVaR es una medida de riesgo más prudente que el VaR ya que tiene en cuenta toda la cola mala de la distribución, además se trata de una medida coherente que no tiene problemas a la hora de diversificar riesgos, incluso en distribuciones no elípticas es fácil de controlar y optimizar.

Una medida de riesgo derivada de la anterior es el Weighted CVaR (WCVaR). Esta medida permite involucrar más de un nivel de confianza, por lo que no sólo se centra en la cola mala de la distribución, sino que, en función de los pesos asignados, se fijará más en toda la distribución de y.

$$WCVaR_{(1-\alpha_1,\dots 1-\alpha_n;k_1\dots k_n)}(y) = \sum_{i=1}^{n} k_i \, CVaR_{1-\alpha_i}(y)$$

Lógicamente, al igual que el CVaR, el WCVaR es una medida coherente de riesgo.

2.2.3 Riesgos dados por funciones de distorsión

Estas medidas de riesgo, a diferencia del WCVaR, incorporan toda la distribución de y. Una función de distorsión es una función continua, creciente y cóncava:

$$g: [0,1] \to [0,1]$$

Tal que $g(0) = 0$ y $g(1) = 1$ y, además, g es derivable en $(0,1)$.

Si se cumplen estas propiedades, entonces una medida de riesgo dada por una función de distorsión quedará definida por:

$$\rho_g(y) = \int_0^1 VaR_{1-t}(y) g'(t) dt$$

Los dos casos más importantes son la Medida de Wang (W_a) y la Dual Power Transform (DTP_a).

La Medida de Wang de parámetro $a > 0$ aparece cuando:

$$g(t) = \phi(\phi^{-1}(t) + a)$$

Siendo ϕ la función de distribución $N(0,1)$.

La Dual Power Transform de parámetro $a > 1$ se da cuando:

$$g(t) = 1 - (1-t)^a$$

Para ambas medidas el parámetro a es un parámetro de aversión al riesgo, es decir, cuanto mayor sea el parámetro, se tendrá una mayor aversión al riesgo y, por tanto, la cola mala se tendrá más en cuenta a la hora del cálculo de la medida.

2.2.4 Medidas espectrales

Las medidas de riesgo espectrales aparecen cuando sustituimos el VaR por el CVaR en las medidas de riesgo dadas por funciones de distorsión.

Si $0 < \alpha_i < 1,\ k_i > 0,\ \sum_{i=1}^{n} k_i \leq 1$, y además

$$g: [0,1] \rightarrow [0,1]$$

Es una función continua tal que $g(0) = 0$, g es derivable en $(0,1)$, g es no decreciente ($g'(t) \geq 0$) y $g(1) + \sum_{i=1}^{n} k_i = 1$, podremos definir la medida de riesgo como:

$$\rho_{(1-\alpha,k,g)}(y) = \sum_{i=1}^{n} k_i \, CVaR_{(1-\alpha_i)}(y) + \int_0^1 CVaR_{1-t}(y) g'(t) dt$$

Estas medidas se diferencian de las anteriores en dos puntos muy importantes: no imponen que la función g sea estrictamente creciente ni que sea cóncava.

Los dos casos más importantes son:

- Spectral Dual Power Transform ($SDPT_a$): si $n = 0$ y $g(t) = 1 - (1-t)^a$ con $a > 0$
- Spectral Wang (SW_a): si $n = 0$ y $g(t) = \phi(\phi^{-1}(t) + a)$ con a arbitrario.

3
Modelos predictivos

La técnica de tarificación habitual utilizada por las compañías de seguros no vida es aquella basada en el uso de los modelos predictivos.

Estas técnicas se basan en el análisis de los datos actuales recogidos en las bases de datos para poder realizar predicciones sobre futuros sucesos. Dichas predicciones distan de ser verdades absolutas.

Los modelos predictivos explotan los patrones de comportamiento identificados en el pasado para poder cuantificar riesgos futuros. Estos modelos capturan la relación entre una serie de variables independientes (factores de riesgo) y la variable a explicar.

Antes de profundizar en la explicación, es conveniente explicar la diferencia entre las dos funciones básicas de un modelo:

- Explicar: a diferencia de los modelos de predicción, los modelos explicativos, se centran en describir de la mejor manera posible, la relación subyacente entre la variable dependiente y las variables independientes. Lógicamente, cuantos más factores de riesgo utilicemos en el modelo, mejor quedará perfilado el riesgo.
- Predecir: estos modelos actúan en sentido contrario. En ellos, lo que se pretende es poder predecir un evento futuro en base a una serie de datos históricos. Al contrario que los modelos explicativos, los modelos predictivos, requieren del uso de pocos factores de riesgo (de acuerdo al principio de parsimonia) para poder tener clústers muy nutridos y que nuestras predicciones puedan ser representativas en media.

Las técnicas predictivas más utilizadas en el campo actuarial son, entre otras:

- Modelos Lineales Generalizados (GLM): se trata de la técnica de modelización por excelencia en las compañías de seguros no vida. Es una generalización flexible de la regresión lineal que relaciona la distribución aleatoria de la variable dependiente con la parte sistemática a través de la función de enlace.
- Redes neuronales: las redes neuronales artificiales (RNA) son un sistema de interconexión de neuronas que colaboran entre sí para producir un estímulo de salida.
- Árboles de decisión: es un modelo de predicción muy utilizado en el ámbito de la inteligencia artificial. Se trata de una técnica que permite analizar decisiones secuenciales basada en el uso de resultados y probabilidades asociadas.

3.1 Modelos Lineales Generalizados (GLM)

Los Modelos Lineales Generalizados son una de las técnicas de modelización más utilizadas en las entidades aseguradoras.

El modelo de regresión clásico se basa en los siguientes supuestos:

- Los errores se distribuyen normalmente.
- Homocedasticidad.
- La variable dependiente se relaciona linealmente con la(s) variable(s) explicativa(s).

Formalmente, tendríamos:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \xi$$

Donde,

$$\xi \sim N(0, \sigma)$$

Si tomamos esperanzas:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Como es lógico observar, la relación entre la variable dependiente y el predictor lineal, es la función identidad.

Sin embargo, en la práctica, las restricciones del modelo de regresión simple son muy fuertes y pocas veces se cumplen. Por ello, en la práctica se utilizan los Modelos Lineales Generalizados, que son una extensión del modelo de regresión simple que permiten utilizar distribuciones no normales de los residuos.

Algunas de las situaciones más comunes en el uso de los Modelos Lineales Generalizados son:

- Cuando se trabaja con variables de conteo (por ejemplo, el número de siniestros).
- Cuando la variable de conteo viene expresada como una proporción (porcentaje de siniestros graves).
- Si trabajamos con variables binarias (renovación o no renovación).

Los Modelos Lineales Generalizados permiten utilizar distintos tipos de distribuciones a la normal, por ejemplo:

- Poisson, muy utilizada para conteos.
- Binomial, cuando tratamos de estimar proporciones.
- Gamma, utilizada cuando en nuestros datos, la varianza aumenta según lo hace la media de la muestra.
- Exponencial, muy útil en análisis de la supervivencia.

3.1.1 Estructura de un GLM

Un Modelo Lineal Generalizado tiene tres componentes básicos:

- Componente aleatorio: el componente aleatorio consiste en la variable aleatoria y con observaciones independientes $(y_1, y_2 \dots y_n)$. Los elementos de este vector aleatorio observable son independientes y están idénticamente distribuidos con una función de distribución perteneciente a la familia exponencial uniparamétrica:

$$f_\theta(y) = \exp[\{y\theta - b(\theta)\}/a(\phi) + c(y,\theta)]$$

Donde $a(*), b(*) y\ c(*)$ son funciones arbitrarias, ϕ un parámetro arbitrario de escala y θ el parámetro canónico de la distribución.

Por ejemplo, para comprobar que la distribución normal pertenece a la familia exponencial:

$$f_\mu(y) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right] = \exp\left[\frac{-y^2+2y\mu-\mu^2}{2\sigma^2} - \log\left(\sigma\sqrt{2\pi}\right)\right] =$$

$$\exp\left[\frac{y\mu - \frac{\mu^2}{2}}{\sigma^2} - \frac{y^2}{2\sigma^2} - \log\left(\sigma\sqrt{2\pi}\right)\right]$$

Donde $\theta = \mu,\ b(\theta) = \theta^2/2 = \mu^2/2,\quad a(\phi) = \phi = \sigma^2$ y $c(\phi, y) = -\frac{y^2}{2\phi} - \log(\sqrt{\phi 2\pi}) = -\frac{y^2}{2\sigma^2} - \log(\sqrt{\sigma 2\pi})$

Es posible obtener la expresión para la media y varianza de una distribución perteneciente a la familia exponencial en términos de a, b y ϕ.

El logaritmo de verosimilitud de θ, dada una y en particular, es simplemente $\log[f_\theta(y)]$ como una función de θ. Es decir,

$$l(\theta) = [y\theta - b(\theta)]/a(\phi) + c(y,\phi)$$

Por lo que,

$$\frac{\partial l}{\partial \theta} = [y - b'(\theta)]/a(\phi)$$

Considerando l como una variable aleatoria, mediante la sustitución de la observación y por la variable aleatoria Y permite que el valor esperado de $\frac{\partial l}{\partial \theta}$ pueda ser evaluado:

$$E\left(\frac{\partial l}{\partial \theta}\right) = [E(y) - b'(\theta)]/a(\phi)$$

Y sabiendo que:

$$E_0\left(\frac{\partial l}{\partial \theta}\Big|_{\theta_0}\right) = 0$$

Implica que:

$$E(y) = b'(\theta)$$

Si diferenciamos la verosimilitud una vez más:

$$\frac{\partial^2 l}{\partial \theta^2} = -b"(\theta)/a(\phi)$$

Y la trasladamos al resultado general:

$$E\left(\frac{\partial^2 l}{\partial \theta^2}\right) = -E\left[\left(\frac{\partial\, l}{\partial \theta}\right)^2\right]$$

Obtenemos:

$$b"(\theta)/a(\phi) = E[(y - b'(\theta))^2]/a(\phi)^2$$

Que, finalmente, conduce a:

$$Var(y) = b"(\theta)\, a(\phi)$$

- Componente sistemático: son las variables explicativas utilizadas en la función predictora lineal, es decir, se relacionan como:

$$\alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

Alternativamente, puede expresarse de forma vectorial $(\eta_1, \eta_2 \ldots \eta_n)$ de manera que:

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

Donde x_{ij} es el valor del j-ésimo predictor en el i-ésimo individuo.

- Función link o enlace: se trata de la función del valor esperado de la variable a explicar, dada como una combinación lineal de las variables predictoras, es decir, especifica una función $g(*)$ que relaciona la $E(y)$ con el predictor lineal:

$$g(\mu) = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n$$

De esta manera, la función link relaciona las componentes aleatoria y sistemática.

$$\eta_i = g(\mu_i) = \sum_j \beta_j x_{ij}$$

La función $g(\mu)$ es una función conocida, monótona y diferenciable de η, así:

$$\mu_i = g^{-1}(\eta_i)$$

En la práctica las funciones de enlace más utilizadas son:

Función de enlace	Fórmula	Uso
Identidad	μ	Datos continuos con errores normales
Logarítmica	$\log(\mu)$	Conteos con errores Poisson
Logit	$\log(\mu/1-\mu)$	Proporciones (datos entre 0 y 1) con errores binomiales
Recíproca	$1/\mu$	Datos continuos con errores gamma
Raíz cuadrada	$\sqrt{\mu}$	Conteos
Exponencial	μ^n	Funciones de potencia

3.1.2 Estimación de los parámetros de un GLM

A la hora de estimar los parámetros del modelo, si la variable dependiente se distribuye conforme a una distribución de la familia exponencial, se suele utilizar el método de máxima verosimilitud.

Teniendo un vector de observaciones $y' = (y_1, y_2 \dots y_n)$, la función de verosimilitud cuantifica la probabilidad de que un vector $\beta \in \Re^p$ haya generado el vector observado.

La función de verosimilitud viene dada por la función de densidad conjunta de las variables aleatorias independientes e idénticamente distribuidas $y_1, y_2 \dots y_n$:

$$L(\beta) = \prod_{i=1}^{n} f_{\theta_i}(y_i)$$

Siendo θ_i el parámetro canónico determinado por μ_i, por lo que el logaritmo de la verosimilitud será:

$$l(\beta) = \sum_{i=1}^{n} \log[\, f_{\theta_i}(y_i)] = \sum_{i=1}^{n} [y_i\theta_i - b_i(\theta_i)]/a_i(\phi) + c_i(\phi, y_i)$$

En la práctica, se consideran aquellos casos donde $a_i(\phi) = \phi/\omega_i$, siendo ω_i una constante. Por lo que:

$$l(\beta) = \sum_{i=1}^{n} \omega_i[y_i\theta_i - b_i(\theta_i)]/\phi + c_i(\phi, y_i)$$

A continuación, se procederá a maximizar $l(\beta)$ tomando primeras diferencias parciales, igualando la expresión a cero y resolviendo para β:

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \omega_i \left(y_i \frac{\partial \theta_i}{\partial \beta_j} - b'_i(\theta_i) \frac{\partial \theta_i}{\partial \beta_j} \right)$$

Que por la regla de la cadena:

$$\frac{\partial \theta_i}{\partial \beta_j} = \frac{\partial \theta_i}{\partial \mu_i} \frac{\partial \mu_i}{\partial \beta_j}$$

Luego, diferenciando, obtenemos:

$$\frac{\partial \mu_i}{\partial \theta_i} = b"_i(\theta_i) \Rightarrow \frac{\partial \theta_i}{\partial \mu_i} = \frac{1}{b"_i(\theta_i)}$$

Implicando:

$$\frac{\partial l}{\partial \beta_j} = \frac{1}{\phi} \sum_{i=1}^{n} \frac{[y_i - b'_i(\theta_i)]}{b"_i(\theta_i)/\omega_i} \frac{\partial \mu_i}{\partial \beta_j}$$

Finalmente, si sustituimos en esta última expresión, obtendremos la ecuación para estimar los β:

$$S = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \ \ \forall j$$

3.1.3 Deviance

En la práctica, cuando se trabaja con Modelos Lineales Generalizados, resulta muy útil tener una medida que se pueda interpretar de manera análoga a la suma residual de cuadrados en los modelos de regresión lineal clásicos. Esta medida recibe el nombre de deviance.

Se podría definir la deviance como:

$$D = 2\left[l\left(\hat{\beta}_{max}\right) - l\left(\hat{\beta}\right)\right]\phi = \sum_{i=1}^{n} 2\omega_i[y_i(\tilde{\theta}_i - \hat{\theta}_i) - b(\tilde{\theta}_i) + b(\hat{\theta}_i)]$$

Donde, $l\left(\hat{\beta}_{max}\right)$ hace referencia a la máxima verosimilitud del modelo saturado (aquel con un parámetro por cada observación), $\tilde{\theta}_i$ y $\hat{\theta}_i$ son los estimadores de máxima verosimilitud de los parámetros canónicos para el modelo saturado y para el modelo de análisis respectivamente.

Un concepto relacionado con la deviance, es la deviance en escala, que depende del parámetro de escala:

$$D^* = \frac{D}{\phi}$$

Nuestro modelo, por tanto, tendrá un buen ajuste cuando:

$$D^* \sim \chi^2_{n-p}$$

3.1.4 Residuos en un GLM

Dentro de la modelización, el estudio de los residuos es una de las partes más importantes. En los modelos lineales clásicos, para analizar los residuos, simplemente estudiaremos el modelo residual, que contiene toda la información no explicada por la parte sistemática del modelo.

En el caso de los Modelos Lineales Generalizados, este estudio también es crucial. Sin embargo, no es tan fácil. No podremos observar directamente los residuos producidos por el modelo $\hat{\epsilon}_i = y_i - \hat{\mu}_i$, debido a la dificultad en la comprobación de la validez de la relación media varianza de los mismos. Por ello, es habitual trabajar con residuos estandarizados en los Modelos Lineales Generalizados, ya que si las suposiciones del modelo son correctas, los residuos estandarizados deben tener aproximadamente la misma varianza y comportarse, en la medida de lo posible, como residuos de un modelo de regresión lineal simple.

La manera más utilizada en la práctica, es trabajar con los residuos estandarizados de Pearson:

$$\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Debiendo tener aproximadamente media cero y varianza ϕ, si el modelo es correcto.

Sin embargo, al trabajar con dato real, suele suceder que la distribución de los residuos de Pearson, sea muy asimétrica alrededor de cero, por lo que su comportamiento podría distar bastante de lo esperado en un modelo de regresión ordinario. Es por ello, que se utilizan los residuos de la deviance.

Si denotamos por d_i al componente de la deviance que aporta la i-ésima observación, tendremos:

$$D = \sum_{i=1}^{n} d_i$$

Y, de manera análoga al modelo lineal ordinario, definiremos:

$$\hat{e}_i^d = sign(y_i - \hat{\mu}_i)\sqrt{d_i}$$

3.1.5 Modelos Lineales Generalizados para conteos

El modelo básico para datos de conteo es aquel modelo en el que la variable respuesta se distribuye de acuerdo a una distribución de Poisson y cuya función link es la función logaritmo.

Es sabido que una de las propiedades de la distribución de Poisson es que:

$$E(Y) = VAR(Y) = \mu$$

Como he comentado anteriormente, en este modelo se utiliza el logaritmo de la media para la función link:

$$\log(\mu) = \alpha + \beta x$$

De manera que,

$$\mu = e^{\alpha+\beta x}$$

En un modelo de Poisson, para obtener la deviance, tendremos que calcular:

$$-2\log(p(y))$$

Sin embargo, al trabajar con datos reales, la propiedad de igualdad entre media y varianza no suele cumplirse a menudo. Estaríamos, por tanto, bajo un modelo sobredisperso.

Esta situación se debe, habitualmente, a la existencia de heterogeneidad en la muestra. Podría interpretarse como una mixtura de distribuciones Poisson.

Modelo Cuasi-Poisson

Una manera de solucionar la sobredispersión en los datos de un modelo de Poisson, es introducir un parámetro de dispersión en el modelo, por lo que la varianza condicional de la variable respuesta será ahora:

$$V(y_i/\eta_i) = \phi\mu_i$$

Lógicamente, si $\phi > 1$, la varianza condicional de y, aumentará más rápidamente que la media.

Para el cálculo del parámetro de dispersión utilizaremos un estimador del método de los momentos, de manera que en el modelo Cuasi-Poisson, dicho parámetro será:

$$\tilde{\phi} = \frac{1}{n-k-1}\sum \frac{(y_i - \widehat{\mu_i})^2}{\widehat{\mu_i}}$$

Donde $\widehat{\mu_i} = g^{-1}(\widehat{\eta_i})$ es el valor esperado de y_i

Modelo Binomial Negativo

Nos encontramos ante otro modelo derivado de la distribución de Poisson que funciona muy bien cuando trabajamos con sobredispersión.

Asumimos que la variable respuesta y se distribuye conforme a una Poisson, pero suponemos que su media es una variable aleatoria no observable según una gamma de media μ_i y parámetro de dispersión $\frac{1}{k}$

$$P(y/k,\mu) = \frac{\Gamma(y+k)}{\Gamma(k)\Gamma(y+1)}\left(\frac{k}{\mu+k}\right)^k\left(1-\frac{k}{\mu+k}\right)^y$$

Con,

$$E(y) = \mu$$

$$VAR(y) = \mu + \frac{\mu^2}{k}$$

Si el parámetro de dispersión $\frac{1}{k} \to 0$ entonces $VAR(y) = \mu$, por lo que la distribución binomial negativa convergirá a una Poisson.

3.1.6 Modelos Lineales Generalizados para datos binarios

Son otros de los modelos más utilizados en la práctica. En muchas situaciones, queremos modelizar variables respuestas y de carácter dicotómico, es decir, que puedan tomar sólo dos posibles valores del tipo $éxito = 1$ o $fracaso = 0$.

Por lo que la variable respuesta se distribuirá conforme a una variable aleatoria Bernoulli de parámetro: $y \sim B(1, \pi)$. Luego:

$$f(y/\pi) = \pi^y (1-\pi)^{1-y} = (1-\pi)\left(\frac{\pi}{1-\pi}\right)^y = (1-\pi)\exp\left[y\, log\left(\frac{\pi}{1-\pi}\right)\right]$$

Con $y = 0{,}1$.

El parámetro natural será:

$$Q(\pi) = log\left(\frac{\pi}{1-\pi}\right) = logit(\pi)$$

Y además:

$$E(y) = P(y = 1) = \pi(x)$$

$$VAR(y) = \pi(x)(1-\pi(x))$$

Dependientes de p variables explicativas $x = (x_1, x_2 \dots x_p)$.

Por lo tanto, cuando la variable aleatoria y es una variable binaria, un modelo análogo al de regresión lineal será:

$$\pi(x) = \alpha + \beta x$$

Denominado modelo de probabilidad lineal, ya que la probabilidad de éxito aumenta linealmente con x. Sin embargo, este modelo no está exento de problemas estructurales, ya que aunque la probabilidad está acotada entre 0 y 1, el modelo puede devolver valores $\pi(x) > 1$ y valores $\pi(x) < 0$.

Además las relaciones entre $\pi(x)$ y x no suelen ser lineales, ya que el cambio en x tiene menor impacto cuando π está más próximo a 0 ó 1 que cuando está más centrado, es decir, la relación tiene forma sigmoidal.

Para modelizar este fenómeno, se suele utilizar la siguiente expresión:

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}$$

Conocida como función logística, de la que provienen los modelos Logit:

$$1-\pi(x) = 1 - \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}} = \frac{1}{1+e^{\alpha+\beta x}}$$

Luego,

$$\frac{\pi(x)}{1-\pi(x)} = e^{\alpha+\beta x} \Rightarrow \log\left(\frac{\pi(x)}{1-\pi(x)}\right) = \alpha + \beta x$$

La función de enlace de π se denominará función Logit, asegurando que no exista ningún problema con el rango de posibles valores de π.

Además, el parámetro β, determinará la velocidad de incremento/decremento de la curva.

Existe otro modelo muy utilizado en la práctica para trabajar con datos binarios: los modelos Probit.

La función de enlace de estos modelos, transforma probabilidades en valores de la distribución normal estándar:

$$\pi(x) = \phi(\alpha + \beta x) \Rightarrow \phi^{-1}\big(\pi(x)\big) = \alpha + \beta x$$

Donde,

$$\phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{\left(\frac{-t^2}{2}\right)} dt$$

3.2 Modelos Aditivos Generalizados (GAM)

Como hemos visto anteriormente, los modelos lineales generalizados asumen una relación lineal entre la esperanza de la variable a explicar y las variables explicativas independientes a través de una función link. Sin embargo, T.J. Hastie y R.J. Tibshirani, introdujeron en 1990 otro modelo más flexible: los modelos aditivos generalizados.

Estos modelos reemplazan el predictor lineal $\sum_{j=1}^{n} \beta_j x_j$ por una suma de funciones suaves de la forma: $\sum_{j=1}^{n} f_j(x_j)$.

Estas funciones no paramétricas pueden ser estimadas a través de splines cúbicos mediante un método iterativo conocido como algoritmo de back fitting.

Pero, ¿cuáles son las ventajas de estos modelos frente a los ya estudiados GLM?

GLM	
Ventajas	**Desventajas**
Relajan los supuestos de normalidad y homocedasticidad	Sustanciales pérdidas en la precisión si la familia de la variable a explicar no es conocida
Funciones link alternativas	
Evitan problemas de retransformación	Limitación a causa de la linealidad

GAM	
Ventajas	**Desventajas**
Flexibilidad	Sobreajuste
Eficaz para trabajar con relaciones no lineales	Cálculo intensivo y complejo

Pasemos a estudiar más detalladamente estos modelos, los modelos aditivos generalizados.

3.2.1 Introducción a los modelos aditivos generalizados (GAM)

Como hemos comprobado anteriormente, el modelo de regresión lineal, es una de las herramientas estadísticas más utilizadas en el análisis de datos.

Esta técnica asume una relación lineal entre una variable aleatoria de la que se quiere realizar una predicción y una serie de variables explicativas independientes. Si denotamos la variable a explicar como y, y las variables explicativas como $x_1, x_2 \dots x_n$ tendremos:

$$\mu_{y|x} = E(y|x_1, x_2 \dots x_n) = f(x_1, x_2 \dots x_n) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n =$$

$$= \beta_0 + \sum_{j=1}^{n} \beta_n x_n$$

Y, en el caso de los GLM, como hemos visto en el capítulo anterior:

$$g(\mu_{y|x}) = \beta_0 + \sum_{j=1}^{n} \beta_n x_n$$

Los modelos aditivos generalizados reemplazan esta relación lineal por funciones suaves:

$$E(Y|x_1, x_2 \dots x_n) = f(x_1, x_2 \dots x_n) = f_0 + f_1(x_1) + f_2(x_2) + \cdots + f_n(x_n) =$$

$$= f_0 + \sum_{j=1}^{n} f_n(x_n)$$

Estas funciones $\widehat{f_n}(x_n)$ son estimadas de manera flexible a través de splines cúbicos suavizados.

3.2.2 Suavizado

El suavizado es una herramienta para poder explicar la tendencia de una variable aleatoria y en función de una o más variables independientes $x_1, x_2 \dots x_n$. Ello produce una estimación de la tendencia menos volátil que la variable aleatoria y por si misma, y será conocida como suavizado (smoother). Llamaremos smooth, a la estimación producida por un suavizado.

El suavizado es una herramienta muy útil en la práctica, ya que permite estimar la dependencia de la esperanza de la variable aleatoria a explicar sobre los predictores.

La propiedad más importante del suavizado es su naturaleza no paramétrica, por lo que no se asume una forma rígida de dependencia entre la variable a explicar y los predictores. Ésta es la mayor diferencia con los modelos lineales generalizados.

Splines cúbicos de suavizado

Son una de las mejores soluciones para el siguiente problema de optimización: de entre todas las funciones $f(x_i)$, todas ellas con segundas derivadas continuas, encontrar aquella que minimice los mínimos cuadrados penalizados. Esta función recibirá el nombre de spline cúbico de suavizado.

$$\sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b [f''(x)]^2 dx$$

Siendo λ una constante y $a \leq x_1 \leq \cdots \leq x_n \leq b$.

El nombre de spline cúbico deriva de que el polinomio de orden 3 ha demostrado en la práctica ser suficiente para el ajuste.

El primer término de la ecuación representa el método de mínimos cuadrados. Utilizando sólo esta parte de la ecuación, el resultado sería una curva interpolada que no sería suave.

La integral $\int [f''(x)]^2 dx$, mide la ondulación de la función $f(x)$.

El parámetro λ es un parámetro de suavizado no negativo. Este parámetro regula el equilibrio entre la bondad del ajuste a los datos y la ondulación de la función.

Cuando $\lambda \to \infty$, el término de penalización cobra mayor importancia, forzando $f"(x) = 0$ y siendo el resultado la línea de mínimos cuadrados.

Cuando $\lambda \to 0$, el término de penalización se hace intrascendente, por lo que una solución será la segunda derivada.

Los valores más grandes de λ producen curvas más suaves, y los valores más pequeños, curvas más onduladas.

3.2.3 Selección automática de parámetros de suavizado

Debemos elegir el parámetro de suavizado con el objetivo de minimizar el spline cúbico.

Para ello, no será necesario minimizar el error cuadrático medio de cada x_i, sino que deberemos centrarnos en una medida global conocida como Average Mean Square Error (AMSE):

$$AMSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(E(\widehat{f_\lambda}(x_i) - f(x_i)\right)^2$$

Siendo $y_i = f(x_i) + \varepsilon_i$, y $\widehat{f_\lambda}(x_i)$ un estimador de $f(x)$.

El error cuadrático predictivo medio (PSE) difiere del AMSE por una función constante δ^2:

$$PSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E(y_i^* - \widehat{f_\lambda}(x_i))^2$$

Siendo y_i^* una nueva observación en x_i.

Se puede demostrar que $PSE = MSE + \sigma^2$ con $E(y_i^* - \widehat{f_\lambda}(x_i))$:

$$E(y_i^* - \widehat{f_\lambda}(x_i) + f(x_i) - f(x_i))^2 =$$

$$= E(y_i^* - f(x_i))^2 + E(\widehat{f_\lambda}(x_i) - f(x_i))^2 + 2E(y_i^* - f(x_i))E(f(x_i) - \widehat{f_\lambda}(x_i)) =$$

$$= E(y_i^* - f(x_i))^2 + E(\widehat{f_\lambda}(x_i) - f(x_i))^2 + 2E(\varepsilon_I^*)E(\varepsilon_I) =$$

$$= \delta_i^2 + E(\widehat{f_\lambda}(x_i) - f(x_i))^2$$

Luego,

$$PSE(\lambda) = \frac{1}{n}\sum_{i=1}^{n} E(y_i^* - \widehat{f_\lambda}(x_i))^2 = \frac{1}{n}\sum_{i=1}^{n}\left(\delta_i^2 + E(\widehat{f_\lambda}(x_i) - f(x_i))^2\right) =$$

$$= \delta^2 + \frac{1}{n}\sum_{i=1}^{n} E(\widehat{f_\lambda}(x_i) - f(x_i)^2 = MSE + \sigma^2$$

La validación cruzada es un método estadístico consistente en dividir una muestra de observaciones en varios subconjuntos, calibrar los subconjuntos para el ajuste del modelo y evaluar la adecuación del modelo. Sin embargo, para que no se produzcan errores, la muestra deberá ser suficientemente grande.

Luego, para construir la validación cruzada (CV) de la suma de cuadrados:

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_\lambda^{-i}(x_i))^2$$

Donde, $\hat{f}_\lambda^{-i}(x_i)$ indica el ajuste en x_i calculado, dejando fuera la i-ésima observación.

Luego,

$$E(y_i - \hat{f}_\lambda^{-i}(x_i))^2 = E(y_i - f(x_i) + f(x_i) - \hat{f}_\lambda^{-i}(x_i))^2 =$$

$$E(y_i - f(x_i))^2 + E(f(x_i) - \hat{f}_\lambda^{-i}(x_i))^2 + E\big(y_i - f(x_i)\big)E(f(x_i) - \hat{f}_\lambda^{-i}(x_i)) =$$

$$\delta_i^2 + E(f(x_i) - \hat{f}_\lambda^{-i}(x_i))^2$$

Siendo,

$$E\big(y_i - f(x_i)\big)E(f(x_i) - \hat{f}_\lambda^{-i}(x_i)) = 0$$

Ya que $\hat{f}_\lambda^{-i}(x_i)$ no involucra a y_i

Por otra parte, si asumimos $\hat{f}_\lambda^{-i}(x_i) \approx \widehat{f_\lambda}(x_i)$:

$$E\big(CV(\lambda)\big) = E\left(\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}_\lambda^{-i}(x_i))^2\right) = \delta^2 + E(f(x_i) - \hat{f}_\lambda^{-i}(x_i))^2 =$$

$$= \delta^2 + E(\widehat{f_\lambda}(x_i) - f(x_i))^2 = PSE$$

Por lo que, como minimizar $CV(\lambda)$ es equivalente a minimizar $PSE(\lambda)$, podemos utilizar $CV(\lambda)$ como parámetro de suavizado.

Para calcular el $CV(\lambda)$, parece que debe ajustarse el modelo n veces, una por cada valor y^{-i} no tenido en cuenta. Esto requiere una gran capacidad computacional, sin embargo, en ocasiones, existe una expresión cerrada más fácil de tratar.

$$CV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \widehat{f_\lambda}(x_i)}{1 - S_{ii}}\right)^2$$

Donde $\hat{f} = S_\lambda y$, sabiendo que $S = \{S_{ij}\}$ es una matriz de n x n conocida como matriz de suavizado.

Luego, si:

$$\tilde{y}_{ij} = \begin{cases} y_{j,} & j \neq i \\ \hat{f}_{\lambda}^{-i}(x_i), & j = i \end{cases}$$

Por lo que,

$$\widehat{f}_{\lambda}(x_i) = \sum_{j=1}^{n} S_{ij} y_j$$

$$\hat{f}_{\lambda}^{-i}(x_i) = \sum_{j=1}^{n} S_{ij} \tilde{y}_j = \sum_{j=1}^{n} S_{ij} y_j + S_{ii} \hat{f}_{\lambda}^{-i}(x_i)$$

Y combinando estas dos ecuaciones:

$$\widehat{f}_{\lambda}(x_i) - \hat{f}_{\lambda}^{-i}(x_i) = S_{ii} y_i - S_{ii} \hat{f}_{\lambda}^{-i}(x_i) \Rightarrow$$

$$\Rightarrow (S_{ii} - 1) \hat{f}_{\lambda}^{-i}(x_i) = S_{ii} y_i - \sum_{j \neq i} S_{ij} y_j \Rightarrow \hat{f}_{\lambda}^{-i}(x_i) = \frac{\sum_{j=1, j \neq i} S_{ij} y_j}{1 - S_{ii}}$$

Por lo que,

$$y_i - \hat{f}_{\lambda}^{-i}(x_i) = \frac{(1 - S_{ii}) y_i - \sum_{j=1, j \neq i} S_{ij} y_j}{1 - S_{ii}} = \frac{y_i - S_{ii} y_i - \sum_{j=1, j \neq i} S_{ij} y_j}{1 - S_{ii}} = \frac{y_i - \widehat{f}_{\lambda}(x_i)}{1 - S_{ii}}$$

Luego:

$$CV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \widehat{f}_{\lambda}(x_i)}{1 - S_{ii}} \right)^2$$

Con lo que queda demostrado que el ajuste de $\hat{f}_{\lambda}^{-i}(x_i)$ puede ser calculado con $\widehat{f}_{\lambda}(x_i)$ y ya no es necesario remover el i-ésimo dato para recalcular el smooth. El único paso necesario, es ajustar una sola vez el modelo con todos los datos y posteriormente, calcular los elementos de la diagonal de la matriz de suavizado.

Si reemplazamos S_{ii} por los promedios de todos los elementos de la diagonal, obtendremos la validación cruzada generalizada:

$$GCV(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \widehat{f}_{\lambda}(x_i)}{1 - \frac{trS}{n}} \right)^2$$

La GCV es una versión ponderada de CV con pesos $\frac{(1 - S_{ii})^2}{\left(1 - \frac{trS}{n}\right)^2}$

3.2.4 Ajuste de un modelo aditivo generalizado

Como se mencionó anteriormente, un modelo aditivo generalizado tiene la siguiente forma:

$$y = f_0 + \sum_{j=1}^{n} f_j(x_j) + \varepsilon$$

Donde los errores ε son independientes de las x_j, la esperanza del error es cero, $E(\varepsilon) = 0$ y su varianza, $VAR(\varepsilon) = \sigma^2$

El mecanismo más utilizado en la práctica para ajustar un modelo aditivo generalizado es el algoritmo de back-fitting.

Si se define el j-ésimo conjunto de residuos parciales como:

$$R_j = Y - f_0 - \sum_{k \neq j} f_k(x_k)$$

Entonces,

$$E(R_j | x_j) = f_j(x_j)$$

Esta observación, proporcionará una manera de estimar cada función de suavizado.

Dada una observación (x_i, y_i), se puede especificar un criterio como la suma de cuadrados penalizada para resolver este problema:

$$\sum_{j=1}^{n} \left(y_j - f_0 - f_j(x_j)\right)^2 + \sum_{j=1}^{n} \lambda_j \int [f''_j(t_j)]^2 dt_j$$

A continuación, se desarrollará el algoritmo de back-fitting para los modelos aditivos generalizados:

I. Inicio:

$$\hat{f}_0 = \frac{1}{n} \sum_{i=1}^{n} y_i$$
$$\hat{f}_j^1 \equiv 0$$
$$m = 1$$

II. Iteración:

$$m \Leftarrow m + 1, j = 1,2 \dots p$$

$$f_j^m(x_j) \longleftarrow S_j \left[\left(y_i - \hat{f}_0 - \sum_{k=1}^{j-1} \hat{f}_k^m(x_k) - \sum_{k=j+1}^{p} \hat{f}_k^{m-1}(x_k) \right)_1^n \right]$$

III. Hasta que:

$$RSS = S\left(Y - \hat{f}_0 - \sum_{k=1}^{pl} \hat{f}_k^m(x_k) \right)^2$$

Deje de disminuir, lo que significará que la función $\hat{f}_j$ cambie menos que un umbral predefinido.

El procedimiento GAM en el software estadístico SAS, utiliza el siguiente criterio de convergencia para el algoritmo de back-fitting:

$$\frac{\sum_{j=1}^{p} \sum_{i=1}^{n} \left(f_j^{m-1}(x_{ij}) - f_j^m(x_{ij}) \right)^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} f_j^{m-1}(x_{ij})^2} \leq \varepsilon$$

Donde $\varepsilon = 10^{-8}$ por defecto.

Segunda parte
Descripción de la muestra

Como se comentó en la introducción del presente trabajo, el objetivo del mismo es hacer un estudio sobre la influencia de la localización en la frecuencia siniestral en España.

Dicho estudio, se ha realizado para una cartera de seguros del ramo de hogar.

El número de pólizas que se han utilizado para el estudio es de 3.388.929, y no se han tenido en cuenta siniestros con carga negativa.

El primer paso que se ha llevado a cabo, es realizar un análisis exhaustivo de la variable a explicar, en este caso, el número de siniestros. Este paso es muy importante, ya que para modelizar dicha variable, deberemos comprobar si se cumplen una serie de hipótesis. Es fundamental hacer un estudio de la distribución de la variable a explicar, ya que al modelizar con un GLM, debería estar distribuida conforme a una distribución de la familia exponencial, en caso contrario, podrían presentarse problemas, ya que si se utiliza el modelo con fines predictivos y la variable a explicar no se distribuye conforme a una distribución de la familia exponencial, nuestras predicciones podrían presentar anomalías.

En el caso de que nuestra variable dependiente, no se distribuyese conforme a una distribución de la familia exponencial, nos privaremos de tener caracterizada la distribución de la variable dependiente y el modelo lineal generalizado que planteemos será un modelo matemático que explique el comportamiento de la frecuencia en función de los factores de riesgo elegidos.

A continuación, se han seleccionado una serie de variables que, a priori, podrían explicar la frecuencia en el ramo de hogar. Posteriormente, se ha analizado estas variables, trameándolas en el caso de que fuesen continuas, y, finalmente, codificándolas para facilitar su tratamiento.

Antes de plantear un modelo, se ha comprobado el posible grado de asociación/dependencia entre las variables explicativas para que no distorsionen el modelo.

Realizados todos estos pasos, se ha planteado un modelo lineal generalizado y se ha hecho un análisis exhaustivo entre los recargos por frecuencia entre cada factor de riesgo así como un estudio entre la frecuencia real y la frecuencia estimada por el modelo.

Finalmente, asumiendo que los residuos del modelo vienen explicados por la localización del riesgo, se ha planteado un modelo aditivo generalizado con la variable código postal.

1
Análisis del número de siniestros

El número de siniestros es la variable clave de nuestra base de datos. Sin embargo, como es fácil suponer, no es lo mismo que un clúster de 100 expuestos tengan 20 siniestros a que los tengan uno de 30 expuestos. Por ello, utilizaremos la variable exposición como soporte para nuestro modelo lineal generalizado.

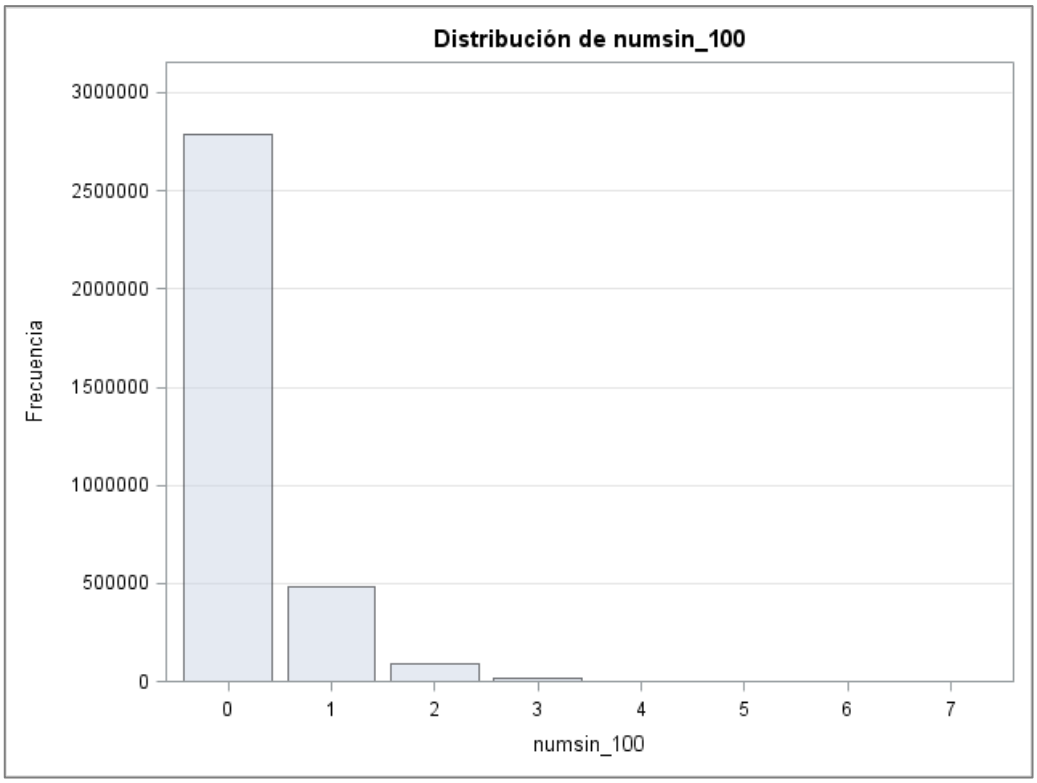
La exposición de la póliza ira en función del número de días en los que ha estado expuesta al riesgo, es decir, si la póliza ha estado los 365 días en vigor, tendrá exposición 1, si ha estado medio año, tendrá 0,5 de exposición y así sucesivamente.

Lógicamente, al querer explicar el número de siniestros deberemos plantear un modelo lineal generalizado para conteos. En estos modelos, la variable a explicar, debe estar distribuida conforme a una distribución de la familia exponencial, más concretamente como una Poisson.

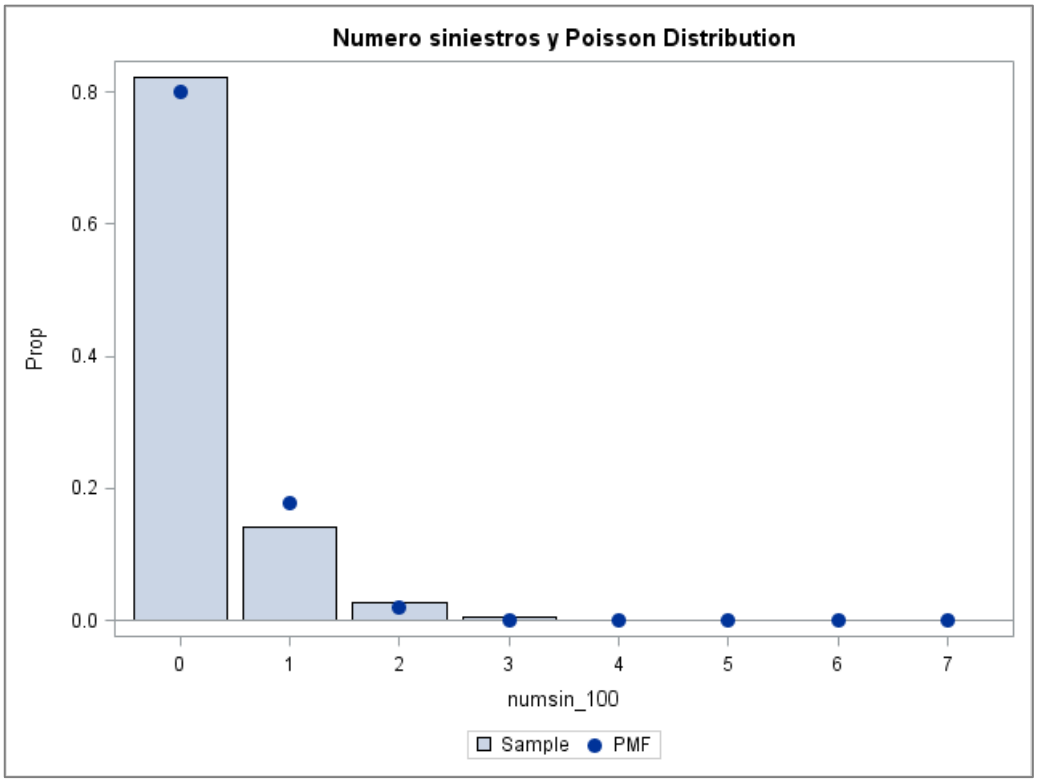
A continuación, se realizará un estudio de la variable del número de siniestros:

Siniestros	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
0	2.787.608	82.26	2.787.608	82.26
1	483.363	14.26	3.270.971	96.52
2	93.220	2.75	3.364.191	99.27
3	18.843	0.56	3.383.034	99.83
4	4.316	0.13	3.387.350	99.95
5	1.119	0.03	3.388.469	99.99
6	334	0.01	3.388.803	100.00
7	126	0.00	3.388.929	100.00

Como se puede comprobar según la distribución de frecuencias, la mayor parte de las pólizas no tienen siniestralidad, algo lógico, ya que el ramo de hogar se caracteriza por presentar bajas tasas de siniestralidad, sin embargo ¿se ajustará esta distribución a una distribución de Poisson?

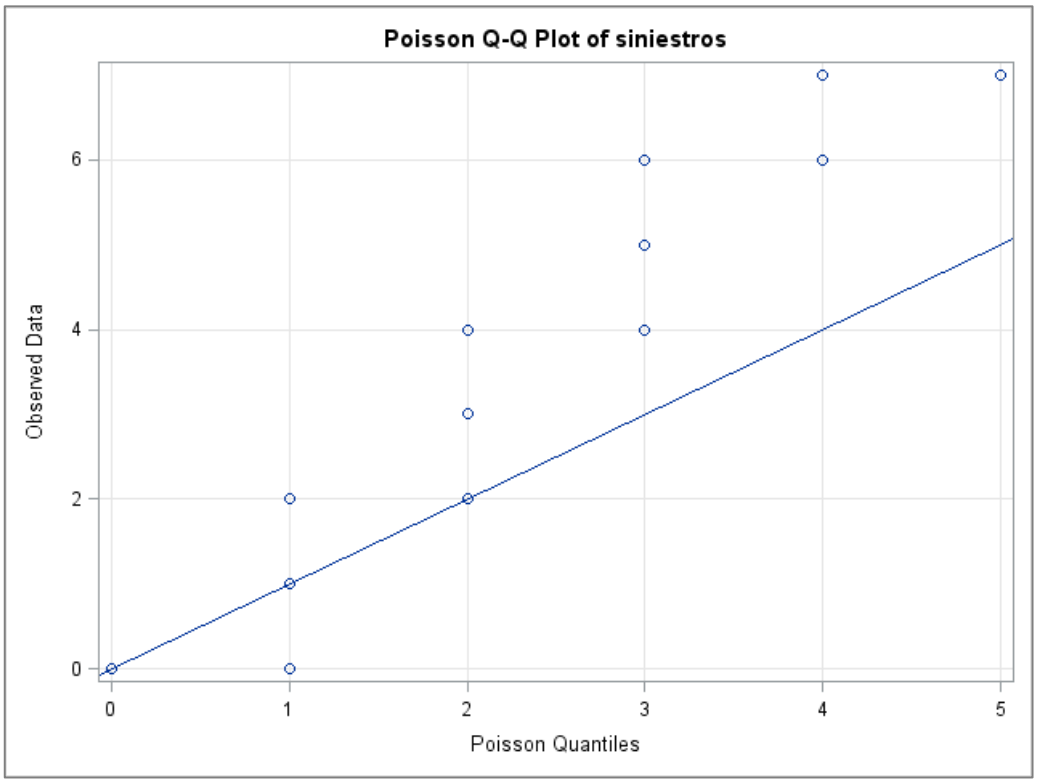


A primera vista, parece tener forma de una distribución Poisson. Una manera de comprobarlo, es representar junto a la distribución empírica de los datos, la distribución teórica de Poisson, calculando para ello el parámetro lambda. En este caso, es fácil ya que dicho parámetro corresponde a la media muestral:



En un principio, no parece ajustar bien, ya que en 0 y 1 siniestros parece haber diferencias. Además parece haber problemas de colas, por lo que podría haber sobredispersión en los datos. Para corroborarlo, se utilizará el gráfico Q-Q.

Este gráfico es un método utilizado para el diagnóstico de diferencias entre la distribución de probabilidad de una población y una distribución elegida como posible distribución de ajuste. Comprobémoslo:



Después de estudiar el gráfico Q-Q, hemos podido comprobar cómo, definitivamente, la distribución de Poisson no se ajusta bien a nuestros datos. Una de las principales causas estaría en la cola de la distribución empírica, ya que ésta es más larga que la cola de la distribución teórica, lo que implicaría sobredispersión en los datos. Por ello, plantearemos otra distribución que suele funcionar muy bien ante la presencia de sobredispersión: la binomial negativa.

El primer paso será calcular la media y varianza muestral, obteniendo:

$$\bar{X} = \sum_{i=1}^{8} x_i p(x_i) = 0{,}2219$$

$$VAR(X) = \sum_{i=1}^{8} p_i (x_i - \bar{X})^2 = 0{,}2875$$

Otro de los motivos por los que suponemos la existencia de sobredispersión, es debido a que la media y la varianza no son iguales, siendo la varianza un 30% superior.

Para comprobar si una binomial negativa se pudiera ajustar a la distribución empírica, realizaremos un test de bondad del ajuste, en este caso un test chi-cuadrado.

Esta prueba no paramétrica compara la distribución de frecuencias observadas (F_0) con la distribución teórica a ajustar (F_E), indicando en qué medida las diferencias entre ambas se deben al azar en el siguiente contraste de hipótesis:

$$H_0: F_0 = F_E$$

$$H_1: F_0 \neq F_E$$

El estadístico utilizado en la prueba será:

$$\chi^2 = \sum_i \frac{(Observada_i - Esperada_i)^2}{Esperada_i}$$

Cuanto mayor sea el valor de χ^2, menos verosímil es que la hipótesis nula sea correcta. De manera análoga, cuanto más tienda a cero el valor de la chi-cuadrado, más se ajustarán ambas distribuciones.

Los grados de libertad vendrán dados por el número de filas menos el número de parámetros a estimar menos uno.

Para realizar la prueba con una distribución binomial negativa, se nos presentan dos dificultades: la programación de la función de cuantía de la distribución binomial negativa y la estimación de los parámetros de la misma.

Se ha utilizado el método de Moment matching para estimar los dos parámetros de la distribución: p y r.

$$\bar{X} = \frac{pr}{1-p}$$

$$VAR(X) = \frac{pr}{(1-p)^2}$$

Utilizando la media y varianza muestral, hemos obtenido unos parámetros de:

$$p = 0{,}2279$$

$$r = 0{,}7515$$

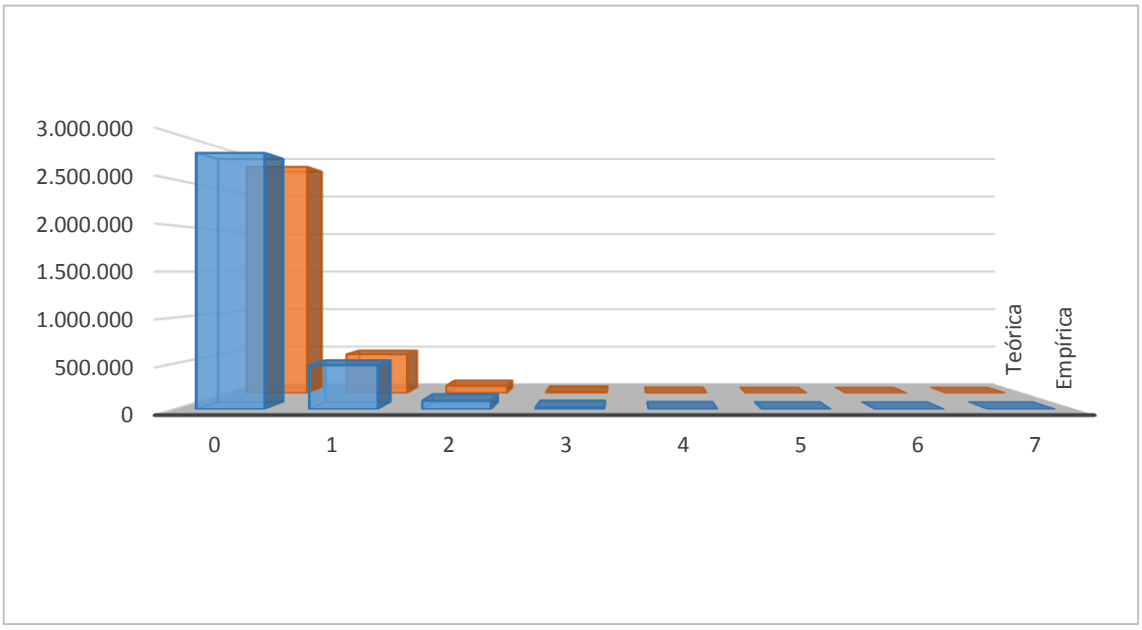
A continuación se ha programado utilizando Visual Basic la función de cuantía de la distribución binomial negativa:

$$pmf = \binom{k+r-1}{k}(1-p)^r p^k$$

```
Function pmf_BN (k, r, p) As Double

    Dim inter As Double
    Inter = WorksheetFunction.GammaLn_Precise (k + r) -
    WorksheetFunction.GammaLn_Precise (k + 1) _
     - WorksheetFunction.GammaLn_Precise(r)
     Inter = inter + r * WorksheetFunction.Ln (1 - p) + k * WorksheetFunction.Ln (p)
    pmf_BN = Exp (inter)

End Function
```

Y, utilizando los parámetros calculados anteriormente, hemos obtenido la siguiente distribución teórica:

Siniestros	Frecuencia Observada (F_0)	Frecuencia Esperada (F_E)
0	2.787.608	2.788.623
1	483.363	482.522
2	93.220	92.937
3	18.843	19.055
4	4.316	4.267
5	1.119	1.094
6	334	317
7	126	114



Gráficamente, ambas distribuciones son prácticamente iguales, al contrario que sucedía cuando utilizábamos la distribución de Poisson. Esto corrobora la idea de que cuando hay sobredispersión en los datos, la binomial negativa funciona bastante bien. Sin embargo, aún no hemos probado si ajusta o no:

Si fijamos un nivel de confianza del 5% y buscamos en tablas el valor de la chi cuadrado para 5 grados de libertad encontramos que el valor del estadístico debe ser inferior a 11,07 para no rechazar la hipótesis nula.

Resultando:

$$\chi^2 = \sum_i \frac{(Observada_i - Esperada_i)^2}{Esperada_i} = 8{,}27$$

Por lo tanto no hay razones estadísticas para rechazar la hipótesis nula y la binomial negativa se ajustaría a la distribución empírica.

Como alternativa, se ha decidido probar con otra distribución como es la beta binomial. Esta distribución surge cuando la probabilidad de éxito de cada ensayo Bernoulli es aleatoria. Dicha probabilidad estaría distribuida conforme a una beta.

La razón de utilizar la beta binomial es que es una distribución que también suele funcionar bien cuando hay dispersión en los datos, ya que el componente aleatorio proporcionado por la beta, induce más difusión que la distribución binomial con p constante.

Al igual que cuando hemos intentado ajustar la binomial negativa, debemos programar la función de cuantía de la distribución beta binomial y calcular los parámetros de la misma.

Para el cálculo de los parámetros, se ha vuelto a hacer uso de la técnica de Moment matching:

$$\bar{X} = \frac{n\alpha}{\alpha + \beta}$$

$$VAR(X) = \frac{n\alpha\beta(\alpha + \beta + n)}{(\alpha + \beta)^2(\alpha + \beta + 1)}$$

Utilizando la media y varianza muestral, hemos obtenido unos parámetros de:

$$\alpha = 0{,}5315$$

$$\beta = 16{,}2357$$

Siendo $n = 7$

A continuación se ha programado en Visual Basic la función de cuantía de la distribución beta binomial, pero antes, debemos programar la función beta:

$$B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a + b)}$$

```
Function betam(a, b) As Double

    Dim inter As Double
    Inter = WorksheetFunction.GammaLn_Precise(a + b) -
    WorksheetFunction.GammaLn_Precise(a) _
     - WorksheetFunction.GammaLn_Precise (b)
    Betam = Exp (-inter)

End Function
```

Una vez programada la función beta, estaremos en disposición de poder calcular la función de cuantía de la distribución beta binomial:

$$pmf = \binom{n}{k} \frac{B(k+\alpha, n-k+\beta)}{B(\alpha,\beta)}$$

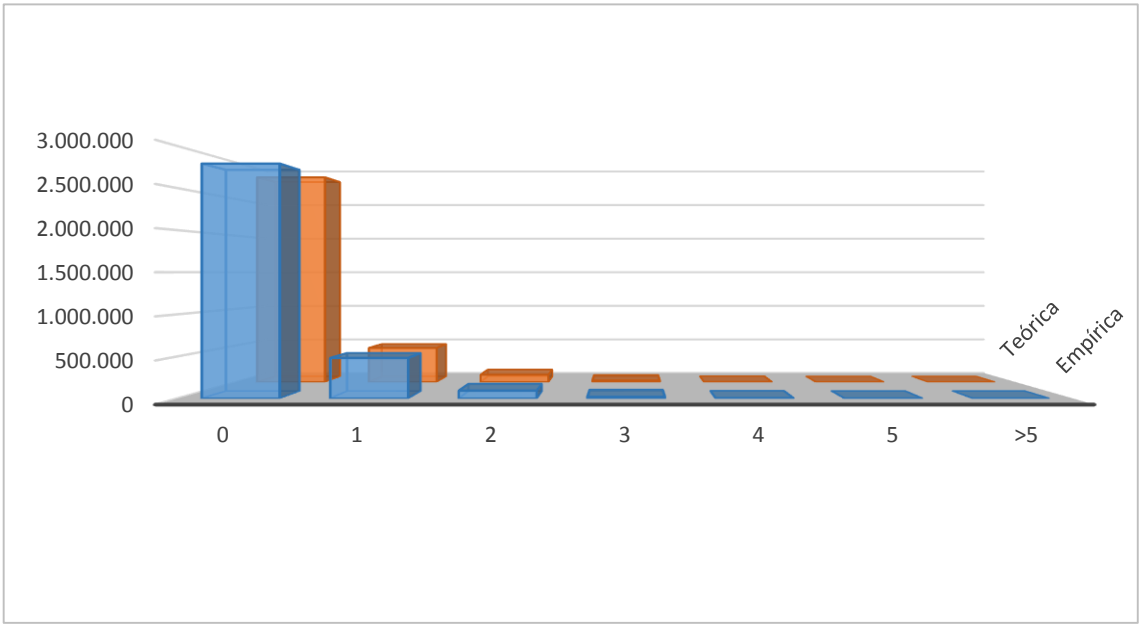
```
Function BB_pmf (k, N, a, b) As Double

    Dim inter As Double
    Inter = WorksheetFunction.GammaLn_Precise(N + 1) -
    WorksheetFunction.GammaLn_Precise (k + 1) _
     - WorksheetFunction.GammaLn_Precise(N - k + 1)
     Inter = Exp (inter)
    BB_pmf = inter * betam (k + a, N - k + b) / betam(a, b)

End Function
```

Y, utilizando los parámetros calculados anteriormente, hemos obtenido la siguiente distribución teórica:

Siniestros	Frecuencia Observada (F_0)	Frecuencia Esperada (F_E)
0	2.787.608	2.789.260
1	483.363	483.643
2	93.220	92.183
3	18.843	18.197
4	4.316	4.173
5	1.119	1.068
6	334	402
7	126	4

Observando las distribuciones anteriores, observamos que hay frecuencias esperadas inferiores a 5. Uno de los requisitos sobre los que se asienta el test chi-cuadrado de la bondad del ajuste es que no haya frecuencias inferiores a 5, por lo que procedemos a agrupar de la siguiente manera:

Siniestros	Frecuencia Observada (F_0)	Frecuencia Esperada (F_E)
0	2.787.608	2.789.260
1	483.363	483.643
2	93.220	92.183
3	18.843	18.197
4	4.316	4.173
5	1.119	1.068
>5	460	406



Gráficamente, parece que la beta binomial ajusta peor que la binomial negativa. Para cerciorarnos, realizaremos el test de bondad del ajuste.

Fijando el mismo nivel de confianza del 5% y para cuatro grados de libertad (ya que al tener frecuencias inferiores a cinco nos hemos visto obligados a tramear las mismas, por lo que tenemos una fila menos) obtenemos de tablas el valor 9,48.

Calculando el estadístico, resulta:

$$\chi^2 = \sum_i \frac{(Observada_i - Esperada_i)^2}{Esperada_i} = 50{,}46$$

Por lo que, como se esperaba observando el gráfico de frecuencias, la distribución beta binomial no ajustaría correctamente y, por tanto se rechazaría la hipótesis nula.

Este estudio que se ha realizado de la variable número de siniestros es fundamental. Muchas compañías de seguros se suelen saltar este análisis, modelizando directamente la variable a estudio. Esto podría acarrear algún problema, ya que si se utiliza el modelo con fines predictivos y la variable a explicar no se distribuye conforme a una distribución de la familia exponencial, nuestras predicciones podrían presentar anomalías.

En el caso que ocupa, parece que la distribución binomial negativa ajusta bien.

En un principio, la distribución binomial negativa no pertenece a la familia exponencial, sin embargo, esta distribución no es más que una Poisson sobredispersa en la que dicha sobredispersión viene dada por el parámetro Lambda, que no es fijo para cada asegurado, sino que se distribuye conforme a una gamma que además es máximo entrópica, es decir, aporta máxima desinformación.

Por lo tanto, podremos modelizar dicha variable y, si el modelo es adecuado, utilizarlo con fines predictivos.

2
Análisis de los factores de riesgo

En este apartado, se realizará un análisis descriptivo de las variables explicativas utilizadas en nuestro modelo lineal generalizado.

Estas variables, han sido elegidas por su potencial influencia sobre el número de siniestros de una póliza del ramo hogar:

- Forma de pago
- Superficie inmueble
- Año estudio
- Capital del edificio
- Capital mobiliario
- Tipo de vivienda
- Uso de la vivienda
- Ubicación de la vivienda
- Antigüedad de la póliza

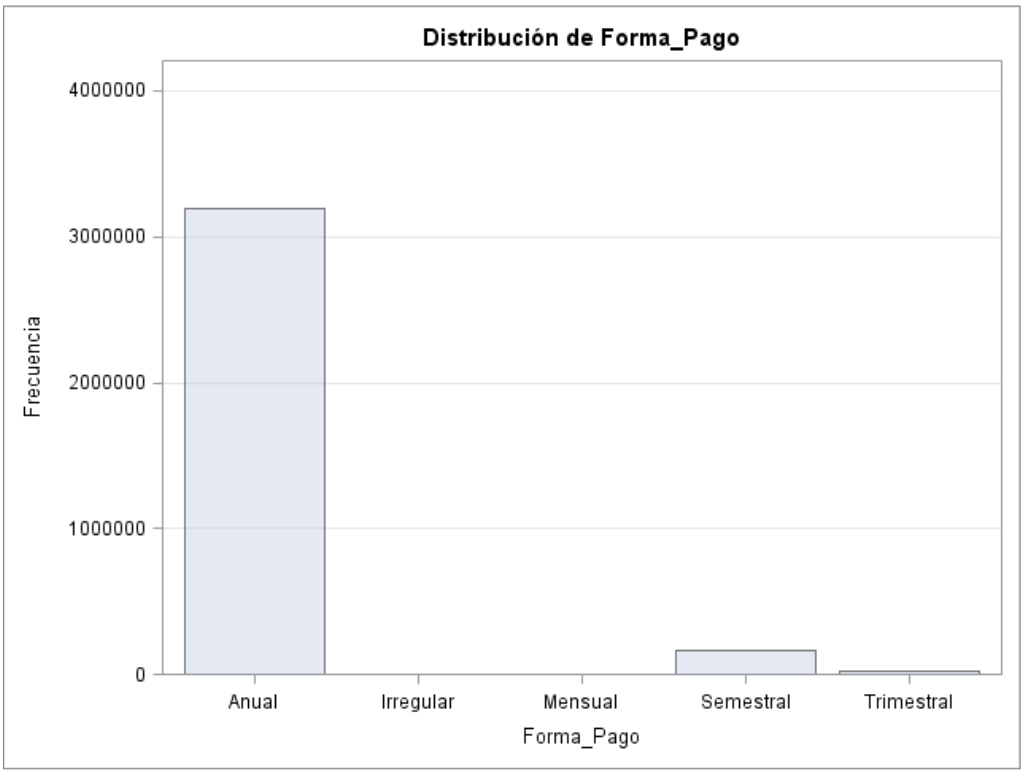
De aquellas variables en las que no había masa suficiente, se ha procedido a realizar agrupaciones con el fin de que los clúster finales de nuestro GLM sean lo suficientemente nutridos para que no se produzcan errores predictivos.

Finalmente, una vez estudiadas dichas variables, se ha realizado un estudio sobre el posible grado de asociación de las mismas, para evitar problemas en el modelo lineal generalizado propuesto.

2.1 Forma de pago

Esta variable hace referencia a la forma en la que el asegurado paga su prima correspondiente. Nos encontramos con la siguiente casuística:

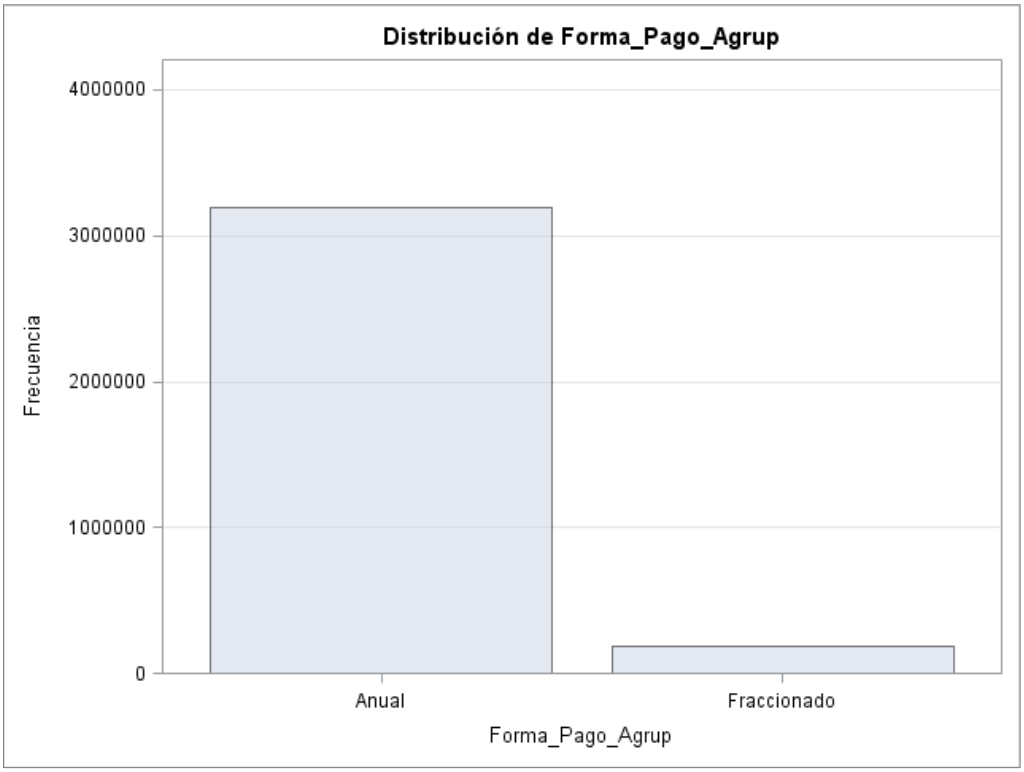
Forma_Pago	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
Anual	3.197.449	94.35	3.197.449	94.35
Irregular	113	0.00	3.197.562	94.35
Mensual	422	0.01	3.197.984	94.37
Semestral	165.521	4.88	3.363.505	99.25
Trimestral	25.424	0.75	3.388.929	100.00



Podemos observar como la gran mayoría de asegurados escoge la modalidad de pago anual, por lo que parece sensato separar la forma de pago en dos modalidades: anual y fraccionado.

Por lo tanto, tendremos:

Forma_Pago _Agrup	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
Anual	3.197.449	94.35	3.197.449	94.35
Fraccionado	191.480	5.65	3.388.929	100.00



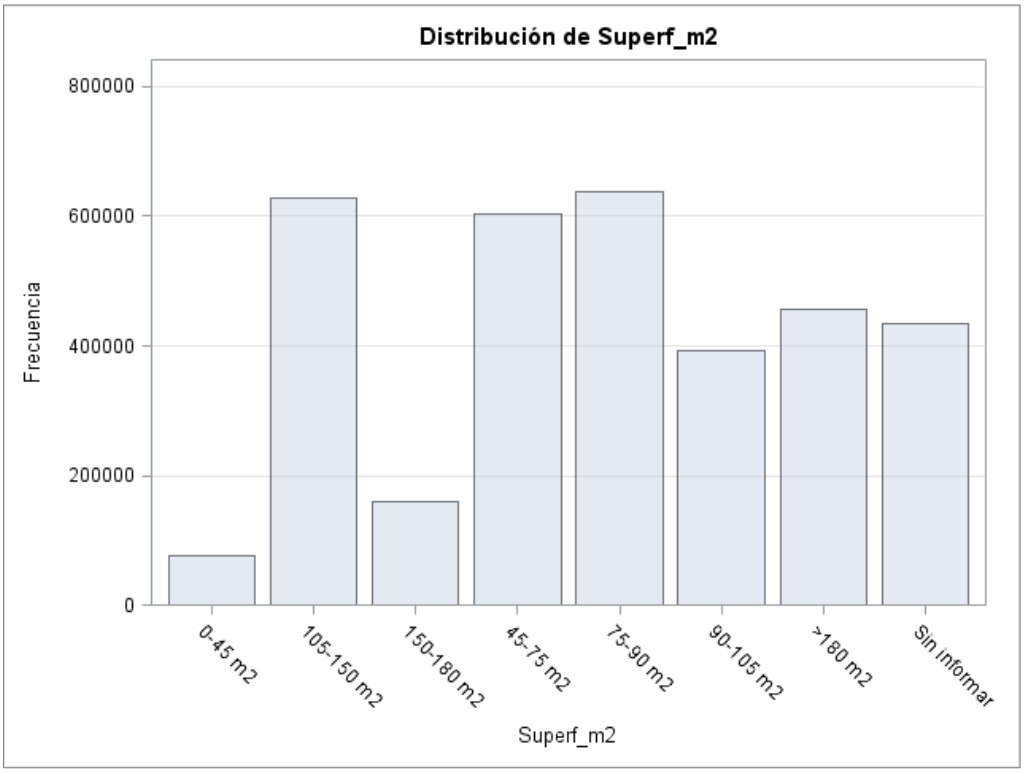
En la modelización del número de siniestros, esta variable se codificará como una variable de clasificación de la siguiente manera:

- `Anual`= **`99`**
- `Fraccionado`= **`2`**

2.2 Superficie

La variable superficie mide el número de metros cuadrados útiles del inmueble. En nuestra base de datos encontramos:

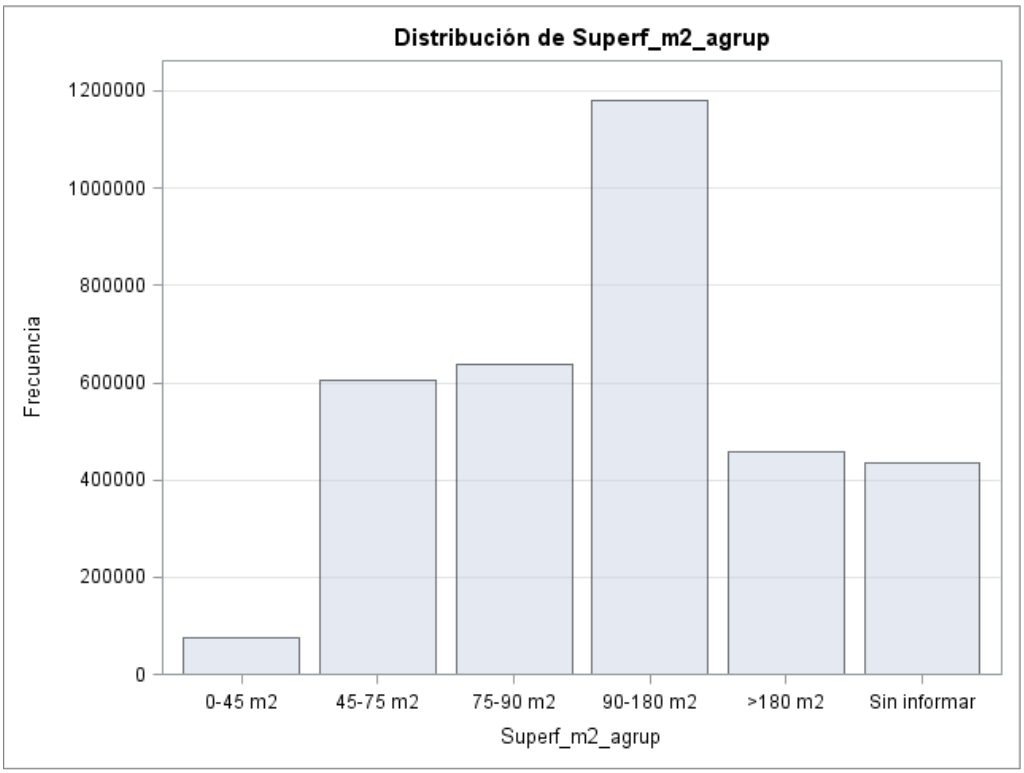
Superf_m2	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
0-45 m2	75.301	2.22	75.301	2.22
105-150 m2	627.688	18.52	702.989	20.74
150-180 m2	160.261	4.73	863.250	25.47
45-75 m2	604.034	17.82	1.467.284	43.30
75-90 m2	636.793	18.79	2.104.077	62.09
90-105 m2	392.482	11.58	2.496.559	73.67
>180 m2	457.385	13.50	2.953.944	87.16
Sin informar	434.985	12.84	3.388.929	100.00



La superficie más repetida de nuestra base de datos es la de 75-90 metros cuadrados. Algo esperado siendo ésta la superficie más habitual de los inmuebles en España.

Hemos agrupado aquellas superficies comprendidas entre 90 y 180 metros cuadrados de la siguiente manera:

Superf_m2_agrup	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
0-45 m2	75.301	2.22	75.301	2.22
45-75 m2	604.034	17.82	679.335	20.05
75-90 m2	636.793	18.79	1.316.128	38.84
90-180 m2	1.180.431	34.83	2.496.559	73.67
>180 m2	457.385	13.50	2.953.944	87.16
Sin informar	434.985	12.84	3.388.929	100.00



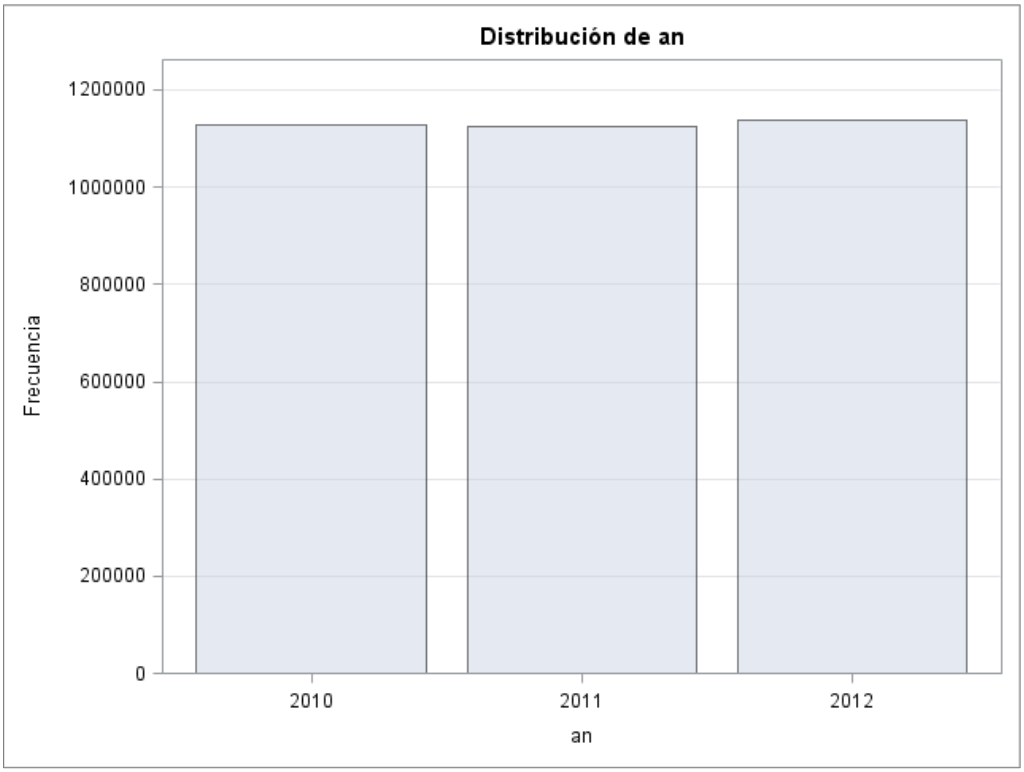
Finalmente, la codificación de la variable agrupada para su tratamiento como variable de clasificación en la modelización ha sido:

- `Sin informar=1`
- `0-45 m2=2`
- `45-75 m2=3`
- `90-180 m2=5`
- `>180 m2=8`
- `75-90 m2=99`

2.3 Año de estudio

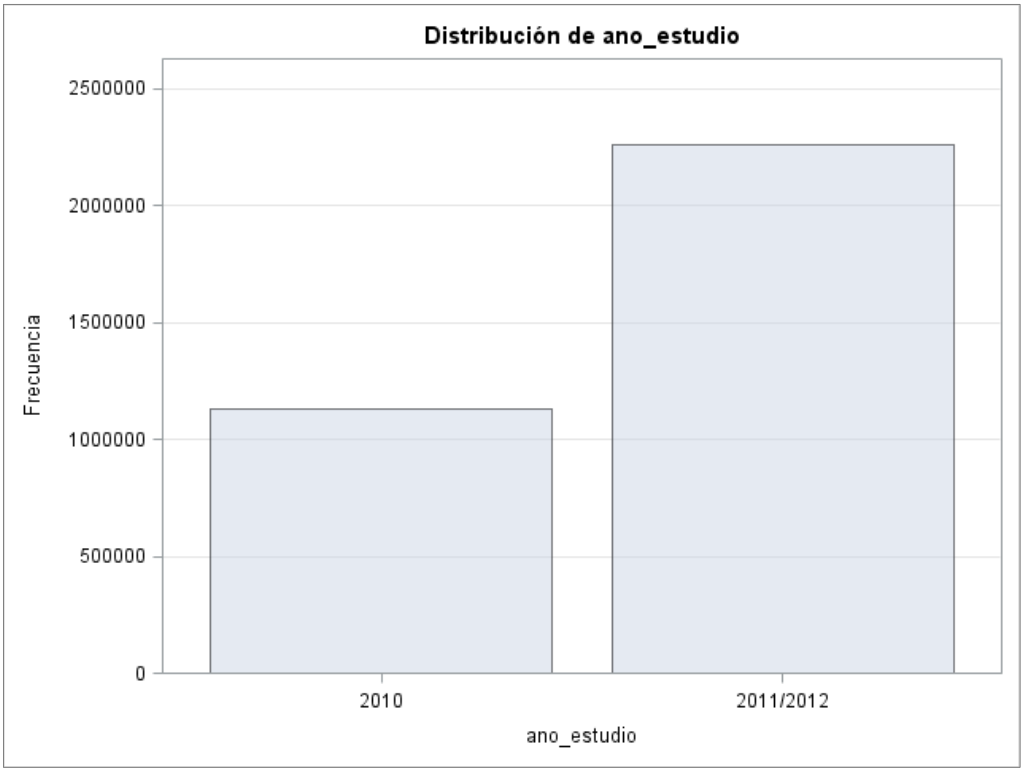
Para nuestro análisis de la frecuencia, se han tenido en cuenta tres carteras del ramo de hogar. Las carteras son las de los años 2010, 2011 y 2012:

Año	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
2010	1.128.023	33.29	1.128.023	33.29
2011	1.124.428	33.18	2.252.451	66.46
2012	1.136.478	33.54	3.388.929	100.00



Vemos que la distribución de las pólizas es bastante homogénea entre los tres años de estudio. Sin embargo, en 2011, se realizó un cambio en las garantías ofrecidas por la compañía, por lo que por criterios de homogeneidad, se ha procedido a agrupar las carteras de 2011 y 2012 de la siguiente forma:

Año_agrup	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
2010	1.128.023	33.29	1.128.023	33.29
2011/2012	2.260.906	66.71	3.388.929	100.00



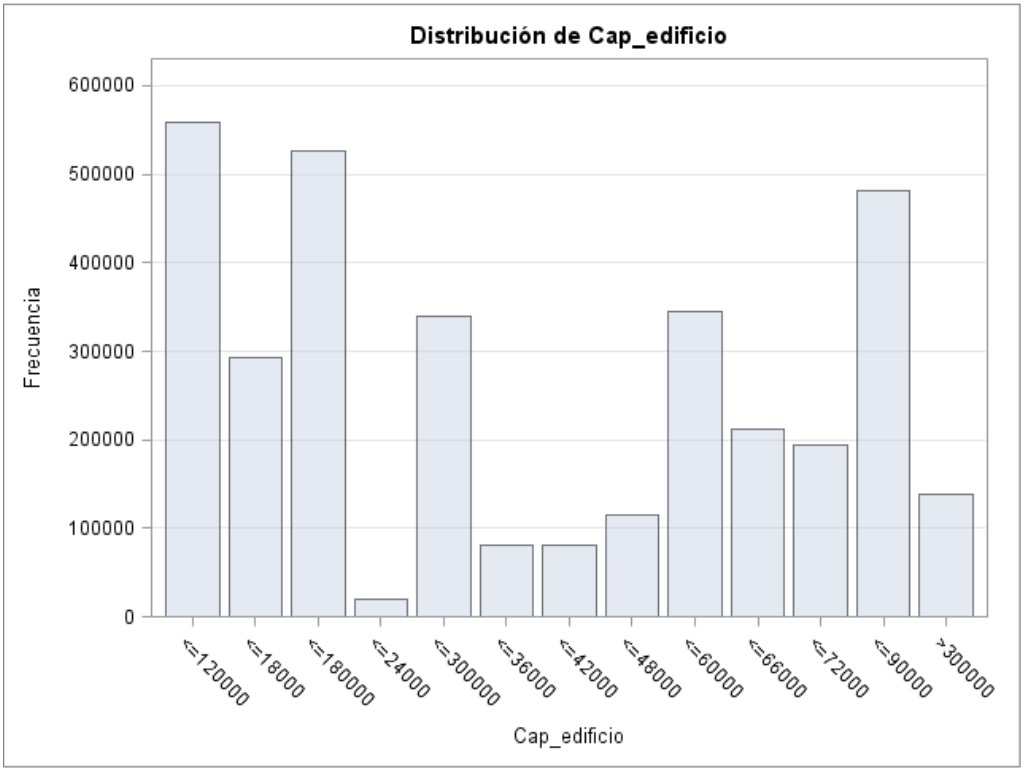
Lógicamente, al agrupar de esta manera, se va a mantener un criterio de homogeneidad en las garantías ofrecidas por la compañía según el valor que tome nuestra variable de clasificación:

- `2010=1`
- `2011/2012=2`

2.4 Capital del edificio

Se trata del capital en el que está asegurado el inmueble. Se trata de una variable continua en la que se han establecido los siguientes intervalos:

Cap_edificio	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
<=120.000	558.597	16.48	558.597	16.48
<=18.000	293.367	8.66	851.964	25.14
<=180.000	525.446	15.50	1.377.410	40.64
<=24.000	20.377	0.60	1.397.787	41.25
<=300.000	340.139	10.04	1.737.926	51.28
<=36.000	81.958	2.42	1.819.884	53.70
<=42.000	81.652	2.41	1.901.536	56.11
<=48.000	114.762	3.39	2.016.298	59.50
<=60.000	344.806	10.17	2.361.104	69.67
<=66.000	212.701	6.28	2.573.805	75.95
<=72.000	194.439	5.74	2.768.244	81.68
<=90.000	481.349	14.20	3.249.593	95.89
>300.000	139.336	4.11	3.388.929	100.00



Para su codificación en variable de clasificación se ha procedido de la siguiente manera:

- <=18.000=1
- <=24.000=2
- <=36.000=3
- <=42.000=4
- <=48.000=5
- <=60.000=6
- <=66.000=7
- <=72.000=8
- <=90.000=9
- <=120.000=10
- <=180.000=11
- <=300.000=12
- >300.000=13

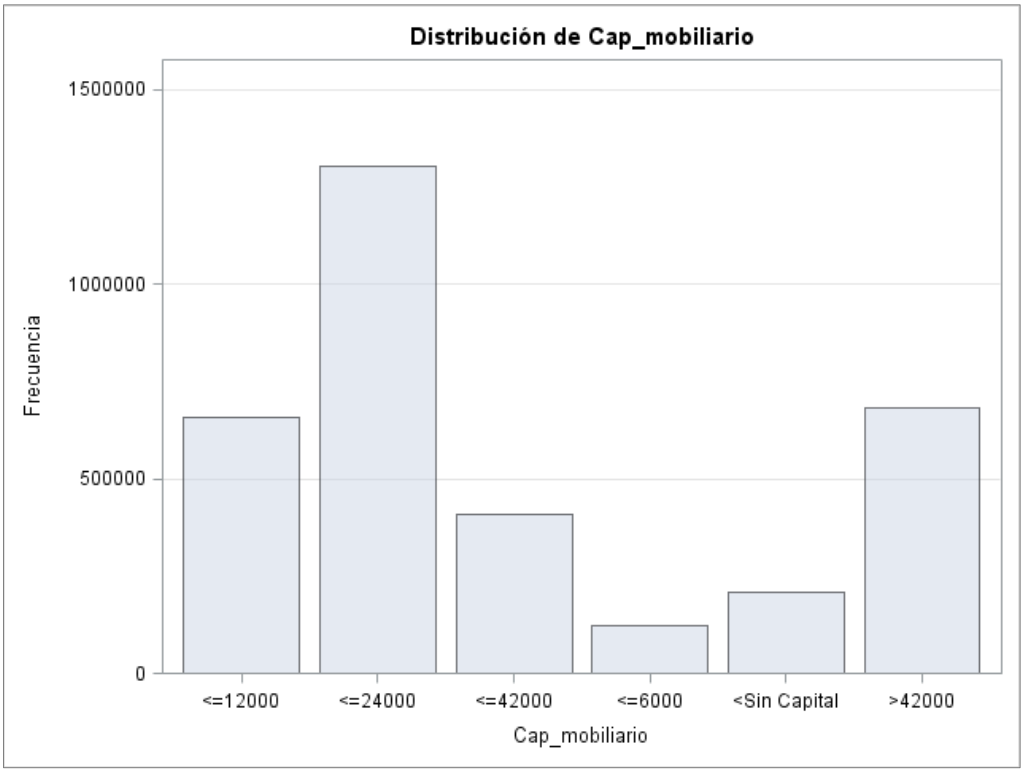
2.5 Capital mobiliario

Esta variable hace referencia a la suma asegurada del contenido, es decir, la cantidad en la que están asegurados los bienes mobiliarios del edificio. Al igual que la variable capital del edificio, esta variable es continua y se ha procedido a realizar intervalos de la siguiente manera:

Cap_mobiliario	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
<=12.000	657.930	19.41	657.930	19.41
<=24.000	1.305.008	38.51	1.962.938	57.92
<=42.000	410.197	12.10	2.373.135	70.03
<=6.000	123.215	3.64	2.496.350	73.66
Sin capital	208.321	6.15	2.704.671	79.81
>42.000	684.258	20.19	3.388.929	100.00

Para su posterior clasificación, se ha codificado la variable capital mobiliario:

- Sin capital=0
- <=6.000=1
- <=12.000=2
- <=24.000=4
- <=42.000=7
- >42.000=8



2.6 Tipo de vivienda

El tipo de vivienda es una variable que especifica el tipo de inmueble. Hemos encontrado en nuestra base de datos la siguiente casuística:

Tipo_Vivienda	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
Casa tradicional	226.120	6.67	226.120	6.67
Chalet adosado	388.730	11.47	614.850	18.14
Chalet independiente	388.847	11.47	1.003.697	29.62
Otras viviendas	214.900	6.34	1.218.597	35.96
Piso	2.170.332	64.04	3.388.929	100.00

Como era de esperar, el mayor número de pólizas se encuentra en los pisos, ya que es el tipo de vivienda más habitual.

No utilizaremos esta variable por sí sola para modelizar la frecuencia, sino que crearemos una variable conjunta con el tipo de vivienda, su uso y su ubicación.

2.7 Uso de la vivienda

El uso de vivienda hace referencia a la manera en la que el asegurado utiliza dicha vivienda:

Uso_Vivienda	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
Alquilada	278.371	8.21	278.371	8.21
Desocupada	54.241	1.60	332.612	9.81
Habitual	2.429.693	71.70	2.762.305	81.51
Secundaria	626.624	18.49	3.388.929	100.00

El uso habitual es el más frecuente con un 71% de peso, seguido del uso secundario con un 18%.

2.8 Ubicación de la vivienda

Dicha variable describe el lugar donde está situada la vivienda del asegurado:

Ubicacion_Vivienda	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
Casco urbano	3.007.583	88.75	3.007.583	88.75
Despoblado	149.148	4.40	3.156.731	93.15
Urbanización	232.198	6.85	3.388.929	100.00

Como podemos observar, la mayor parte de las viviendas están situadas en caso urbano.

A continuación, tal y como se mencionó anteriormente, se han unificado las tres últimas variables para formar una variable conocida como Tipo-Uso-Ubicación de la vivienda.

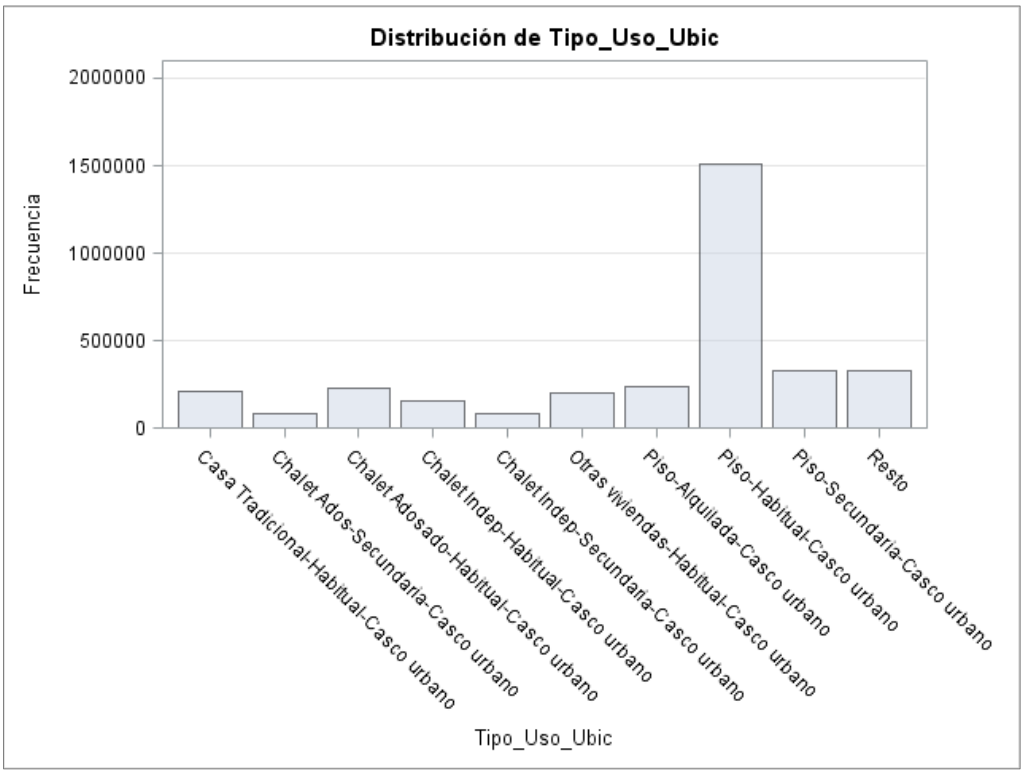
Esta variable nos será de gran utilidad ya que agrupa la potencia explicativa de las variables anteriores en una sola.

A priori, habiendo analizado por separado las tres variables que formarán el tipo-uso-ubicación de la vivienda, se espera que el perfil de Piso-Habitual-Caso urbano sea el más frecuente.

2.9 Tipo-Uso-Ubicación de la vivienda

Como se explicó con anterioridad, esta variable es una conjunción del tipo de la vivienda, su uso y su ubicación:

Tipo_Uso_Ubic	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
Casa Tradicional-Habitual-Casco urbano	214.498	6.33	214.498	6.33
Chalet Ados-Secundaria-Casco urbano	88.223	2.60	302.721	8.93
Chalet Adosado-Habitual-Casco urbano	230.634	6.81	533.355	15.74
Chalet Indep-Habitual-Casco urbano	159.163	4.70	692.518	20.43
Chalet Indep-Secundaria-Casco urbano	87.555	2.58	780.073	23.02
Otras viviendas-Habitual-Casco urbano	207.478	6.12	987.551	29.14
Piso-Alquilada-Casco urbano	238.931	7.05	1.226.482	36.19
Piso-Habitual-Casco urbano	1.505.478	44.42	2.731.960	80.61
Piso-Secundaria-Casco urbano	330.962	9.77	3.062.922	90.38
Resto	326.007	9.62	3.388.929	100.00



Tal y como se esperaba, el piso habitual situado en casco urbano es con diferencia, la casuística más frecuente.

Para su codificación, se ha procedido de la siguiente manera:

- Resto=0
- Piso-Habitual-Casco urbano=1
- Piso-Secundaria-Casco urbano=2
- Chalet Adosado-Habitual-Casco urbano=3
- Piso-Alquilada-Casco urbano=4
- Casa Tradicional-Habitual-Casco urbano=5
- Chalet Indep-Habitual-Casco urbano=6
- Chalet Ados-Secundaria-Casco urbano=8
- Chalet Indep-Secundaria-Casco urbano=9
- Otras viviendas-Habitual-Casco urbano=10

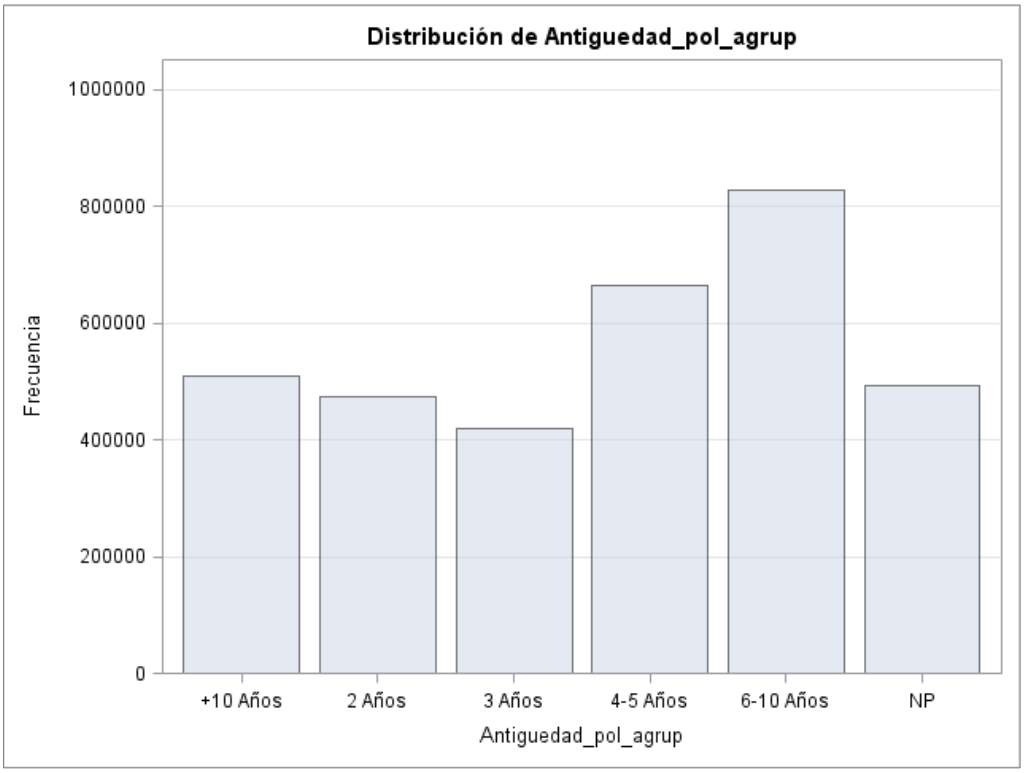
2.10 Antigüedad de la póliza

El último factor de riesgo que utilizaremos para modelizar la frecuencia siniestral es la antigüedad de la póliza del asegurado. Esta puede ser:

Antiguedad_pol	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
10 Años	106.116	3.13	106.116	3.13
11 Años	77.189	2.28	183.305	5.41
12 Años	432.528	12.76	615.833	18.17
13 Años	3	0.00	615.836	18.17
2 Años	475.181	14.02	1.091.017	32.19
3 Años	418.725	12.36	1.509.742	44.55
4 Años	357.398	10.55	1.867.140	55.10
5 Años	307.871	9.08	2.175.011	64.18
6 Años	251.325	7.42	2.426.336	71.60
7 Años	197.345	5.82	2.623.681	77.42
8 Años	149.905	4.42	2.773.586	81.84
9 Años	122.577	3.62	2.896.163	85.46
NP	492.766	14.54	3.388.929	100.00

Vamos a agrupar por intervalos esta variable de la siguiente manera:

Antiguedad_pol_agrup	Frecuencia	Porcentaje	Frecuencia acumulada	Porcentaje acumulado
+10 Años	509.720	15.04	509.720	15.04
2 Años	475.181	14.02	984.901	29.06
3 Años	418.725	12.36	1.403.626	41.42
4-5 Años	665.269	19.63	2.068.895	61.05
6-10 Años	827.268	24.41	2.896.163	85.46
NP	492.766	14.54	3.388.929	100.00



Las pólizas de 6-10 años de antigüedad son las más frecuentes en la compañía.

Al igual que el resto de variables, se ha codificado la antigüedad para poder trabajarla como una variable de clasificación:

- NP=1
- 2 Años=2
- 3 Años=3
- 4-5 Años=4
- 6-10 Años=6
- > 10 Años=10

2.11 Estudio de la asociación entre las variables del modelo

Antes de plantear el modelo lineal generalizado, se procederá a estudiar el posible grado de asociación entre las variables explicativas del mismo.

Como nuestras variables explicativas están codificadas como variables de clasificación, utilizaremos el Cramérs V para analizar el posible grado de asociación entre las mismas.

Esta medida explica la intensidad en la asociación entre dos o más variables cuantitativas o cualitativas. El Cramérs V está basado en la χ^2

El Cramérs V es un valor de medida independiente del tamaño de la muestra. Cramérs V es una medida simétrica para la intensidad de la relación entre dos o más variables de escala nominal, cuando (por lo menos) una de las dos variables tiene por lo menos dos formas (valores posibles).

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

Siendo:

- n: el volumen de la muestra
- $\min[r,c]$: el menos entre el número de filas y el número de columnas

El Cramérs V, independientemente del número de filas y columnas, toma valores entre 0 y 1. Puede utilizarse para tablas de contingencia de cualquier tamaño. La interpretación de dicha medida sería:

- Cramérs V = 0; no hay relación entre las dos variables
- Cramérs V = 1; hay una relación perfecta entre las dos variables
- Cramérs V = 0,6; hay una relación relativamente fuerte entre las dos variables

V_Cramér	Forma de pago	Superficie	Año estudio	Capital edificio	Capital mobiliario	Uso-Ubicación-Tipo	Antigüedad póliza
Forma de pago		0.13	0.00	0.15	0.11	0.15	0.05
Superficie	0.13		0.04	0.41	0.20	0.34	0.26
Año estudio	0.00	0.04		0.04	0.02	0.05	0.04
Capital edificio	0.15	0.41	0.04		0.20	0.20	0.06
Capital mobiliario	0.11	0.20	0.02	0.20		0.21	0.07
Uso-Ubicación-Tipo	0.15	0.34	0.05	0.20	0.21		0.14
Antigüedad póliza	0.05	0.26	0.04	0.06	0.07	0.14	

Como puede apreciarse, no existe relación relativamente fuerte entre ninguna variable.

La más alta correspondería a la superficie con el capital asegurado del edificio con un grado de asociación del 0,41, sin embargo, esta relación no es lo suficientemente grande, por lo que un primer paso, incluiremos ambas variables en el modelo lineal generalizado y estudiaremos posteriormente si debemos desparametrizar el modelo.

Tercera parte
Metodología

En la siguiente parte del presente trabajo se detallará la metodología seguida durante la construcción del modelo lineal generalizado, así como su posterior evaluación.

De forma análoga, se realizará un estudio de impacto comparativo entre las frecuencias de la siniestralidad de cada clúster con las frecuencias devueltas por el modelo.

A continuación, se analizarán los posibles recargos que se producirán a nivel de instancias de cada variable, algo fundamental, ya que gracias a este estudio, se podrá saber cuánto se deberá recargar en función de determinados clúster.

Finalmente, se construirá un modelo aditivo generalizado, para explicar los residuos del modelo lineal generalizado en función de la localización del riesgo suscrito.

Según este modelo, asumiremos que todos los errores del modelo lineal generalizado, vendrán determinados por la variable localización del riesgo (código postal).

Por lo tanto, plantearemos un modelo GAM en el que la variable dependiente será los residuos del modelo lineal generalizado y la única variable explicativa, el código postal.

De forma resumida, el presente capitulo comprenderá:

- Construcción y evaluación del modelo lineal generalizado
- Estudio comparativo entre la frecuencia real y la estimada por el modelo
- Impacto en negocio
- Construcción y evaluación del modelo aditivo generalizado

En la siguiente parte del estudio, se utilizará el software de Microsoft MapPoint para detallar el impacto de nuestro estudio.

Para la construcción del modelo lineal generalizado y del modelo aditivo generalizado se ha utilizado el paquete estadístico SAS Base 9.3.

1
Construcción y evaluación del GLM

El número de siniestros es la variable clave. Por lo tanto, trabajaremos con datos de conteo. El modelo básico para datos de conteo es aquel modelo en el que la variable respuesta se distribuye de acuerdo a una distribución de Poisson y cuya función link es la función logaritmo.

Descripción	Valor
Data Set	TFM.CLASES_RIESGO
Distribution	Poisson
Link Function	Log
Dependent Variable	numsin_100
Offset Variable	lnduree

Como se muestra en la salida SAS generada al construir el modelo, la distribución utilizada ha sido una distribución de Poisson y la función de enlace la logarítmica.

El sistema estadístico SAS incluye una opción al realizar el procedimiento GENMOD (aquel utilizado para realizar modelos GLM) conocida como Offset. Esta opción se utiliza para que si asignamos una variable offset, los betas devueltos por el modelo, vengan escalados en un cierto valor, es decir, en nuestro caso, como se comentó anteriormente, no es lo mismo que un clúster de 2.000 expuestos tenga 20 siniestros, a que tenga el mismo número de siniestros un clúster de 10 expuestos. Por lo tanto, utilizaremos como variable offset la variable exposición, sin embargo, previamente deberemos calcular su logaritmo para que el modelo tenga en cuenta la transformación.

Las posibles instancias de cada variable utilizada en el modelo son:

Class level information		
Class	**Levels**	**Values**
Forma_Pago	2	2 99
Superficie	6	1 2 3 5 8 99
Año_estudio	2	1 2
Capital_edificio	13	1 2 3 4 5 6 7 8 9 10 11 12 13
Capital_mobiliario	6	0 1 2 4 7 8
Tipo_uso_ubicacion	10	0 1 2 3 4 5 6 8 9 10
Antiguedad_pol	6	1 2 3 4 6 10

Los parámetros estimados por el modelo son:

Parameter	Level	DF	Estimate	Standard Error	95% Lower Confidence Limit	95% Upper Confidence Limit	Wald Chi-Square	Pr > ChiSq	tval	rel	pval
Intercept		1	-1,42	0,01	-1,44	-1,39	12.626,79	0,00	-112,37	0,24	0,00
Forma_pago	2	1	0,32	0,00	0,31	0,33	4.215,03	0,00	64,92	1,38	0,00
Forma_pago	99	0	0,00	0,00	0,00	0,00				1,00	
Superficie	1	1	-0,14	0,01	-0,16	-0,13	407,84	0,00	-20,20	0,87	0,00
Superficie	2	1	-0,19	0,01	-0,22	-0,16	208,80	0,00	-14,45	0,83	0,00
Superficie	3	1	-0,10	0,01	-0,11	-0,09	402,65	0,00	-20,07	0,90	0,00
Superficie	5	1	0,07	0,00	0,06	0,08	266,54	0,00	16,33	1,07	0,00
Superficie	8	1	0,18	0,01	0,17	0,19	791,59	0,00	28,14	1,19	0,00
Superficie	99	0	0,00	0,00	0,00	0,00				1,00	
Año estudio	1	1	0,05	0,00	0,04	0,05	274,16	0,00	16,56	1,05	0,00
Año estudio	2	0	0,00	0,00	0,00	0,00				1,00	
Capital_ed	1	1	-0,59	0,01	-0,61	-0,57	3.937,57	0,00	-62,75	0,55	0,00
Capital_ed	2	1	-0,26	0,02	-0,30	-0,22	134,23	0,00	-11,59	0,77	0,00
Capital_ed	3	1	-0,17	0,01	-0,20	-0,14	161,07	0,00	-12,69	0,84	0,00
Capital_ed	4	1	-0,14	0,01	-0,16	-0,11	113,78	0,00	-10,67	0,87	0,00
Capital_ed	5	1	-0,13	0,01	-0,16	-0,11	131,78	0,00	-11,48	0,88	0,00
Capital_ed	6	1	-0,12	0,01	-0,14	-0,11	181,48	0,00	-13,47	0,88	0,00
Capital_ed	7	1	-0,13	0,01	-0,15	-0,11	178,44	0,00	-13,36	0,88	0,00
Capital_ed	8	1	-0,15	0,01	-0,17	-0,13	247,51	0,00	-15,73	0,86	0,00
Capital_ed	9	1	-0,15	0,01	-0,17	-0,14	337,16	0,00	-18,36	0,86	0,00
Capital_ed	10	1	-0,15	0,01	-0,16	-0,13	342,46	0,00	-18,51	0,86	0,00
Capital_ed	11	1	-0,14	0,01	-0,15	-0,12	342,47	0,00	-18,51	0,87	0,00
Capital_ed	12	1	-0,11	0,01	-0,12	-0,10	248,63	0,00	-15,77	0,90	0,00
Capital_ed	13	0	0,00	0,00	0,00	0,00				1,00	
Capital_mo	0	1	-0,66	0,01	-0,68	-0,65	6.318,28	0,00	-79,49	0,52	0,00
Capital_mo	1	1	-0,34	0,01	-0,36	-0,32	1.219,46	0,00	-34,92	0,71	0,00
Capital_mo	2	1	-0,27	0,01	-0,28	-0,26	2.764,40	0,00	-52,58	0,76	0,00
Capital_mo	4	1	-0,11	0,00	-0,12	-0,10	805,83	0,00	-28,39	0,90	0,00
Capital_mo	7	1	-0,06	0,00	-0,07	-0,05	181,60	0,00	-13,48	0,94	0,00
Capital_mo	8	0	0,00	0,00	0,00	0,00				1,00	
Tipo_uso_ub	0	1	-0,09	0,01	-0,11	-0,08	106,83	0,00	-10,34	0,91	0,00
Tipo_uso_ub	1	1	0,18	0,01	0,16	0,20	528,07	0,00	22,98	1,20	0,00
Tipo_uso_ub	2	1	-0,31	0,01	-0,32	-0,29	1.019,32	0,00	-31,93	0,74	0,00
Tipo_uso_ub	3	1	0,28	0,01	0,26	0,30	980,84	0,00	31,32	1,32	0,00
Tipo_uso_ub	4	1	0,14	0,01	0,12	0,16	214,15	0,00	14,63	1,15	0,00
Tipo_uso_ub	5	1	0,19	0,01	0,17	0,20	402,25	0,00	20,06	1,20	0,00
Tipo_uso_ub	6	1	0,24	0,01	0,22	0,26	629,88	0,00	25,10	1,27	0,00
Tipo_uso_ub	8	1	-0,36	0,01	-0,38	-0,33	761,39	0,00	-27,59	0,70	0,00
Tipo_uso_ub	9	1	-0,23	0,01	-0,25	-0,21	351,90	0,00	-18,76	0,79	0,00
Tipo_uso_ub	10	0	0,00	0,00	0,00	0,00				1,00	
Ant_pol	1	1	0,43	0,01	0,42	0,44	4.346,81	0,00	65,93	1,54	0,00
Ant_pol	2	1	0,37	0,01	0,36	0,38	3.831,32	0,00	61,90	1,45	0,00
Ant_pol	3	1	0,29	0,01	0,28	0,30	2.171,15	0,00	46,60	1,33	0,00
Ant_pol	4	1	0,22	0,01	0,21	0,23	1.432,24	0,00	37,84	1,24	0,00
Ant_pol	6	1	0,12	0,01	0,11	0,13	502,33	0,00	22,41	1,13	0,00
Ant_pol	10	0	0,00	0,00	0,00	0,00				1,00	

Esta tabla muestra los resultados del proceso iterativo de las estimaciones de los parámetros.

Para cada parámetro del modelo, el procedimiento GENMOD de SAS, muestra una serie de columnas con el nombre del parámetro, los grados de libertad asociados al mismo, el error estándar cometido en la estimación del parámetro, los intervalos de confianza al 95% y, por último, el estadístico chi cuadrado de Wald que prueba la importancia del parámetro para el modelo.

En el modelo propuesto, se puede comprobar como todos los parámetros de cada instancia de las variables son significativos al 95% de nivel de confianza, ya que todos los p-valores asociados al estadístico chi cuadrado son inferiores a 0,05, por lo que se rechazaría la hipótesis nula donde el parámetro estimado sería cero.

Cuando se generó el modelo GLM, se utilizó como perfil base el siguiente: asegurado con forma de pago anual, superficie entre 75-90 metros cuadrados, año de estudio 2011/2012, capital asegurado superior a 300.000 euros para el edificio y a 42.000 euros para el mobiliario, clasificación del inmueble como otras viviendas residencia habitual en casco urbano y una antigüedad de póliza superior a 10 años.

El siguiente paso fue calcular las relatividades de la siguiente manera:

```
*Estimación de las relatividades;
Data TFM.parmest;
      Set TFM.parmest;
      If stderr <> 0 then
      tval=estimate/stderr;
      Rel=round (exp (estimate) ,0.0001);
      Pval=probchisq;
      Format pval percent7.1;
Run;
```

Es decir, tomando exponenciales de los parámetros estimados.

Observando el output de parámetros de SAS, vemos como la columna de las relatividades de las instancias asignadas al perfil base son uno, por lo tanto no modifican la frecuencia base del asegurado base, correspondiendo está frecuencia al 24% del intercept.

Lógicamente, si cambiásemos de asegurado y éste tuviese otras características, su frecuencia variaría, por ejemplo, ceteris paribus, excepto para la forma de pago. Supongamos que un asegurado escoge la modalidad de pago fraccionado, observado la tabla generada, comprobamos como su frecuencia estimada habría que multiplicarla (ya que se trata de un modelo multiplicativo) por 1,38, quedando su frecuencia en un 33%.

El siguiente paso será comprobar el ajuste y evaluación del modelo. Para ello el paquete estadístico SAS genera una serie de salidas donde podemos evaluar los criterios de bondad del ajuste del modelo.

Criterion	DF	Value	Value/DF
Deviance	45.802	65.045,66	1,42
Scaled Deviance	45.802	45.802,00	1,00
Pearson Chi-Square	45.802	70.413,51	1,54
Scaled Pearson X2	45.802	49.581,78	1,08
Log Likelihood		2.058.870,66	
Full Log Likelihood		-83.138,65	
AIC (smaller is better)		166.355,30	
AICC (smaller is better)		166.355,37	
BIC (smaller is better)		166.695,88	

Los criterios para evaluar la bondad del ajuste mostrados en este output, contienen estadísticos que resumen el ajuste del modelo especificado. Estos estadísticos son útiles para juzgar la adecuación del modelo y poder comparar con otros modelos propuestos.

Teniendo en cuenta los grados de libertad, el resultado de la deviance es bastante bueno, ya que es muy próximo a 1.

Sin embargo, si queremos analizar más en detalle la deviance, SAS nos permite realizar dos tipos de análisis para comprobar la aportación de cada variable a la deviance total del modelo. Estos análisis son el Type 1 y el Type 3.

Source	Deviance	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq
Intercept	175.192,38						
Forma_pago	161.366,67	1	45.802	9.735,40	0,00000	9.735,40	0,00000
Superficie	125.828,76	5	45.802	5.004,81	0,00000	25.024,07	0,00000
Año_estudio	125.172,33	1	45.802	462,22	0,00000	462,22	0,00000
Cap_edif	116.887,48	12	45.802	486,15	0,00000	5.833,79	0,00000
Cap_mob	96.753,32	5	45.802	2.835,50	0,00000	14.177,50	0,00000
Uso_tip_ub	75.517,21	9	45.802	1.661,49	0,00000	14.953,44	0,00000
Ant_pol	65.045,66	5	45.802	1.474,71	0,00000	7.373,55	0,00000

Esta tabla muestra los resultados del análisis Type 1. Este análisis consiste en el ajuste de una secuencia de modelos, comenzando por el modelo más sencillo posible con un solo parámetro (el intercept) y continuando a través de un modelo de complejidad especificada, encajando un efecto adicional en cada paso. Este proceso requiere un gran tiempo computacional, ya que es necesario ir calculando la deviance del modelo cada vez que se introduce una variable nueva al mismo.

Este tipo de análisis a veces se llama análisis de la deviance, ya que si el parámetro de dispersión se mantiene fijo para todos los modelos, es equivalente al cálculo de las diferencias de deviance a escala.

Como podemos observar, cada entrada en la columna de la deviance representa la deviance para el modelo que contiene el efecto para esa fila y todos los efectos anteriores en la tabla. Por ejemplo, la deviance correspondiente a la fila de forma de pago en la tabla es la deviance del modelo que contiene el intercept y la forma de pago. A medida que más términos se incluyen en el modelo, la deviance disminuye, señal de que cada variable aporta a nuestro modelo. Esta última afirmación, podemos corroborarla si observamos los p-valores asociados a los estadísticos de chi cuadrado. Todos ellos son inferiores a 0,05 por tanto, cada variable produce un efecto añadido altamente significante en el modelo.

Source	Num DF	Den DF	F Value	Pr > F	Chi-Square	Pr > ChiSq	Method
Forma_pago	1	45.802	3.892,45	0,00000	3.892,45	0,00000	LR
Superficie	5	45.802	549,10	0,00000	2.745,48	0,00000	LR
Año_estudio	1	45.802	272,82	0,00000	272,82	0,00000	LR
Cap_edif	12	45.802	562,45	0,00000	6.749,34	0,00000	LR
Cap_mob	5	45.802	1.722,15	0,00000	8.610,76	0,00000	LR
Uso_tip_ub	9	45.802	1.603,46	0,00000	14.431,17	0,00000	LR
Ant_pol	5	45.802	1.474,71	0,00000	7.373,55	0,00000	LR

Esta tabla muestra los resultados del análisis Type 3. Este análisis es similar a la suma de cuadrado Type III utilizada en un procedimiento GLM en SAS, excepto que se usan los ratios de similitud en lugar de la suma de los cuadrados.

La máxima verosimilitud es calculada bajo la restricción de que la función Type III es cero, utilizando un proceso de optimización forzado.

Luego el parámetro estimado es $\tilde{\beta}$ y el logaritmo de la verosimilitud $l(\tilde{\beta})$, siendo el estadístico del ratio de verosimilitud:

$$S = 2\left(l(\hat{\beta}) - l(\tilde{\beta})\right)$$

Donde $\hat{\beta}$ es el parámetro sin restricciones y sigue una distribución chi cuadrado asintótica bajo la hipótesis de que el análisis Type III es igual a cero con un número de grados de libertad igual al número de parámetros asociados.

En resumen, el output del análisis Type 3 generado presenta las mismas conclusiones que el análisis Type 1. La hipótesis testada en este caso es la significancia de cada variable, dado que la anterior variable está en el modelo, es decir, mide la contribución adicional de cada variable en el modelo.

Si observamos los p-valores asociados a cada estadístico chi cuadrado calculado, podemos comprobar como todas las variables tienen una contribución significativa en nuestro modelo GLM.

El último paso para evaluar la adecuación de nuestro modelo, será analizar los residuos del mismo.

En los modelos GLM es habitual trabajar con residuos estandarizados, ya que si las suposiciones del modelo son correctas, éstos deben tener aproximadamente la misma varianza y comportarse, en la medida de lo posible, como residuos de un modelo de regresión lineal simple.

La manera más utilizada en la práctica, es trabajar con los residuos estandarizados de Pearson:

$$\hat{\epsilon}_i^p = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}$$

Debiendo tener aproximadamente media cero y varianza ϕ, si el modelo es correcto.

Sin embargo, al trabajar con dato real, suele suceder que la distribución de los residuos de Pearson, sea muy asimétrica alrededor de cero, por lo que su comportamiento podría distar bastante de lo esperado en un modelo de regresión ordinario. Es por ello, que se utilizan los residuos de la deviance.

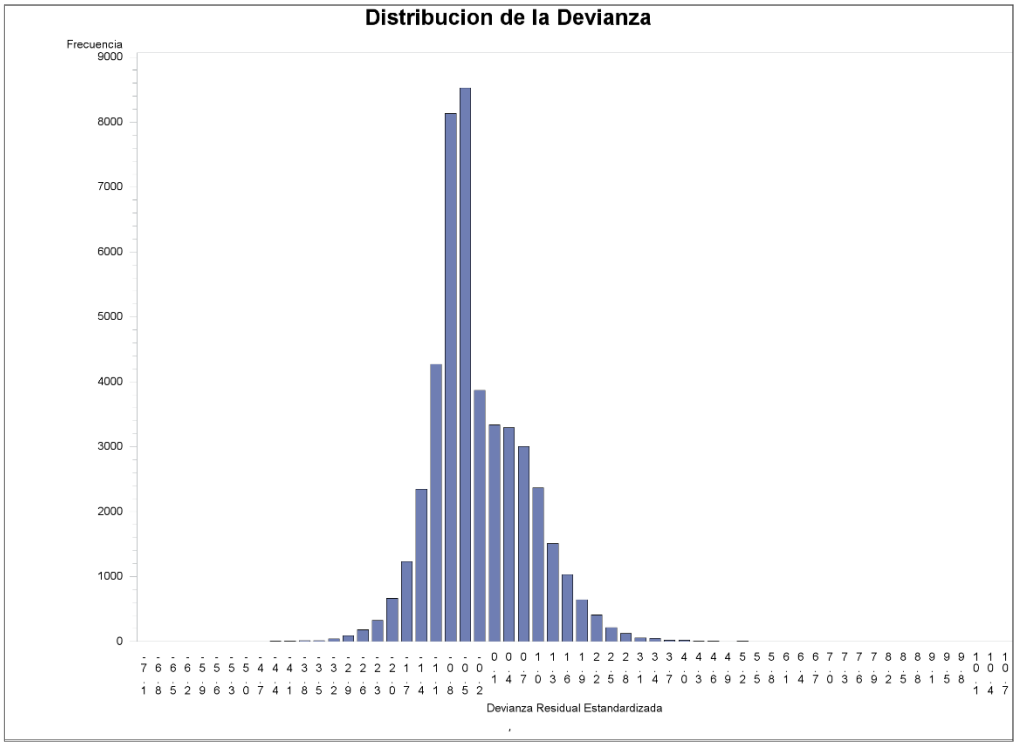
Si denotamos por d_i al componente de la deviance que aporta la i-ésima observación, tendremos:

$$D = \sum_{i=1}^{n} d_i$$

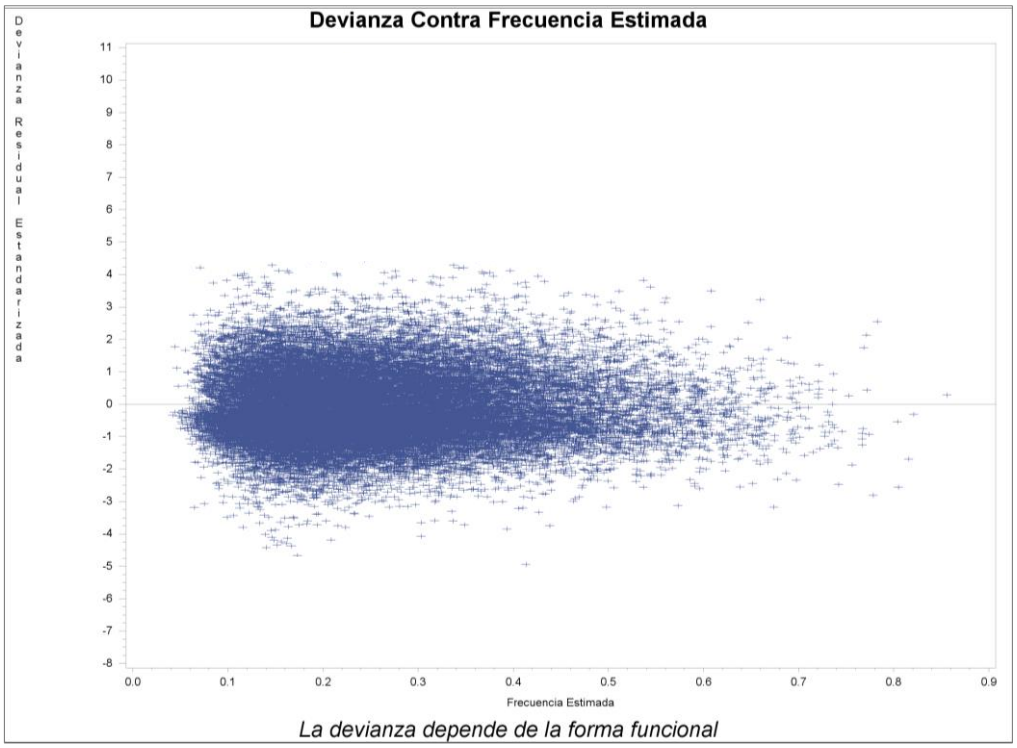
Y, de manera análoga al modelo lineal ordinario, definiremos:

$$\hat{e}_i^d = sign(y_i - \hat{\mu}_i)\sqrt{d_i}$$

Vamos a comprobar la simetría de los mismos representándolos en un histograma:



La deviance residual tiene media cero y, además, parece presentar simetría. Vamos a comprobarlo representando la deviance residual de cada observación en una nube de puntos:



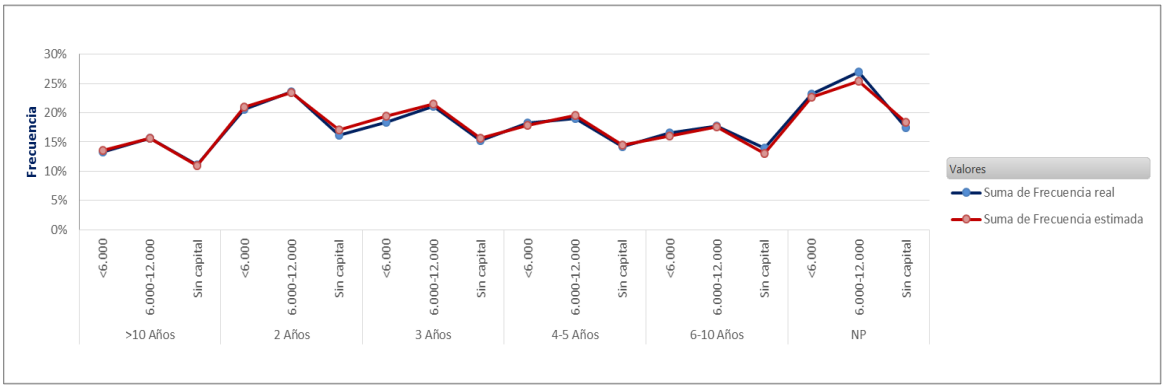
Observamos como aparece una casi perfecta simetría entre los valores de la deviance, sin embargo, en el siguiente apartado, estudiaremos los resultados de nuestro modelo haciendo una comparativa entre las frecuencias observadas empíricamente y las frecuencias estimadas.

2
Estudio de la frecuencia real y la estimada

En este apartado realizaremos una comparativa entre las frecuencias empíricas y las estimadas por nuestro modelo en base a una clusterización de asegurados.

Es importante señalar que en un GLM para conteos, a diferencia de un GLM logístico, lo que se intenta predecir son medias, por lo tanto cuanta más exposición tengamos en nuestros clúster, más parecido será el valor explicado con el valor real.

Por ejemplo, si queremos comparar las frecuencias reales con las estimadas en función de la antigüedad de la póliza y el capital mobiliario asegurado, tendremos:

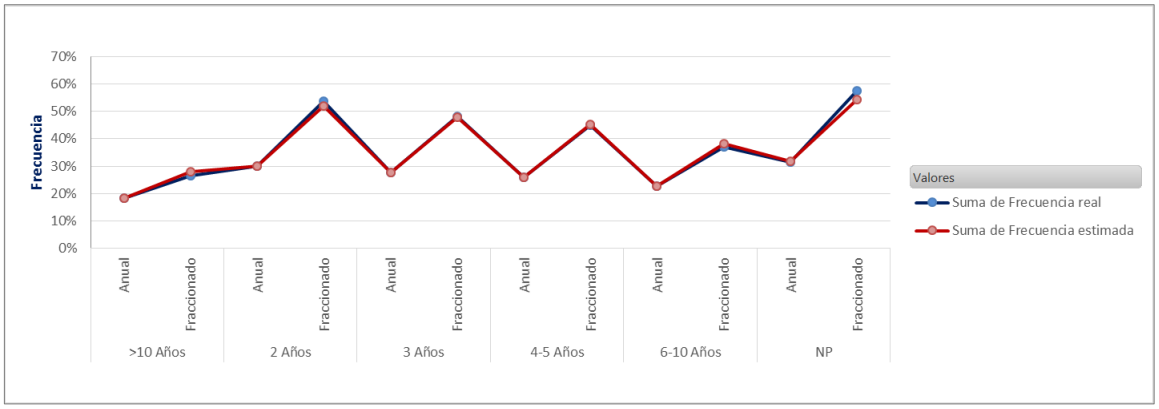


Podemos comprobar como al tener clúster tan nutridos, nuestro GLM trabaja muy bien, ya que en todos los clúster, los valores explicados son muy similares a los empíricos.

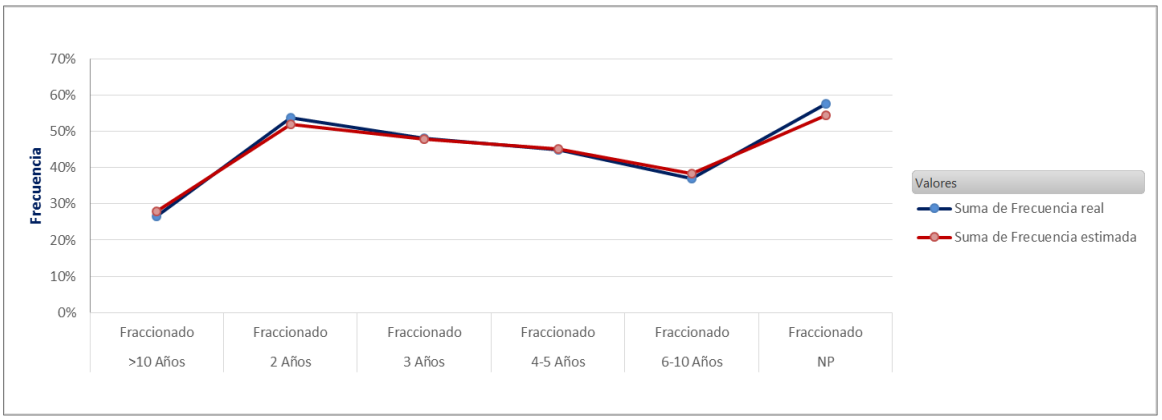
La mayor diferencia se produciría en la nueva producción con capital asegurado entre 6.000 y 12.000 euros, ya que la frecuencia real sería del 27% y la estimada del 25%.

Lógicamente, en función de que trabajemos con clúster con menor exposición, la media será menos representativa, por lo que se producirán más diferencias. Además, podría ser peligroso para realizar predicciones, ya que los valores predichos estarían sujetos a una enorme variabilidad.

Comparemos ahora ambas frecuencias utilizando la antigüedad de la póliza con la forma de pago:

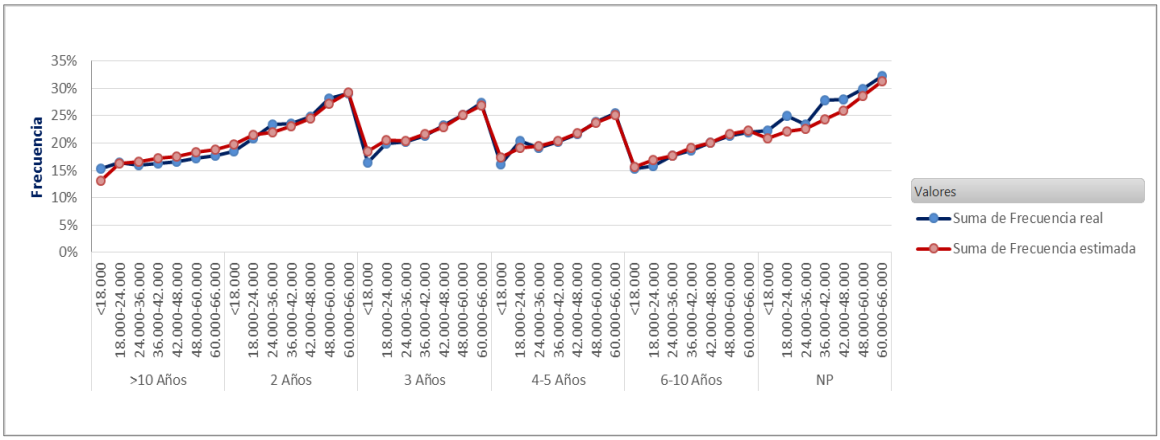


Vemos como de nuevo el GLM trabaja muy bien, sin embargo, observamos que las diferencias más grandes se producen en forma de pago fraccionada. Esto es algo esperado ya que sólo el 5% de la cartera escogía esta modalidad. Estudiemos sólo la frecuencia para el pago fraccionado:



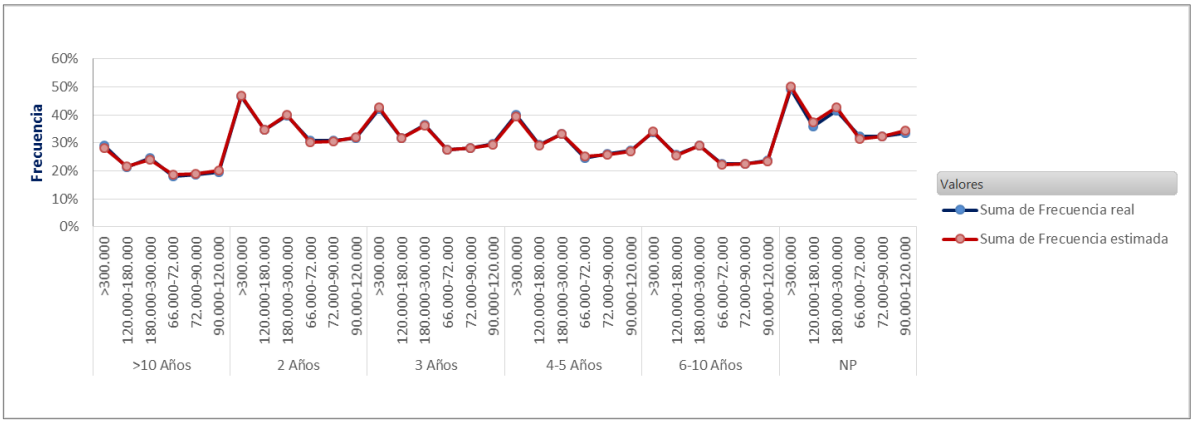
Las diferencias se producen en la nueva producción con una frecuencia real del 58% frente a una frecuencia estimada del 54%. En pólizas de 2 años de antigüedad, también se producen divergencias, ya que la frecuencia real es del 54% y la estimada del 52%.

Para el siguiente análisis, introduciremos el capital asegurado del edificio con la antigüedad de la póliza. En un primer paso, analizaremos capitales asegurados inferiores a 66.000 euros:



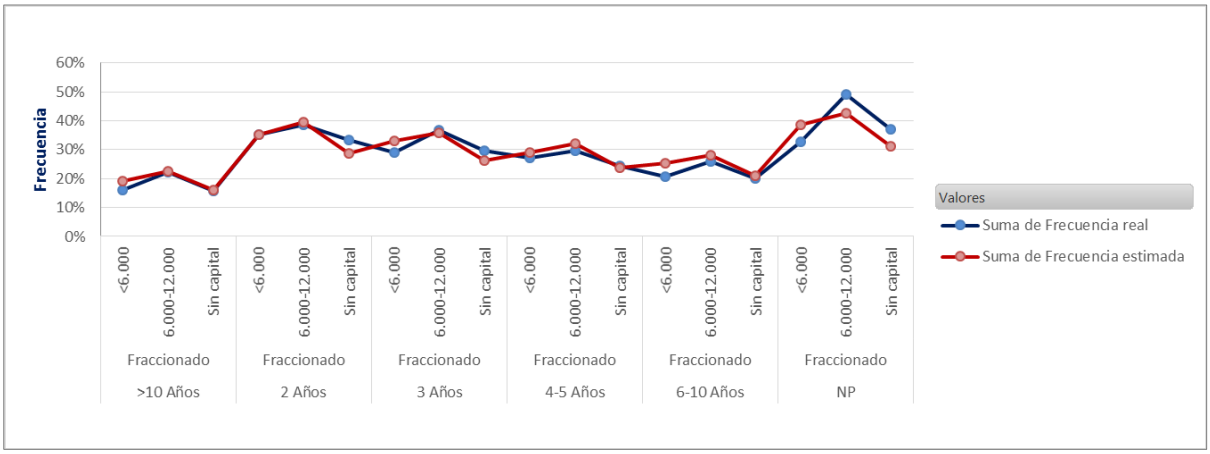
De nuevo, las diferencias más importantes se producen en la nueva producción, más concretamente en los tramos de capital de 18.000-24.000 euros, donde la frecuencia real es del 25% y la estimada del 22%, y de 36.000-42.000 euros, donde la frecuencia real es del 28% frente a la estimada del 24%.

A continuación se realizará la misma comparativa pero con tramos de capital mayores de 66.000 euros:



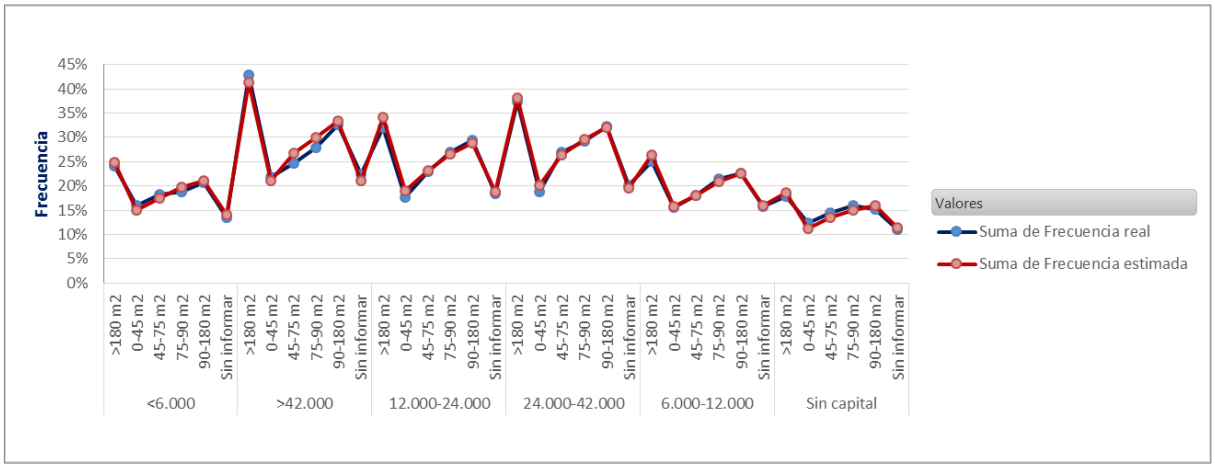
Para capitales superiores a 66.000 euros, el modelo trabaja mejor, ya que los clúster tienen mucha más exposición.

Un análisis interesante sería introducir a la antigüedad de la póliza y a su forma de pago con modalidad fraccionado (debido a su menor exposición) el capital mobiliario asegurado:



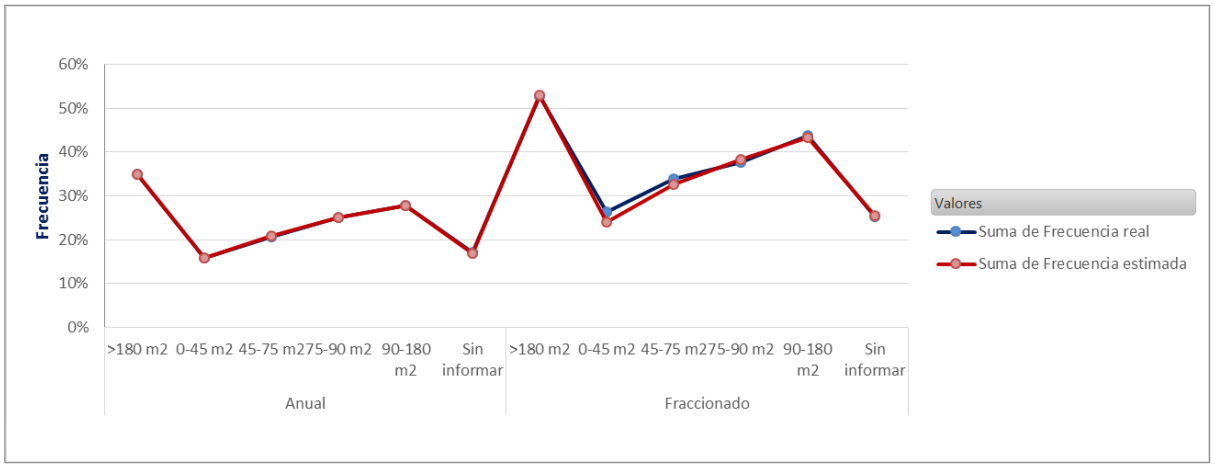
Al segmentar mucho más nuestra cartera, se producen más diferencias entre la frecuencia real y la estimada, sin embargo, el modelo sigue trabajando muy bien, ya que la mayor divergencia se produce, una vez más, en la nueva producción con modalidad de pago fraccionado y capital entre 6.000 y 12.000 euros, donde la frecuencia real es del 49% y la estimada del 43%.

El análisis previo ha sido realizado teniendo en cuenta la variable antigüedad de la póliza como referencia. Vamos a realizar el mismo estudio a continuación, pero utilizando ahora la variable superficie:



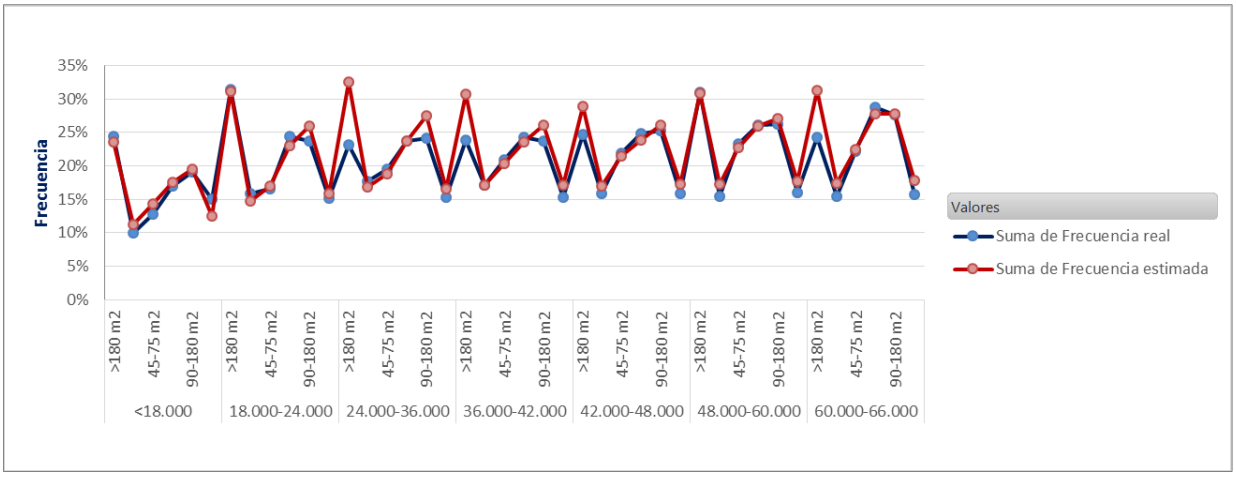
En líneas generales, el modelo explica muy bien, ya que ningún clúster presenta diferencias entre las frecuencias real y estimada superiores a tres puntos porcentuales.

A continuación, se analizarán las frecuencias estimadas de la modalidad de pago escogida en función de la superficie del inmueble:



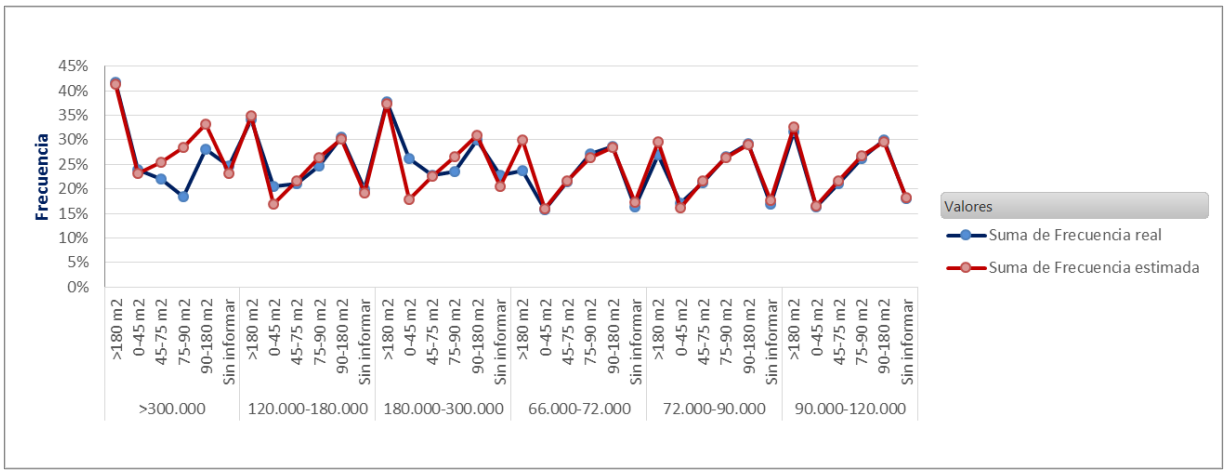
Al igual que sucedía cuando estudiamos la variable antigüedad con la forma de pago, las diferencias se producen cuando se utiliza el pago fraccionado, sin embargo éstas no son importantes ya que nunca exceden del 2%.

El siguiente análisis se realizará teniendo en cuenta capitales asegurados del edificio inferiores a 66.000 euros para cada superficie del inmueble:



Es interesante comprobar una vez más como las diferencias más importantes se producen en aquellos clúster con menor exposición. En este caso las diferencias son muy pronunciadas, ya que en aquellos inmuebles de más de 180 metros cuadrados, las diferencias entre la frecuencia real y la estimada pueden llegar al 10%. Esto es algo totalmente comprensible, ya que los inmuebles de dicha superficie tan sólo representan el 13% de la cartera y el capital asegurado de 24.000 euros donde se produce esa diferencia representa solamente el 0,60% de la cartera.

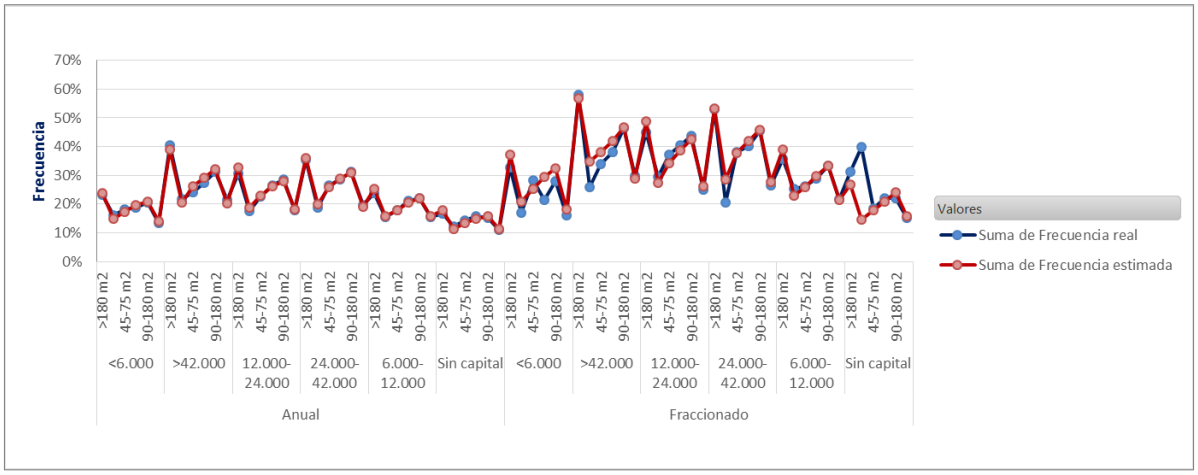
Realizaremos a continuación el mismo estudio pero para capitales superiores a 66.000 euros:



Para capitales superiores a 300.000 euros es donde mayores diferencias se dan, más concretamente en superficies comprendidas entre 75 y 90 metros cuadrados.

Los capitales asegurados superiores a 300.000 euros tan sólo representan un 4% de la cartera, de ahí que se produzca en ese clúster la mayor diferencia entre la frecuencia real y la estimada (10 puntos porcentuales).

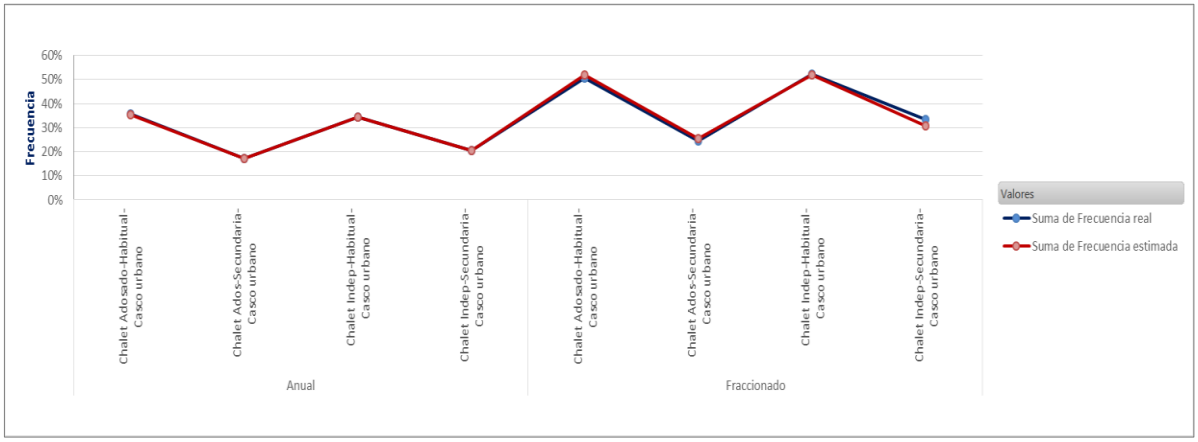
Por último, al igual que hicimos con la variable antigüedad de la póliza, realizaremos un análisis de la superficie del inmueble en función de la forma de pago elegida y el capital mobiliario asegurado:



Cuando se fija la modalidad de pago anual el modelo trabaja francamente bien, pero si analizamos la modalidad de pago fraccionado observamos alguna diferencia importante, como en superficies de más de 180 metros cuadrados y sin capital asegurado, donde la frecuencia real es del 40% y la estimada del 14%. Esta diferencia está producida básicamente por la poca exposición del clúster, ya que solamente el pago fraccionado representa el 5% de la cartera.

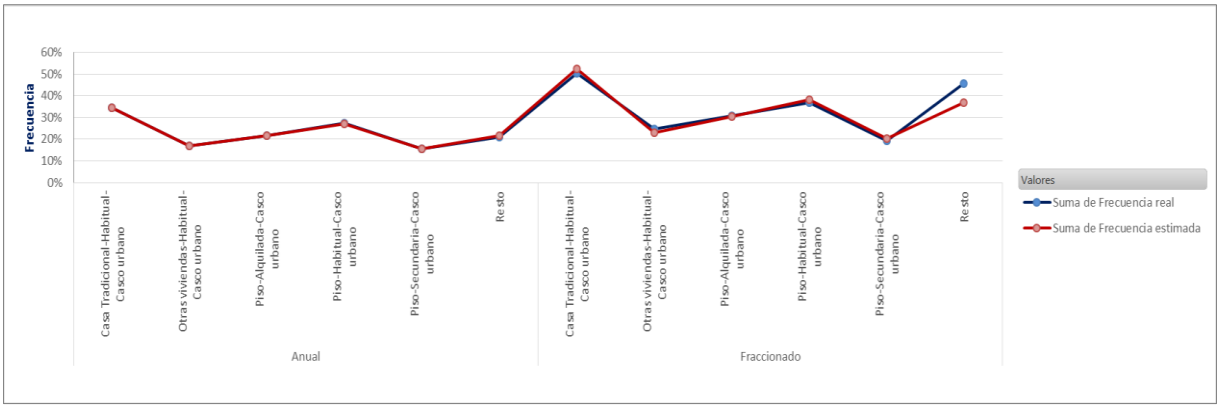
Por último, realizaremos un análisis del tipo del inmueble, su uso y su ubicación en función de la tipología de capital asegurado y forma de pago.

Vamos a observar la frecuencia real y la estimada de los chalets en función de la forma de pago:



En líneas generales el modelo devuelve frecuencias prácticamente iguales a las reales. La única diferencia se produce en chalets independientes utilizados como segunda vivienda en casco urbano con modalidad de pago fraccionado, donde la frecuencia real es del 33% y la estimada del 31%.

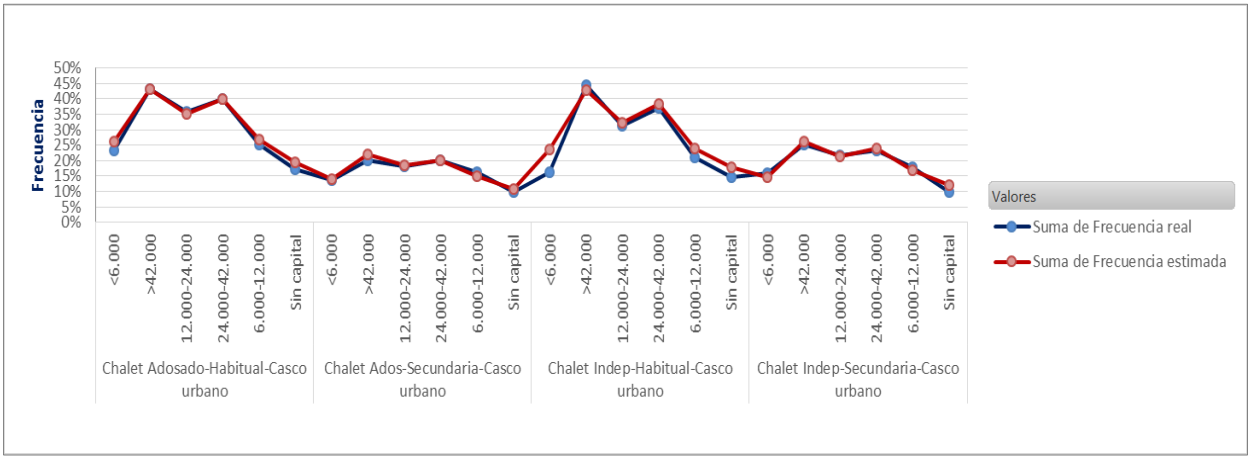
Si realizásemos el mismo análisis para pisos y resto de viviendas que no sean chalets, tendremos:



La única diferencia entre frecuencia real y estimada se produce para viviendas clasificadas como resto cuando se utiliza la modalidad de pago fraccionado. Esta diferencia es de 9 puntos porcentuales (46% de frecuencia real frente a un 37% de frecuencia estimada).

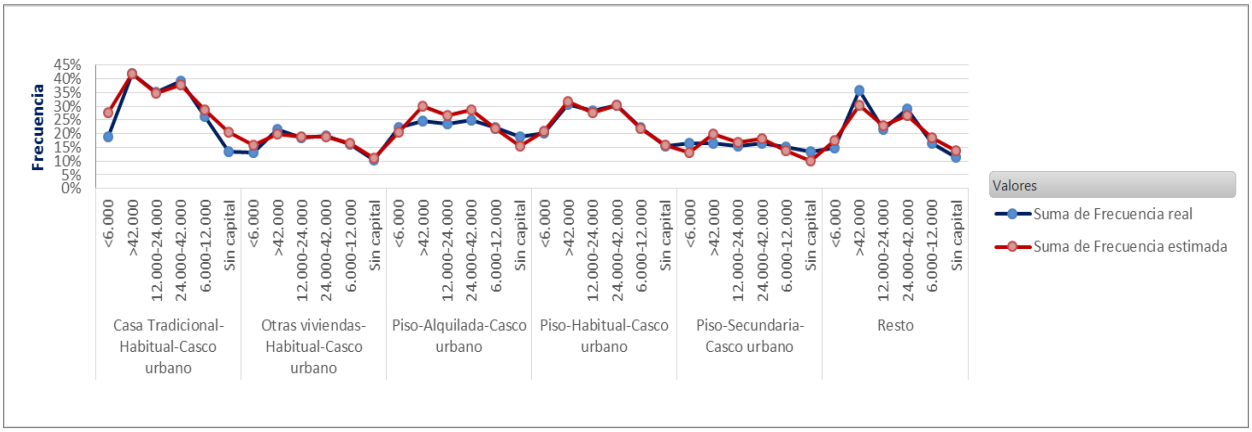
Introduciremos a continuación el capital mobiliario para ver cómo se comportan las frecuencias en función del tipo, uso y ubicación del inmueble.

Primero realizaremos el estudio para chalets:



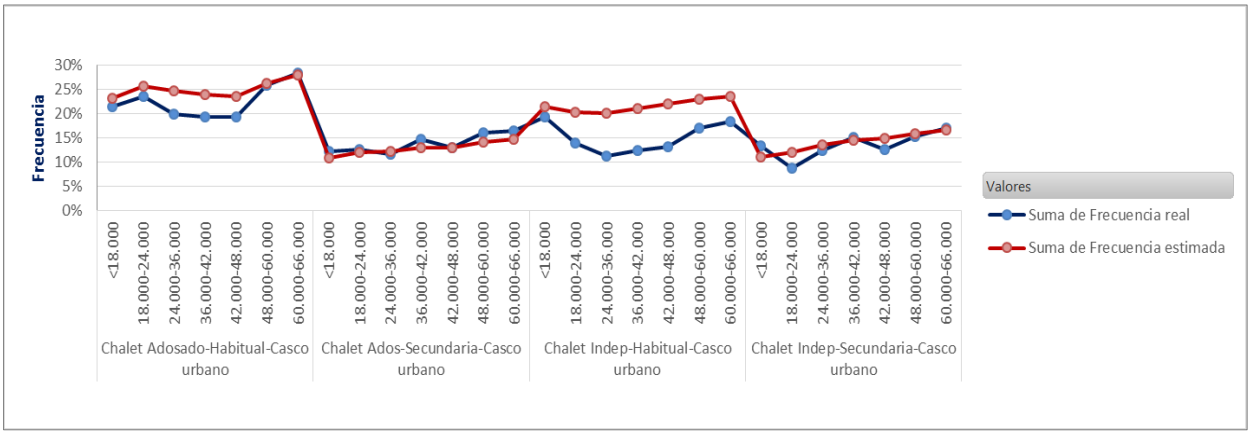
En chalets independientes utilizados como residencia habitual y situados en casco urbano con un capital mobiliario asegurado inferior a 6.000 euros se produce la mayor diferencia entre frecuencia real y estimada. Esta diferencia, de 7 puntos porcentuales, corresponde a una frecuencia real del 16% y una estimada del 23%. En el resto de casuísticas, el modelo explica bastante bien.

A continuación realizaremos el mismo análisis para el resto de tipología de viviendas:



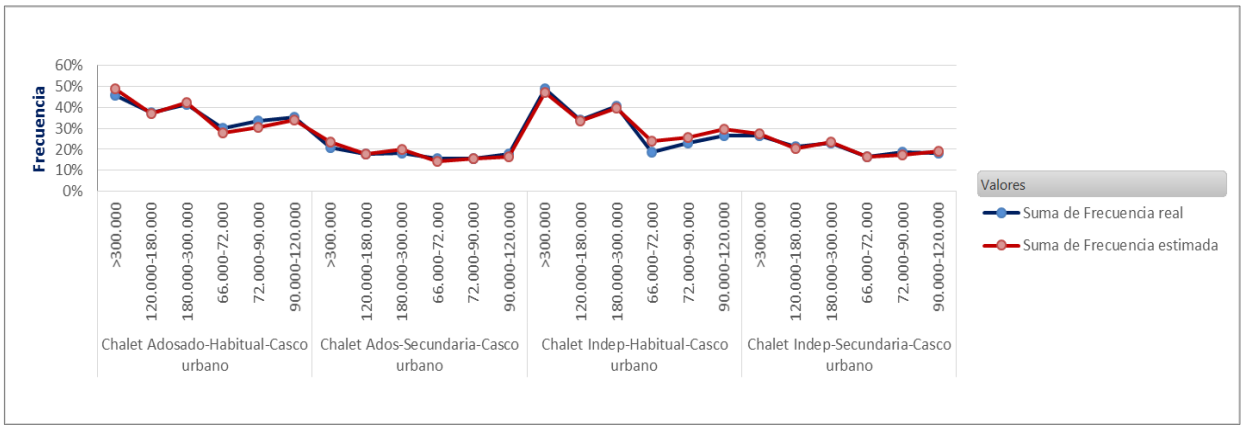
En casas tradicionales utilizadas como vivienda habitual y situadas en casco urbano es donde se producen las mayores diferencias entre frecuencia real y estimada, siendo la mayor de ellas de 9 puntos porcentuales para capitales mobiliarios inferiores a 6.000 euros.

A continuación, en lugar de utilizar el capital mobiliario, haremos uso del capital asegurado del edificio. En un primer paso analizaremos chalets asegurados en menos de 66.000 euros:

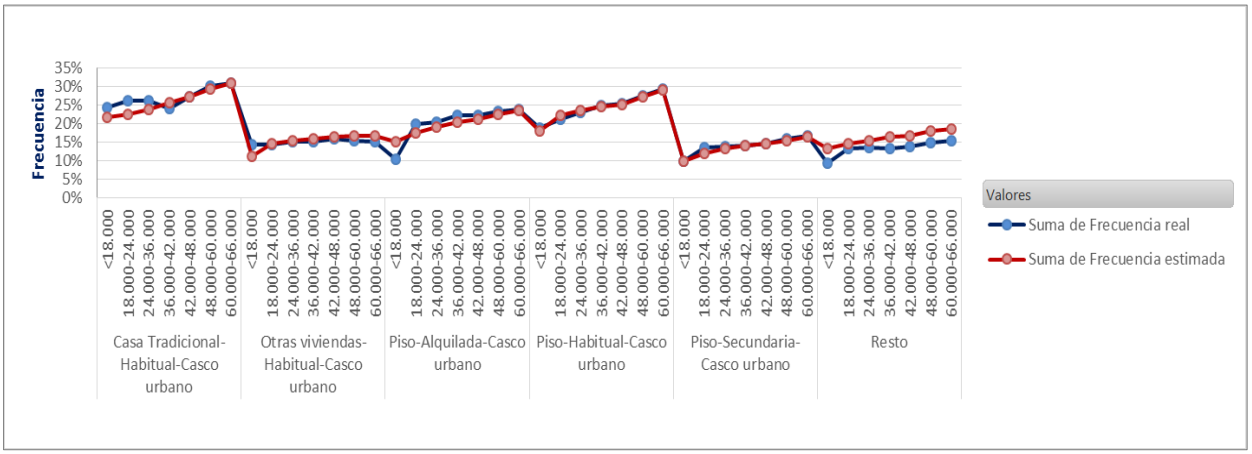


En chalets independientes utilizados como residencia habitual y situados en casco urbano vuelven a producirse diferencias importantes en torno a 10 puntos porcentuales. Es algo totalmente comprensible, ya que esta tipología de inmueble tan sólo representa un 4% de la cartera.

Podemos comprobar cómo se produce el mismo efecto para capitales superiores a 66.000 euros, aunque el efecto es mucho menos pronunciado, ya que en esta casuística la mayor diferencia es de 6 puntos porcentuales y sólo se produce en capitales asegurados que oscilan entre 66.000 y 72.000 euros:

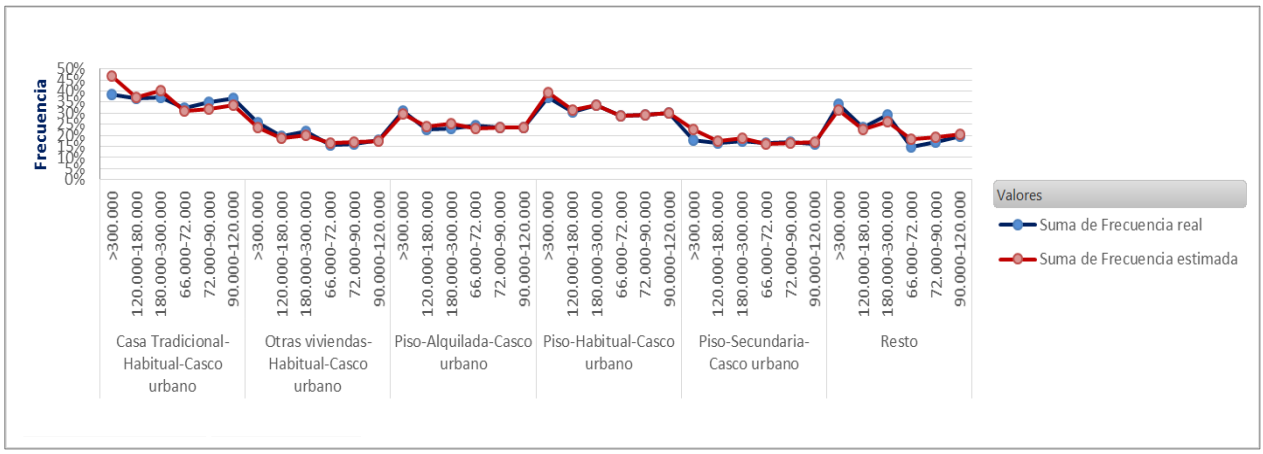


A continuación, realizaremos el mismo estudio para el resto de tipo de viviendas, empezando por capitales asegurados inferiores a 66.000 euros:



Como sucedió con anterioridad, donde se producen más diferencias es en las viviendas clasificadas como resto, sin embargo, éstas no son importantes, ya que nunca exceden de 4 puntos porcentuales.

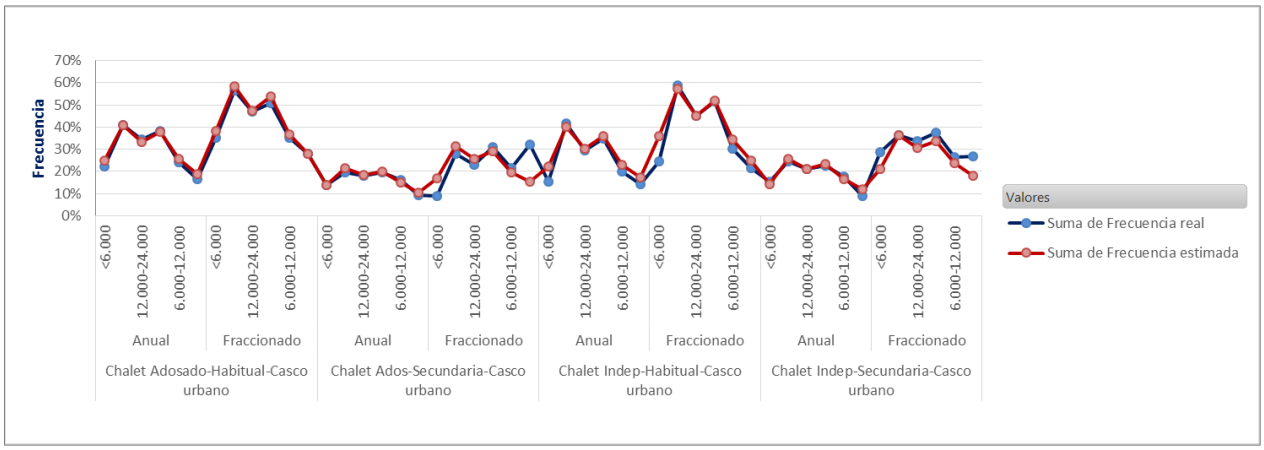
Vamos lo que sucede en capitales asegurados superiores a 66.000 euros:



En este caso, el modelo funciona muy bien, debido a la mayor exposición de los clúster respecto a la casuística anterior.

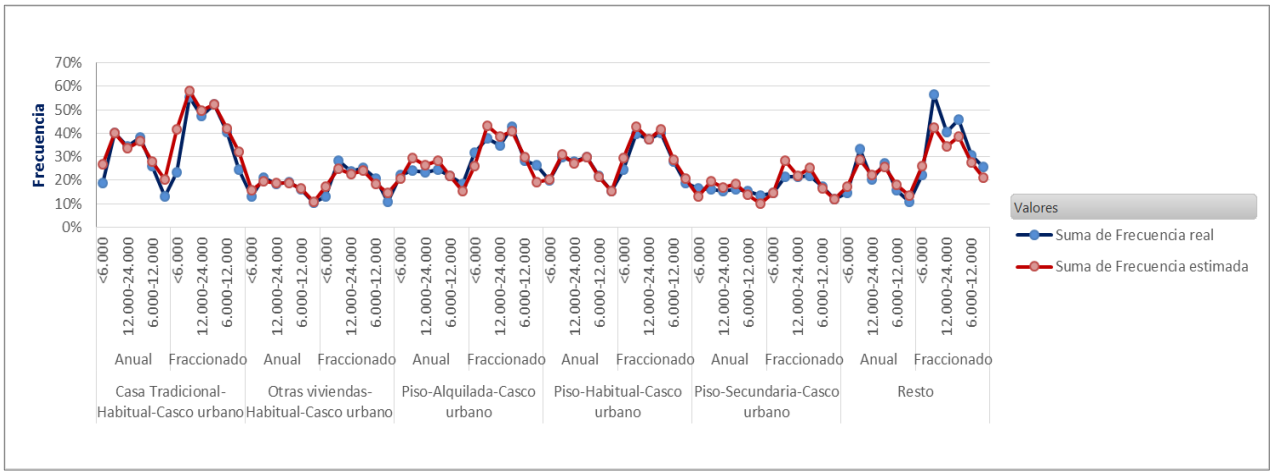
Por último realizaremos un análisis utilizando el tipo uso y ubicación de la vivienda con la forma de pago escogida y el capital mobiliario.

Analizaremos primero los chalets:



El modelo trabaja muy bien en líneas generales, apareciendo diferencias en pago fraccionado como es habitual, siendo la más importante aquella en chalets adosados utilizados como segunda vivienda situados en casco urbano y con capital mobiliario asegurado entre 6.000 y 12.000 euros, ya que la frecuencia real es del 32% y la estimada del 15%.

A continuación, estudiaremos el resto de tipologías de viviendas:



Tal y como se esperaba, en la modalidad de pago fraccionado y la tipología de vivienda clasificada como resto es donde se producen más divergencias entre las frecuencias real y estimada. Más concretamente, en capitales mobiliarios inferiores a 6.000 euros, se produce una diferencia importante de 15 puntos porcentuales, ya que la frecuencia estimada es del 42% y la frecuencia real del 57%.

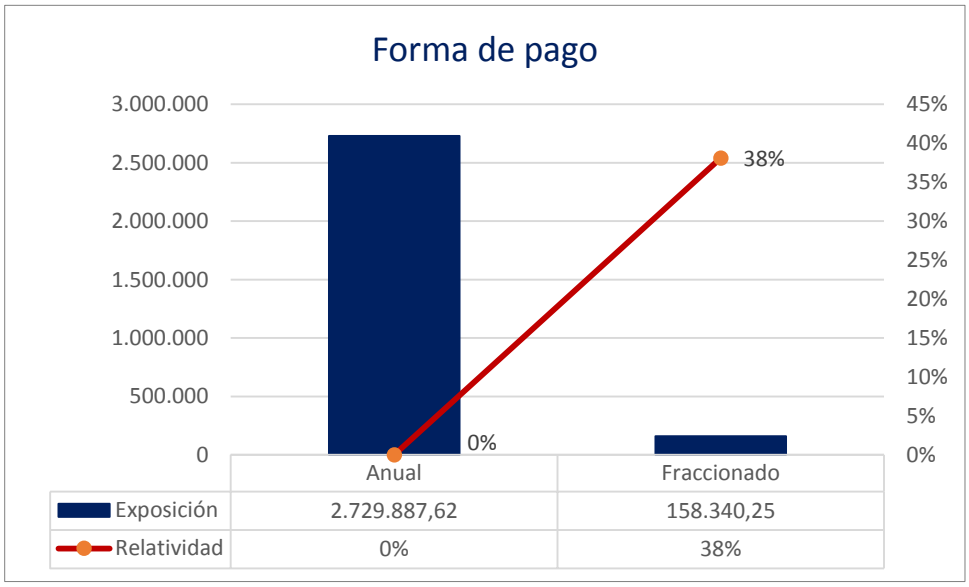
3
Impacto en negocio

El objetivo de este capítulo, es tratar de ilustrar el impacto que va a tener en negocio el modelo propuesto. Para ello, y en función de cada variable, se tratará de explicar el recargo o bonificación respecto a la instancia elegida como base.

Este trabajo es muy común en la práctica aseguradora, ya que la tarificación suele ser un proceso en el que interviene no sólo el departamento técnico, sino que también suele estar involucrado el departamento de negocio.

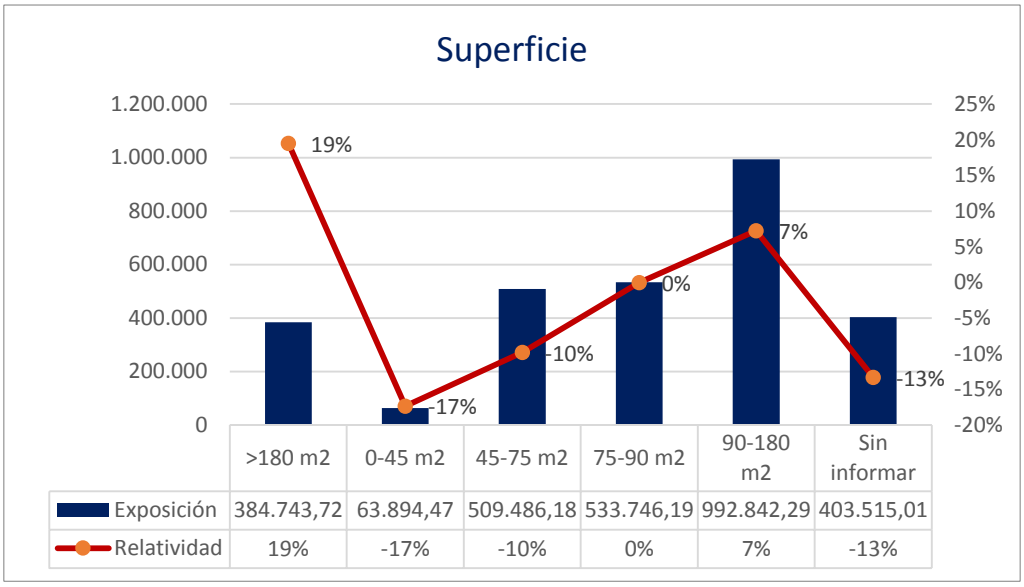
Gracias a este análisis, se pueden detectar incoherencias en el modelo propuesto, ya que por ejemplo, no sería lógico que la frecuencia en la modalidad de pago anual fuese superior a la fraccionada o que la antigüedad de póliza de más de 10 años tuviese mayor frecuencia que la nueva producción entre otros.

3.1 Forma de pago



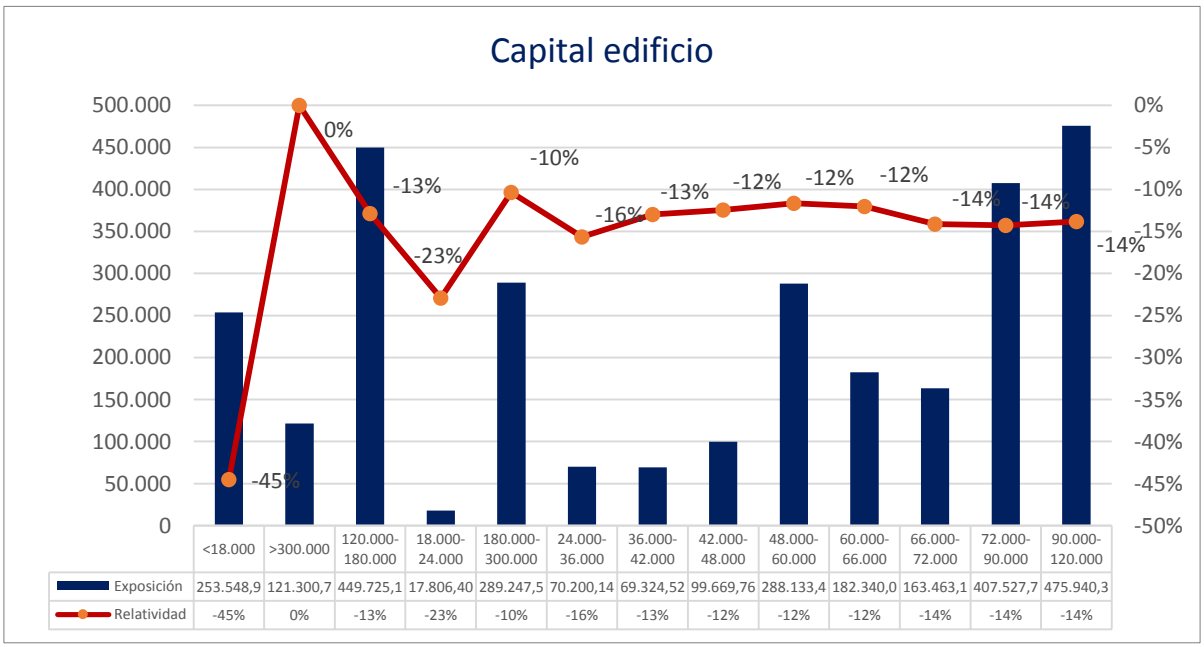
Como se mencionó anteriormente, tan sólo el 5% de la exposición total pertenece al pago fraccionado. Si el asegurado escoge esta modalidad, sufrirá un recargo del 38% respecto a los que elijan la modalidad anual en función de la frecuencia

3.2 Superficie



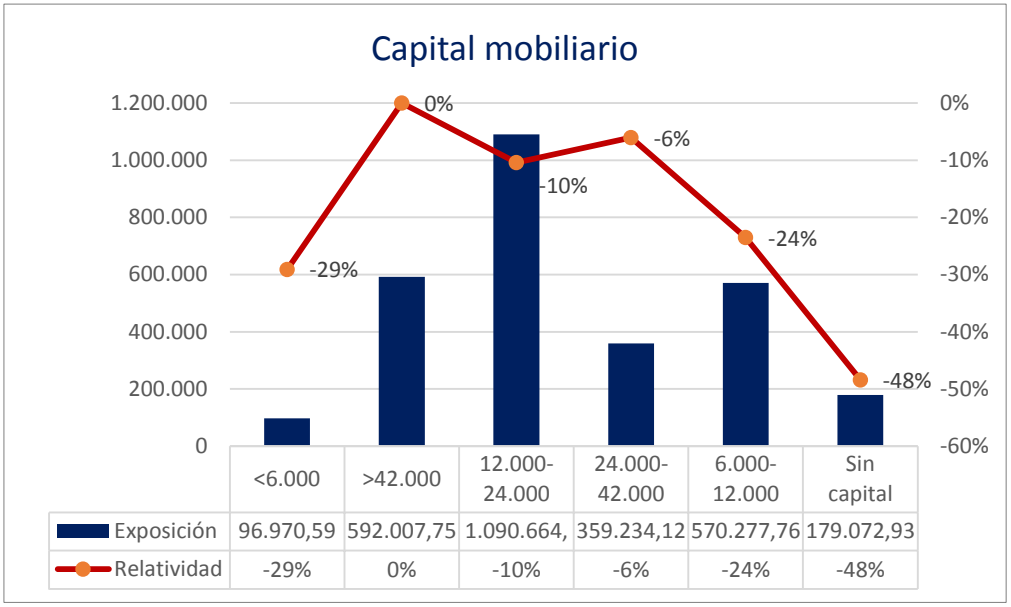
La mayor exposición se encuentra en los inmuebles de entre 75 y 180 metros cuadrados. Se ha elegido como perfil base los inmuebles de entre 75 y 90 metros cuadrados. A partir de ahí, conforme aumente la superficie, se recargará por frecuencia al asegurado. Si disminuye la superficie tomada como base, se bonificará al asegurado hasta un 17% en el caso de inmuebles cuya superficie sea menor o igual a 45 metros cuadrados.

3.3 Capital del edificio



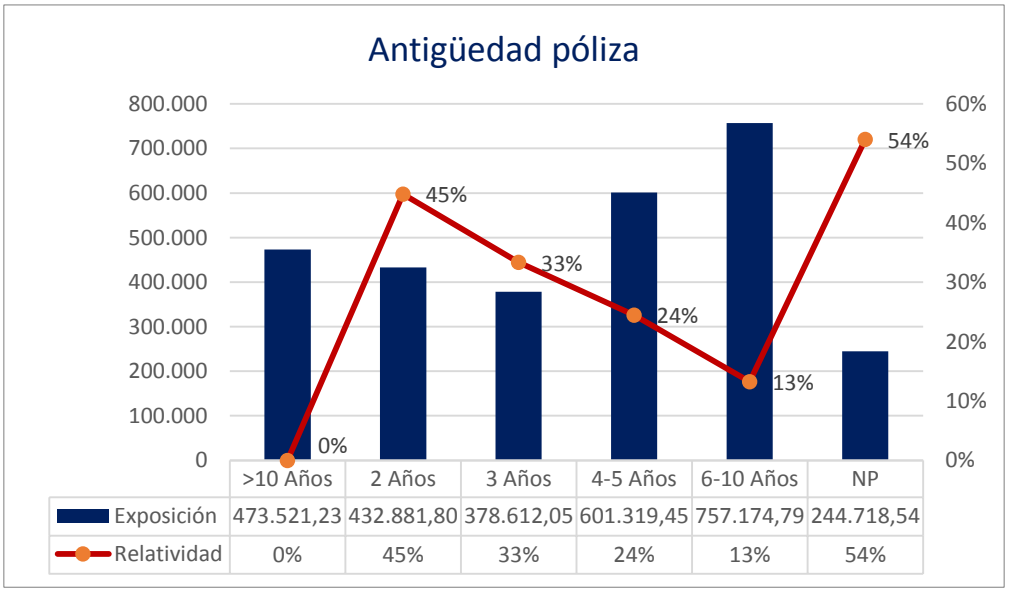
Se ha tomado como base los capitales de más de 300.000 euros. Para todos los asegurados con capitales inferiores, se les practicará bonificaciones en función de la frecuencia que oscilarán entre el 45% para los que aseguren menos de 18.000 euros y el 10% para los que aseguren entre 180.000 y 300.000 euros.

3.4 Capital mobiliario



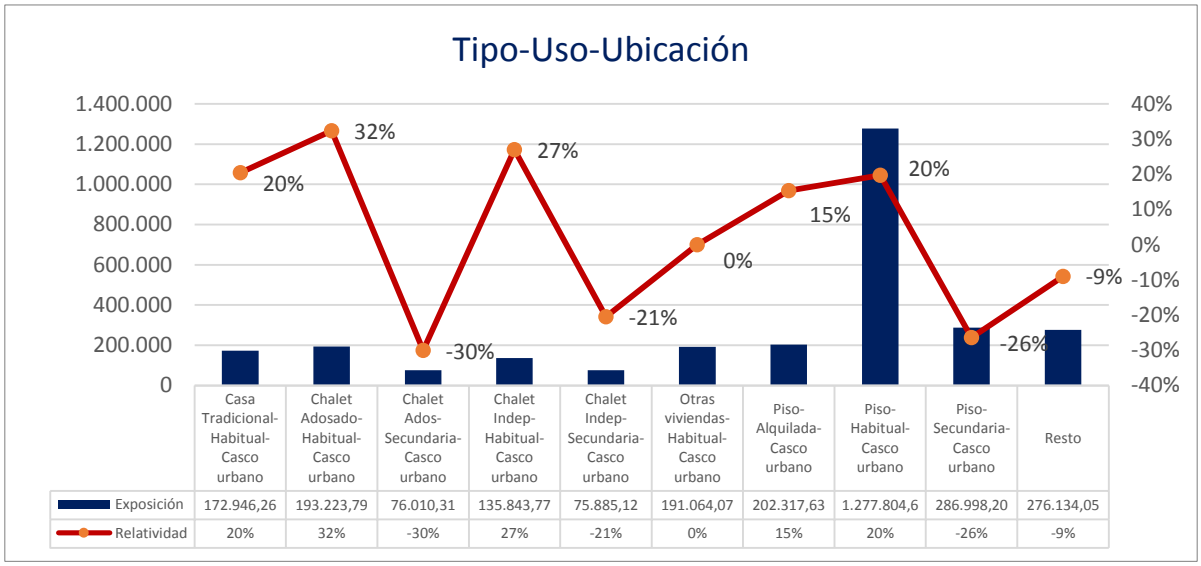
La mayor exposición se produce en aquellos asegurados cuyo capital mobiliario varía entre 12.000 y 24.000 euros. Se ha tomado como base el mayor tramo de capital (más de 42.000 euros). Los asegurados que tengan el mobiliario asegurado por menos valor, recibirán una bonificación que oscilará entre el 48% para los que no tengan capital y el 6% para los que lo aseguren entre 12.000 y 24.000 euros.

3.5 Antigüedad de la póliza



Como es lógico, se ha tomado como base las pólizas de más de 10 años de antigüedad, ya que son las que menor tasa siniestral tienen. La mayor exposición se encuentra en pólizas de entre 6 y 10 años de antigüedad. Las pólizas de nueva producción sufrirán un recargo del 54% y, conforme vayan ganando antigüedad, se les irá aplicando un recargo más suave con el objetivo de hacer un correcto pruning de la cartera.

3.6 Tipo, uso y ubicación del inmueble



Por último, en el tipo, uso y ubicación del inmueble, como es lógico, la mayor exposición se encuentra en inmuebles clasificados como piso que se utilizan como residencia habitual y están situados en el casco urbano.

Se ha utilizado como perfil base otras viviendas situadas en casco urbano y cuya residencia es habitual. Si el asegurado utiliza el inmueble como residencia secundaria, se le aplicará una bonificación, tanto en chalets como en pisos, estén situados o no en casco urbano. Sin embargo, si utiliza el inmueble como residencia habitual se practicarán recargos que oscilarán entre el 32% para los chalets adosados situados en casco urbano y el 15% para los pisos alquilados en casco urbano.

4
Construcción y evaluación del GAM

Tal y como se mencionó en capítulos anteriores, el objetivo de este trabajo es, mediante la utilización de un modelo aditivo generalizado, establecer el recargo o bonificación que se deberá aplicar a la tarifa en función de la frecuencia siniestral por la localización (código postal) del riesgo.

En la actualidad el territorio español está dividido en 11.752 códigos postales. Éstos constan de cinco dígitos, de los cuales, los dos primeros hacen referencia a la provincia por orden alfabético.

Para la construcción del modelo se ha utilizado una base de datos en la que vienen reflejadas las geocordenadas (latitud y longitud) a nivel código postal.

La variable dependiente del modelo será la deviance residual estandarizada del GLM construido en el capítulo anterior, y la variable independiente serán las geocordenadas del riesgo, de manera que asumiremos la hipótesis de que toda la parte no explicada del GLM vendrá determinada por el código postal.

Por tanto el modelo planteado será el siguiente:

The GAM Procedure
Dependent Variable: Stresdev
Smoothing Model Component(s): spline2(longitude latitude)
Frequency Variable: _FREQ_

```
Proc gam data=tfm.BBDDGAM PLOTS= (ALL) plots=components (additive);
      Model Stresdev=SPLINE2 (longitude, latitude, DF=10)/ dist=gaussian;
      Freq _freq_;
      Output out=out2 ALL;
Run;
Quit;
```

Para la modelización de la deviance residual estandarizada se han utilizado los tres millones de expuestos y se ha elegido la función vínculo gaussiana, ya que es la más apropiada para este tipo de estudio

Summary of Input Data Set	
Number of Observations	3.348.425
Number of Missing Observations	0
Distribution	Gaussian
Link Function	Identity

El output generado por el sistema SAS genera varias tablas. En la primera de ellas, resume el criterio de convergencia seguido en el algoritmo de back-fitting:

Iteration Summary and Fit Statistics	
Final Number of Backfitting Iterations	2
Final Backfitting Criterion	2,98E-23
The Deviance of the Final Estimate	149.031,75

Como podemos observar, se han realizado dos iteraciones hasta que el algoritmo de back-fitting ha convergido. Recordamos que el sistema estadístico SAS cuando trabaja con el procedimiento GAM utiliza el siguiente criterio de convergencia para el algoritmo de back-fitting:

$$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\left(f_j^{m-1}(x_{ij}) - f_j^{m}(x_{ij})\right)^2}{\sum_{j=1}^{p}\sum_{i=1}^{n} f_j^{m-1}(x_{ij})^2} \leq \varepsilon$$

Donde $\varepsilon = 10^{-8}$ por defecto.

La siguiente tabla recoge la estimación de los parámetros calculados en el proceso iterativo anterior:

Regression Model Analysis Parameter Estimates				
Parameter	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.08687	0.00011529	753.51	<.0001

Observando el output generado, podemos comprobar como el p-valor asociado al estadístico t calculado es altamente significativo por lo podremos utilizar posteriormente el parámetro estimado para el cálculo de las relatividades.

La siguiente tabla muestra el resumen del ajuste de los componentes del suavizado:

Smoothing Model Analysis Fit Summary for Smoothing Components				
Component	Smoothing Parameter	DF	GCV	Num Unique Obs
Spline2(longitude latitude)	0.017879	10	0.044508	7.540

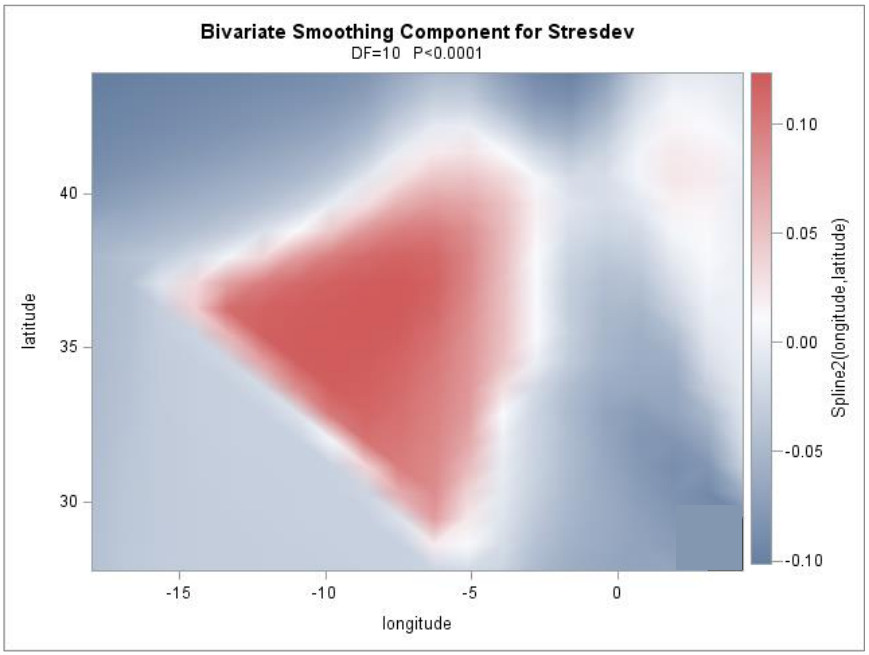
El parámetro de suavizado resultante es 0,017879 y la validación generaliza cruzada es 0,044508. Recordemos que la validación generalizada cruzada es:

$$GCV(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{f}_\lambda(x_i)}{1 - \frac{trS}{n}}\right)^2$$

Por último, se aprecia que aparecen sólo 7.540 observaciones únicas. Esto es debido a que de los 11.000 códigos postales, la compañía solo tiene suficiente exposición en 7.540.

La parte crítica del análisis del output generado con el procedimiento GAM viene con el estudio de la deviance:

Smoothing Model Analysis Analysis of Deviance				
Source	DF	Sum of Squares	Chi-Square	Pr > ChiSq
Spline2(longitude latitude)	10	15.656	351.748,337	<.0001



Para cada efecto de suavizado en el modelo, esta tabla muestra un test χ^2comparativo entre la deviance del modelo final y la deviance del modelo sin la variable de referencia. En este caso, el análisis de los resultados de la deviance indica que el efecto de las geocordenadas (longitud y latitud) es altamente significativo.

Finalmente, se calcularán las relatividades para saber cuánto habrá que recargar la tarifa por código postal. Para ello, utilizaremos el parámetro calculado anteriormente de la siguiente forma:

```
Data relatividades;
      Set union1 nounion1;
      Relatividades=exp (P_Stresdev-(0.08687));*hay que sacar el
      estimador del output "Intercept", este valor es el que buscamos,
      que nos dice cuanto hay que recargar o bonificar la tarifa por
      código postal;
Run;
```

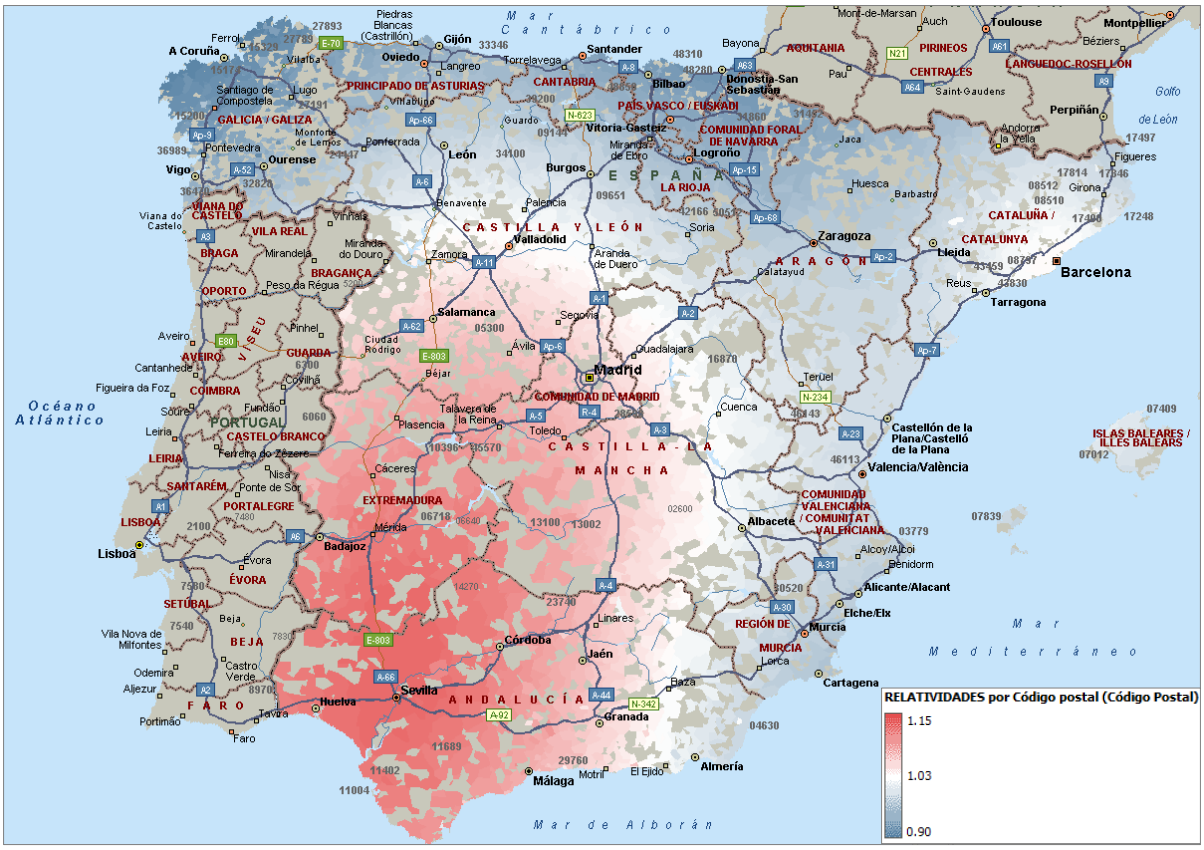
Los resultados finales se detallarán en el siguiente capítulo.

5
Resultados finales del modelo

En el presente capítulo, se expondrán los resultados obtenidos de la modelización de la deviance residual con el modelo GAM.

Se ha utilizado el software MapPoint de Microsoft para representar en un mapa de España las relatividades obtenidas en cada geocordenada (código postal).

El resultado global ha sido el siguiente:



Si observamos la leyenda, podemos comprobar como hay zonas en las que se llegan a aplicar recargos del 15% y zonas en las que se pueden practicar bonificaciones del 10%.

Las zonas en tonos grises son aquellas en las que no hay una exposición relevante. A continuación se realizará un estudio más detallado por regiones.

5.1 Barcelona y alrededores

Barcelona es la ciudad con mayor exposición de España. Podemos comprobar como el comportamiento es bastante homogéneo en toda la región metropolitana de Barcelona, alternándose recargos que oscilan entre el 2% y el 4,5%. El núcleo urbano de Barcelona presenta un recargo de entre un 2,5% y un 2,8%

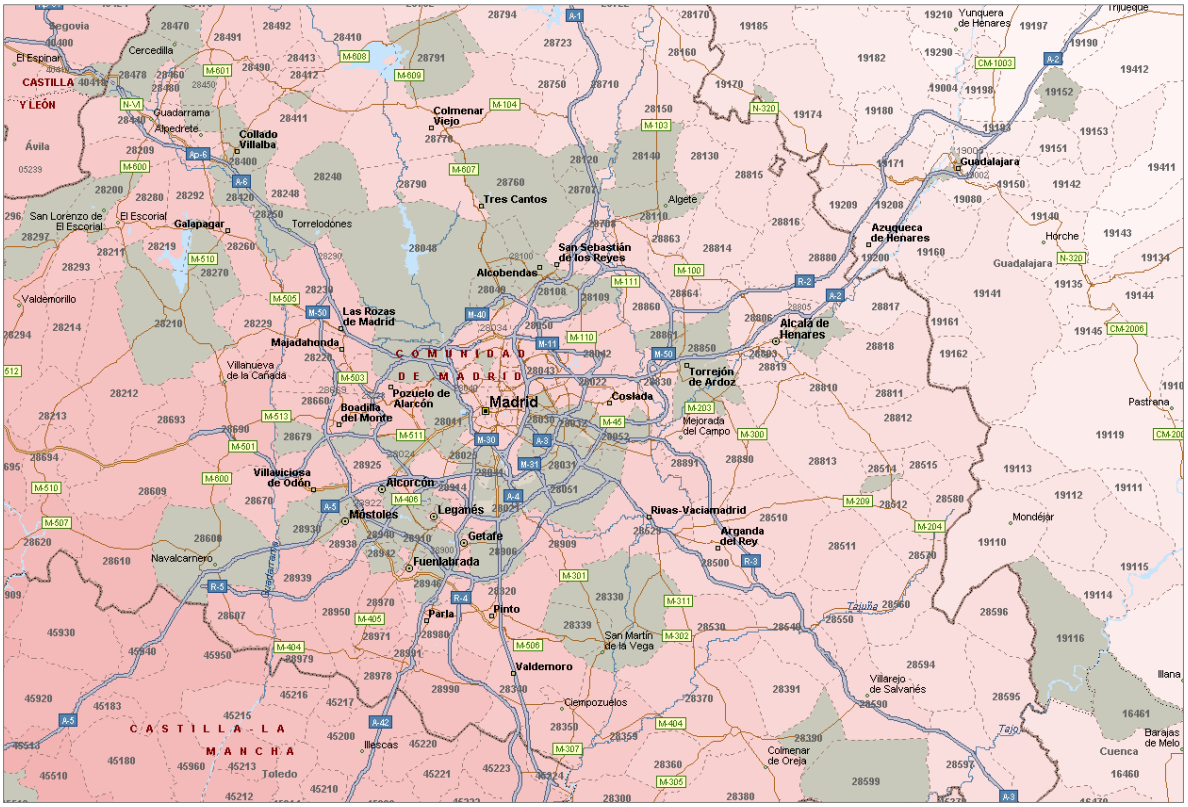
Otro puntos importantes de la región como Sabadell y Terrasa presentan un recargo ligeramente mayor del 4%.

Granollers y Martorrell presentan un comportamiento más próximo al núcleo urbano de Barcelona con un 2,75% y un 3,1% respectivamente.

Otros dos núcleos urbanos importantes de la región de Barcelona son Mataró y Badalona. El primero de ellos presentaría un recargo del 2,5% y el segundo del 3%.

Al resto de la región de Cataluña se aplicaría recargos menores o incluso bonificaciones, como en el caso de Lérida y Gerona donde, en algunos puntos, se bonificaría hasta un 7%.

5.2 Madrid y alrededores



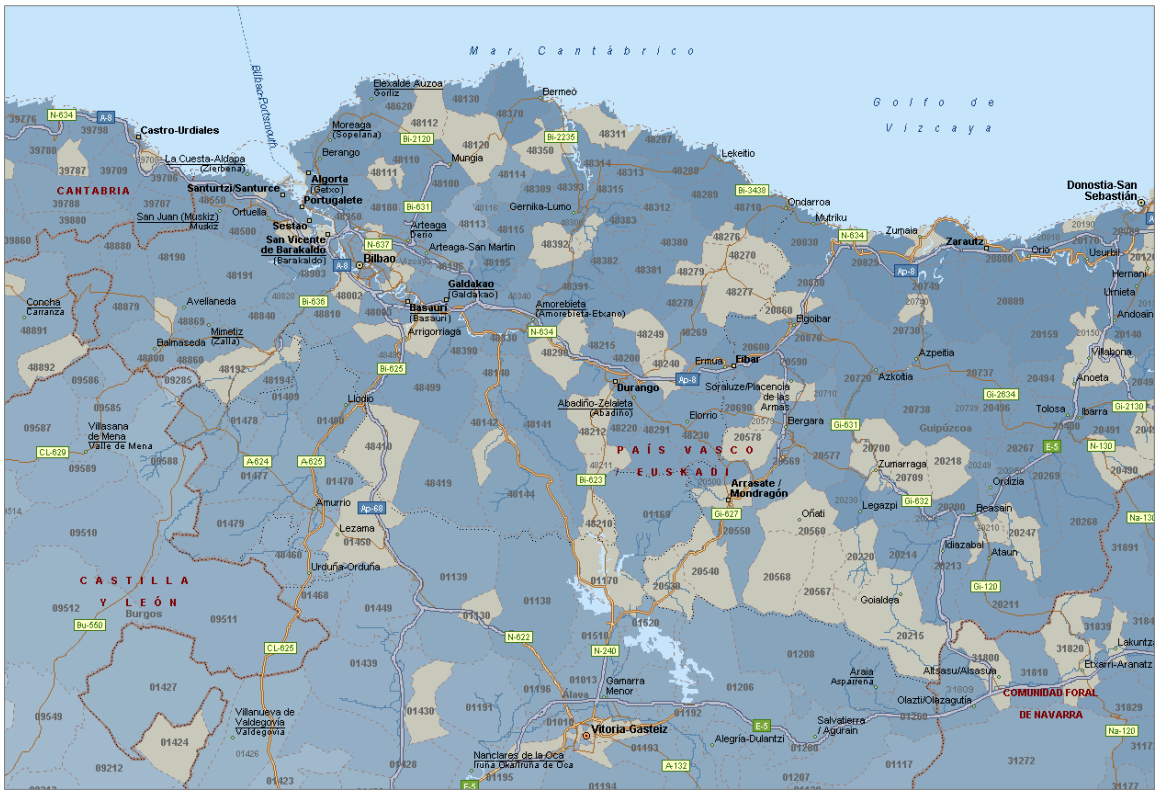
Madrid es la segunda ciudad con mayor exposición. Presenta un peor comportamiento siniestral que Barcelona en cuanto a frecuencia.

Al núcleo urbano de Madrid hay que recargarle entre un 5 y un 6%. Municipios como Alcobendas y San Sebastián de los Reyes presentan una frecuencia un poco más elevada que el núcleo urbano de Madrid, ya que presentan recargos entorno al 6,5%.

Como se puede observar, los municipios más orientales tienen mejor comportamiento, ya que a núcleos como Arganda del Rey, Rivas Vaciamadrid o Villarejo de Salvanés se les aplicaría un recargo que oscilaría entre el 4,5 y el 5%, al igual que provincias limítrofes como es el caso de Guadalajara.

Otros municipios representativos, como es el caso de Majadahonda, Las Rozas, Boadilla del Monte, Móstoles, Valdemoro o Galapagar presentan una casuística similar al núcleo urbano de Madrid, con recargos que oscilan entre el 4,75% y el 5,8%.

5.3 País Vasco y alrededores



Como puede apreciarse en el mapa general de España, el norte peninsular presenta mucho mejor comportamiento que la zona centro-sur de la península.

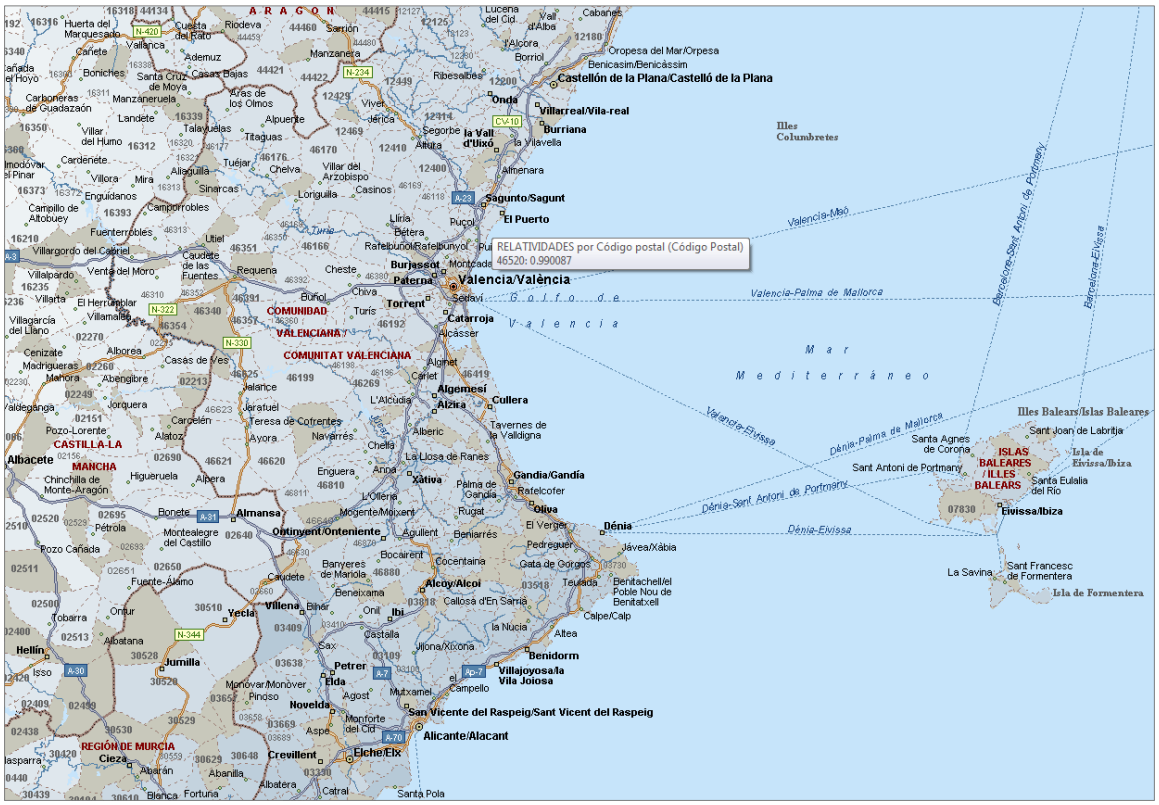
Si nos detenemos en País Vasco, vemos como en los núcleos urbanos de Bilbao y de San Sebastián se aplican bonificaciones que abarcan desde el 6% hasta el 8,5%.

Municipios con bonificaciones similares a Bilbao y San Sebastián, son Portugalete, Durango y Santurce, con un 7,5%, 8% y 7% respectivamente.

Otros municipios importantes como Éibar, Zarauz y Algorta presentan las bonificaciones más elevadas del País Vasco, llegando a alcanzar descuentos del 8-9%.

Finalmente, municipios pertenecientes a otras regiones limítrofes al País Vasco, como por ejemplo Cantabria, presentan también una frecuencia menor que otras regiones, como es el caso de Castro Urdiales, con una bonificación próxima al 9%.

5.4 Levante y Canarias



La zona este peninsular presenta un buen comportamiento en términos de frecuencia. Como se puede comprobar, a núcleos urbanos importantes como el de Valencia, Alicante y Castellón se les aplica bonificaciones del 1%, 3% y 1,5% respectivamente.

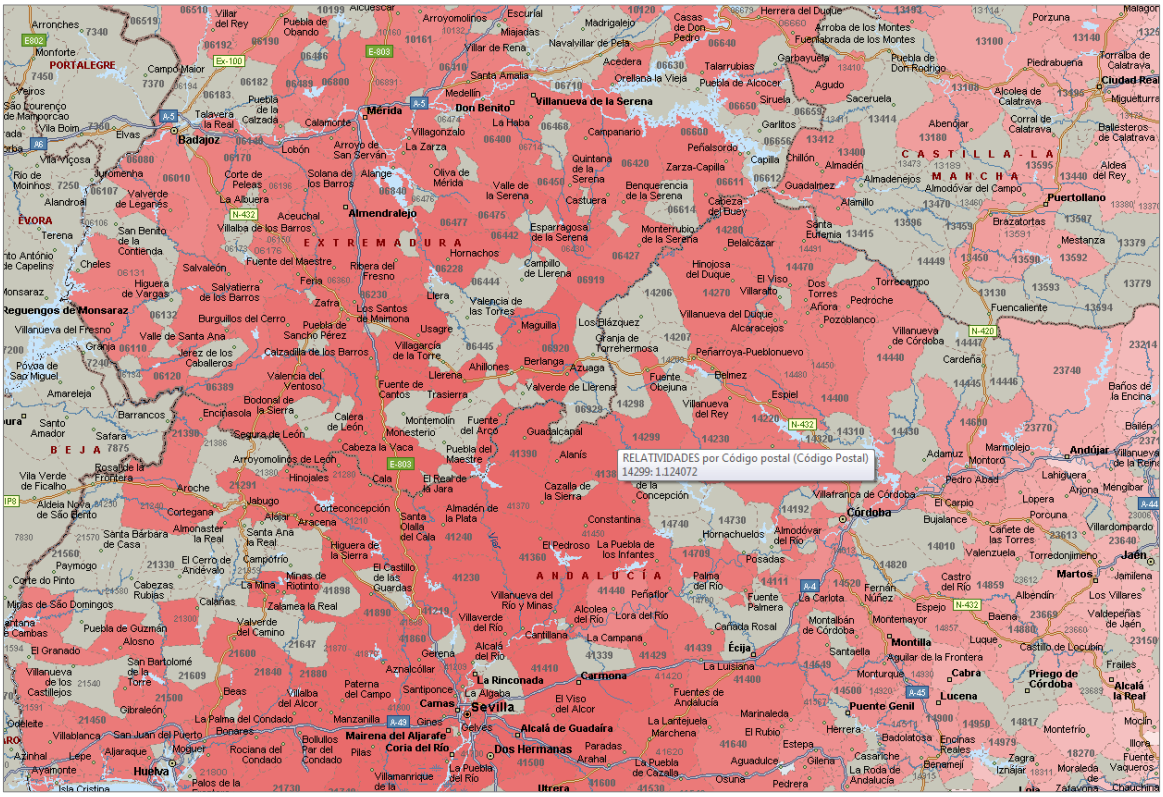
Otros puntos importantes como Villareal, Sagunto, Denia y Alcoy presentarán bonificaciones que oscilarán entre el 1 y el 2,5%.

Sin embargo, no toda la zona recibirá bonificaciones, ya que núcleos como Elche o numerosas zonas de las Islas Baleares sufrirán recargos, a pesar de que nunca superen el 1%.

Los núcleos urbanos pertenecientes a la comunidad murciana se beneficiarán de bonificaciones que variarán entre el 2,3% y el 2,5%.

En las Islas Canarias, hay muy poca exposición. Solamente se aplicarán bonificaciones en Gran Canaria, donde se alcanzarán descuentos de alrededor del 4%.

5.5 Extremadura y Castilla La Mancha



Extremadura es, junto con Andalucía, la comunidad autónoma que presenta peor comportamiento en cuanto a la frecuencia siniestral.

En el núcleo urbano de Cáceres se llegarán a aplicar recargos del 10% e incluso del 12% en Badajoz.

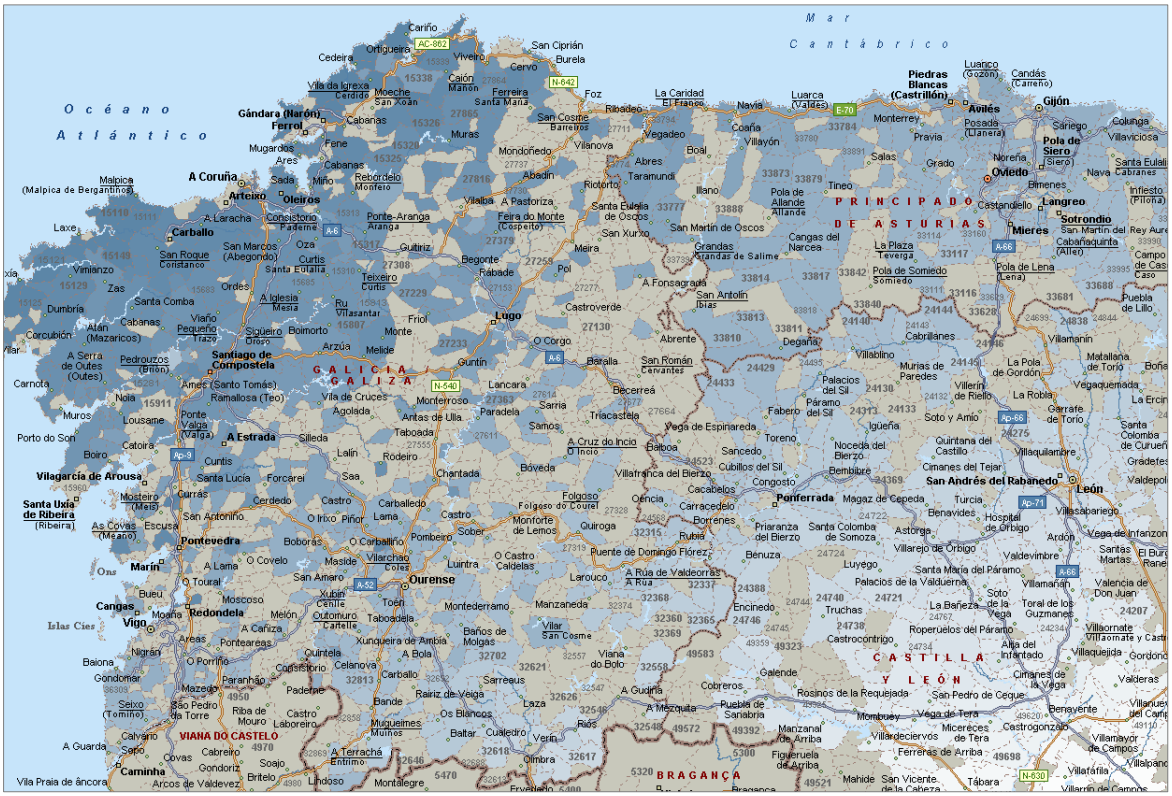
Sin embargo los recargos más elevados se aplicarán en torno a Mérida, donde se llegarán a alcanzar recargos del 12,5%.

Otros núcleos conocidos como Almendralejo, Don Benito y Villanueva de la Serena se comportan de manera similar a la provincia donde se encuentran, Badajoz.

En Castilla La Mancha hay zonas con muy poca exposición. Las zonas que si tienen un número representativo de expuestos presentan un mal comportamiento en la frecuencia, aunque mucho más suave que dos de sus comunidades limítrofes: Extremadura y Andalucía.

A Ciudades importantes como Ciudad Real y Talavera de la Reina, se les aplicarán recargos del 9% y 8% respectivamente, al igual que a Puertollano y al Alcázar de San Juan, con un 9% y un 6%. Sin embargo el este de Castilla La Mancha se comportará mejor, ya que las zonas pertenecientes a Guadalajara, Cuenca y Albacete en ningún caso sufrirán recargos mayores del 4%.

5.6 Galicia y Castilla y León



Como vimos anteriormente al estudiar la frecuencia en el País Vasco, el norte peninsular es la zona donde se practican mayores bonificaciones de toda España, de hecho, Galicia es la comunidad autónoma con mejor comportamiento en la frecuencia siniestral.

En localidades como A Coruña y Santiago de Compostela se aplicarán bonificaciones del 8%-9%. Sin embargo, ciudades como Narón y Malpica presentarán las bonificaciones más elevadas la península, llegando a alcanzar descuentos del 10%.

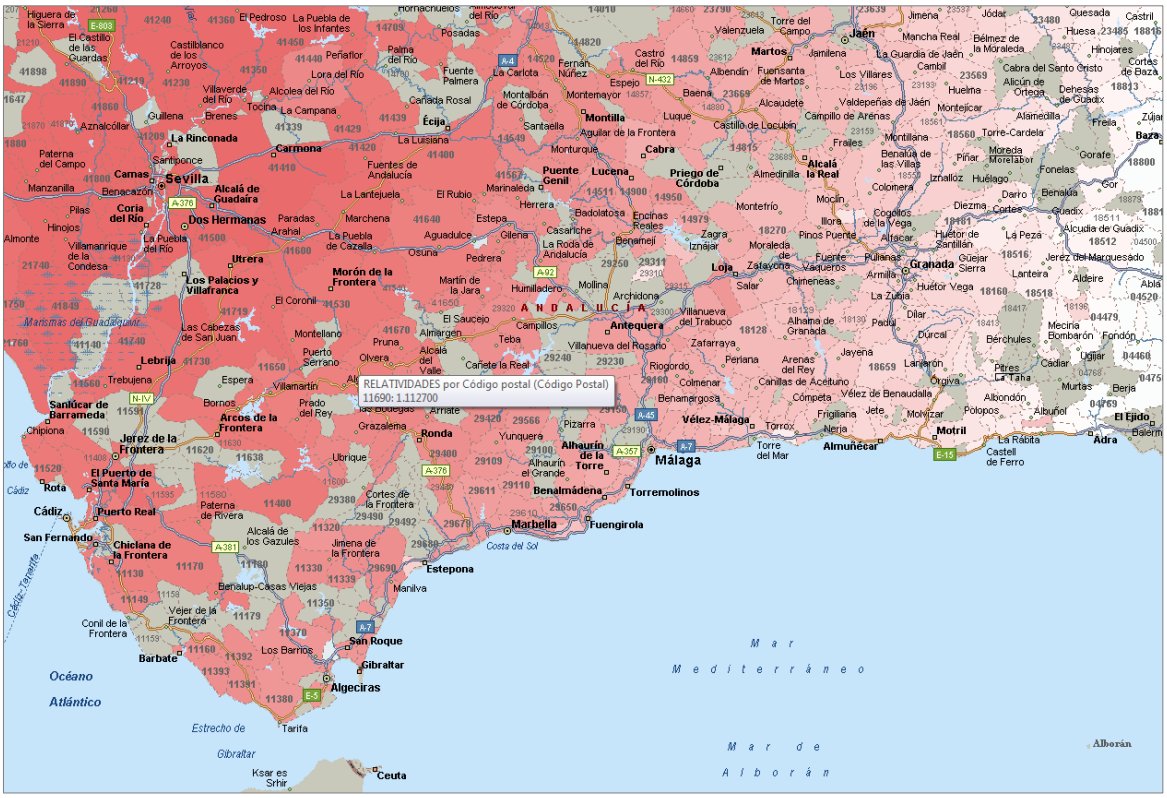
A Oleiros, Vigo y Pontevedra también se les practicarán importantes bonificaciones que oscilarán entre el 8,5 y el 8,75%.

Castilla y León tiene un comportamiento bastante heterogéneo. El sur de la comunidad tiene peor comportamiento, ya que a los núcleos urbanos situados en los alrededores de Salamanca, Ávila y Segovia se les practicarán recargos en torno al 4%-5%.

En León, Palencia, Burgos y Soria se aplicarán bonificaciones nunca mayores del 3,5%.

Por último, a Valladolid y Zamora se les recargará la tarifa un 3% aproximadamente.

5.7 Andalucía



Andalucía es la comunidad autónoma española donde mayores recargos se van a aplicar.

Dentro de Andalucía, Sevilla y Huelva son las provincias que peor se comportan. Concretamente en algunos núcleos urbanos pertenecientes a Sevilla, como Camas, Dos Hermanas o en la propia capital sevillana se aplicarán recargos que podrán llegar a alcanzar el 15%.

Las provincias de Cádiz y Málaga también experimentan muy mal comportamiento a pesar de que no alcancen el nivel de Sevilla y Huelva. El Puerto de Santa María, San Lúcar de Barrameda, Jerez de la Frontera y Arcos de la Frontera sufrirán recargos de entre un 11% y un 12,5%. En cambio, a Málaga, Marbella, o Antequera se les recargará en torno a un 8%-9%.

A medida que nos vayamos aproximando hacia el este de la comunidad, la situación ira mejorando, ya que en zonas de Granada y Almería, como por ejemplo Almuñécar, Motril, Adra o Baza nunca se recargará por encima del 4,5%.

Cuarta parte
Conclusiones

El objetivo fundamental de este estudio era analizar la frecuencia siniestral a nivel de código postal para una cartera del ramo de hogar.

Una parte clave del análisis ha sido el estudio de la variable fundamental sobre la que ha pivotado nuestro análisis: el número de siniestros. Como hemos podido comprobar tras observar su distribución empírica, la variable aleatoria número de siniestros se ajustaba a una binomial negativa y no a una Poisson. Este hecho se da a menudo en el mundo del seguro, ya que la heterogeneidad es algo intrínseco del mismo. Ajustar una distribución de la familia exponencial a nuestra variable aleatoria es algo fundamental como paso previo a modelizar la misma, ya que como se comentó con anterioridad, si utilizamos un modelo GLM para predecir y no ajustásemos una distribución de la familia exponencial, nuestras predicciones no serían fiables, ya que estaríamos violando una de las hipótesis de partida.

Una vez analizada la variable aleatoria que queríamos modelizar, construimos un GLM con las siguientes variables explicativas: superficie, tipo de pago, año de estudio, capital del continente, capital del contenido, tipo-uso-ubicación del inmueble y antigüedad de la póliza. Comprobamos como todas ellas resultaron muy explicativas de la frecuencia siniestral. Analizando los resultados del GLM, pudimos comprobar como éste trabaja mucho mejor en clústers lo suficientemente poblados, ya que el GLM trabaja en medias, con lo cual a mayor exposición, más fiables resultarán nuestras predicciones.

Al observar las relatividades producidas por nuestro GLM, pudimos comprobar como hay determinados perfiles de riesgo en los que se recargaría la tarifa en función de la frecuencia, por ejemplo, solamente por elegir la modalidad de pago fraccionada, se recargaría ésta hasta un 38%.

Finalmente, se asumió que la parte no explicada de nuestro modelo vendría recogida por la localización del riesgo, por ello, se utilizó un modelo aditivo generalizado (GAM) para modelizar la deviance residual estandarizada en función de los splines de las geocordenadas (latitud y longitud). Estudiando los resultados, llegamos a la conclusión de que en el sur peninsular la frecuencia siniestral es mucho mayor que en otras zonas como el norte de España o Levante, debido entre otras cosas, a la forma en la que se construyen los inmuebles y a las calidades de las mismas.

Anexo
Códigos de programación

1. Preparación y tratamiento de la base de datos

```
Libname TFM "C:\TFM\BBDD\";

*Transformamos variables;
Data TFM.BBDD_TFM;
      Set TFM.BBDD_HOGAR;
            If polvig_100=. Then polvig_100=0;
            If polnue_100=. Then polnue_100=0;
            If polanu_100=. Then polanu_100=0;
            If durac_100=.  Then durac_100=0;
            If prima_100=.  Then prima_100=0;
            If carga_100=.  Then carga_100=0;
            If pagos_100=.  Then pagos_100=0;
            If subcg_100=.  Then subcg_100=0;
            If numsin_100=. Then numsin_100=0;
            If numsin_100 > 7 then delete;

            Lnduree=log (durac_100);

            If numsin_100=0 then Cosme=0;
            Else if subcg_100 not in (0,.) then
            Cosme=subcg_100/numsin_100;
            If subcg_100 in (0,.) then PPura=0;
            Else if subcg_100 not in (0,.) then
            PPura=subcg_100/durac_100;

            CAPITAL=CASEDIF+CASMOB;

            *v09= variable de añada;
            If an=2012 then v09=2;*hay que juntar los 2 años xq salen
            mal;

            * antig.poliza=v22;
            v22_b=v22;
            If 4<=v22<=5 then v22_b=4;
            Else if 6<=v22<=10 then v22_b=6;
            Else if v22>10 then v22_b=10;*este me lo tomará como la
            base, que es lo que queremos;
```

```
Select;
  When (CASEDIF<=18000)       v10_a=1;
  When (CASEDIF<=24000)       v10_a=2;
  When (CASEDIF<=36000)       v10_a=3;
  When (CASEDIF<=42000)       v10_a=4;
  When (CASEDIF<=48000)       v10_a=5;
  When (CASEDIF<=60000)       v10_a=6;
  When (CASEDIF<=66000)       v10_a=7;
  When (CASEDIF<=72000)       v10_a=8;
  When (CASEDIF<=90000)       v10_a=9;
  When (CASEDIF<=120000)      v10_a=10;
  When (CASEDIF<=180000)      v10_a=11;
  When (CASEDIF<=300000)      v10_a=12;
  When (CASEDIF>300000)       v10_a=13;
  Otherwise                   v10_a=1;
End;

Select;
  When (capital<=18000)       v10_b=1;
  When (capital<=24000)       v10_b=2;
  When (capital<=36000)       v10_b=3;
  When (capital<=42000)       v10_b=4;
  When (capital<=48000)       v10_b=5;
  When (capital<=60000)       v10_b=6;
  When (capital<=66000)       v10_b=7;
  When (capital<=72000)       v10_b=8;
  When (capital<=90000)       v10_b=9;
  When (capital<=120000)      v10_b=10;
  When (capital<=180000)      v10_b=11;
  When (capital<=300000)      v10_b=12;
  When (capital>300000)       v10_b=13;
  Otherwise                   v10_b=1;
End;

Select;
  When (CASMOB=0)             v10_c=0;
  When (CASMOB<=6000)         v10_c=1;
  When (CASMOB<=12000)        v10_c=2;
  When (CASMOB<=15000)        v10_c=2;
  When (CASMOB<=21000)        v10_c=4;
  When (CASMOB<=24000)        v10_c=4;
  When (CASMOB<=30000)        v10_c=4;
  When (CASMOB<=36000)        v10_c=7;
  When (CASMOB<=42000)        v10_c=8;
  When (CASMOB>42000)         v10_c=8;
  Otherwise                   v10_c=4;*LA MEDIA;
End;

Select;
  When (CASMOB=0)             v10_cc=0;
  Otherwise                   v10_cc=1;
End;
```

```
If v03 in (3,4,5) then v03=3; *Ubicación;

v08_b=0;*mezcla de 3 variables uso-ubicación-tipo;
If v08=99  and v06=99  and v03=99 then v08_b=1;
If v08=99  and v06=2   and v03=99 then v08_b=2;
If v08=2   and v06=99  and v03=99 then v08_b=3;
If v08=99  and v06=3   and v03=99 then v08_b=4;
If v08=4   and v06=99  and v03=99 then v08_b=5;
If v08=3   and v06=99  and v03=99 then v08_b=6;
If v08=5   and v06=99  and v03=3  then v08_b=10;
If v08=2   and v06=2   and v03=99 then v08_b=8;
If v08=3   and v06=2   and v03=99 then v08_b=9;
If v08=5   and v06=99  and v03=99 then v08_b=10;

v04_b=v04;
If v04 IN (5,6,7) then v04_b=5;

v01_b=v01;*Forma_de_pago;
If v01 not IN (99)  then v01_b=2;

*zonificación;

*zona Geográfica: agrupamos las provincias por frec;

Provincia_2=3;
If Provincia=53 then Provincia_2=3;
Else if Provincia=22 Then Provincia_2=1;
Else if Provincia=25 Then Provincia_2=1;
Else if Provincia=42 Then Provincia_2=1;
Else if Provincia=17 Then Provincia_2=1;
Else if Provincia=9  Then Provincia_2=2;
Else if Provincia=20 Then Provincia_2=2;
Else if Provincia=44 Then Provincia_2=2;
Else if Provincia=26 Then Provincia_2=2;
Else if Provincia=31 Then Provincia_2=2;
Else if Provincia=1  Then Provincia_2=2;
Else if Provincia=50 Then Provincia_2=2;
Else if Provincia=15 Then Provincia_2=2;
Else if Provincia=38 Then Provincia_2=2;
Else if Provincia=40 Then Provincia_2=2;
Else if Provincia=24 Then Provincia_2=2;
Else if Provincia=34 Then Provincia_2=2;
Else if Provincia=33 Then Provincia_2=2;
Else if Provincia=27 Then Provincia_2=3;
Else if Provincia=32 Then Provincia_2=3;
Else if Provincia=12 Then Provincia_2=3;
Else if Provincia=7  Then Provincia_2=3;
Else if Provincia=43 Then Provincia_2=3;
Else if Provincia=35 Then Provincia_2=3;
Else if Provincia=39 Then Provincia_2=3;
Else if Provincia=47 Then Provincia_2=3;
Else if Provincia=3  Then Provincia_2=3;
Else if Provincia=5  Then Provincia_2=3;
Else if Provincia=8  Then Provincia_2=4;
Else if Provincia=49 Then Provincia_2=4;
Else if Provincia=30 Then Provincia_2=4;
Else if Provincia=37 Then Provincia_2=4;
Else if Provincia=46 Then Provincia_2=4;
Else if Provincia=48 Then Provincia_2=4;
Else if Provincia=19 Then Provincia_2=4;
Else if Provincia=4  Then Provincia_2=4;
```

```
            Else if Provincia=36 Then Provincia_2=5;
            Else if Provincia=16 Then Provincia_2=5;
            Else if Provincia=28 Then Provincia_2=5;
            Else if Provincia=2  Then Provincia_2=5;
            Else if Provincia=18 Then Provincia_2=5;
            Else if Provincia=29 Then Provincia_2=5;
            Else if Provincia=21 Then Provincia_2=5;
            Else if Provincia=45 Then Provincia_2=5;
            Else if Provincia=13 Then Provincia_2=5;
            Else if Provincia=11 Then Provincia_2=6;
            Else if Provincia=14 Then Provincia_2=6;
            Else if Provincia=10 Then Provincia_2=6;
            Else if Provincia=23 Then Provincia_2=6;
            Else if Provincia=6  Then Provincia_2=6;
            Else if Provincia=41 Then Provincia_2=6;

Run;

Data TFM.BBDD_TFM;
      Set TFM.BBDD_TFM;
            If numsin_100>1 and carga_100=0 then delete;
            If durac_100=0 Then delete;*solo podemos modelizar si hay
            exposición;
            If durac_100=. Then delete;*solo podemos modelizar si hay
            exposición;
            If subcg_100<0 then delete;*cosas raras!;
Run;


Proc Summary nway missing data=TFM.BBDD_TFM;
      Class v01_b v04_b   v09 v10_a v10_c  v08_b  v22_b;
      Var durac_100 numsin_100;
      Output out=TFM.Clases_riesgo sum=;
Run;

Data TFM.Clases_riesgo;
      Set TFM.Clases_riesgo;
        Lnduree=log (durac_100);
Run;
```

2. Test de bondad del ajuste

```
Libname TFM "C:\TFM\BBDD\";

Proc freq data=tfm.bbdd_tfm;
      Tables numsin_100 / out=FreqOut plots=FreqPlot (scale=percent);
Run;

Proc genmod data=tfm.bbdd_tfm;
   Model numsin_100 = / dist=poisson;
   Output out=PoissonFit p=lambda;
Run;

Data _null_;
      Set PoissonFit;
      Call symputx ("Lambda", Lambda);
      Stop;
Run;

Data PMF;
      Do t = 0 to 7; /* 0 to Max(x) */
    Y = pdf ("Poisson", t, &Lambda);
    Output;
    End;
Run;

Data TFM.Discrete;
      Merge FreqOut PMF;
      Prop = Percent / 100; /* convert to same scale as PDF */
Run;

Ods listing close;
Ods html path = "C:\TFM\BBDD\"
Gpath = "C:\TFM\BBDD\png\" (url="png/")
File = "Poisson.htm";
Ods graphics on;

/* Código inicio del user */
Proc sgplot data=tfm.discrete; /* VBARPARM is SAS 9.3 stmt */
      Vbarparm category= numsin_100 response=Prop /
      legendlabel='Sample';
      Scatter x=T y=Y / legendlabel='PMF'
      Markerattrs=GraphDataDefault (symbol=CIRCLEFILLED size=10);
      Title "Numero siniestros y Poisson Distribution";
Run;
/* Código final del user */

Ods graphics off;
Ods html close;
Ods listing;

Proc sort data=tfm.bbdd_tfm; by numsin_100;
Run;    /* 1 */

Data TFM.QQ;
      Set tfm.bbdd_tfm nobs=nobs;
            v = (_N_ - 0.375) / (nobs + 0.25);    /* 2 */
            q = quantile ("Poisson", v, &Lambda); /* 3 */
Run;
```

```
Ods listing close;
Ods html path = "C:\TFM\BBDD\"
Gpath = "C:\TFM\BBDD\png\" (url="png/")
File = "Cuantil.htm";
Ods graphics on;

/* Código inicio del user */
Proc sgplot data=TFM.QQ noautolegend;     /* 4 */
      Scatter x=q y=numsin_100;
      Lineparm x=0 y=0 slope=1; /* SAS 9.3 statement */
      Xaxis label="Poisson Quantiles" grid;
      Yaxis label="Observed Data" grid;
      Title "Poisson Q-Q Plot of siniestros";
Run;
/* Código final del user */

Ods graphics off;
Ods html close;
Ods listing;
```

3. Análisis de los factores de riesgo

```
Libname tfm "C:\TFM\BBDD\";
*Forma de pago;
Data pp;
      Set tfm.bbdd_tfm;
            Format Forma_Pago $12.;
            If v01 = 2  then Forma_Pago = "Semestral";
            If v01 = 3  then Forma_Pago = "Trimestral";
            If v01 = 6  then Forma_Pago = "Mensual";
            If v01 = 9  then Forma_Pago = "Irregular";
            If v01 = 99 then Forma_Pago = "Anual";
            If v01 = 0  then Forma_Pago = "Sin informar";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Forma_Pago / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

Data pp;
      Set tfm.bbdd_tfm;
            Format Forma_Pago $8.;
            If v01_b = 2   then Forma_Pago_Agrup = "Fraccionado";
            If v01_b = 99  then Forma_Pago_Agrup = "Anual";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Forma_Pago_Agrup / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

*Superficie en m2;
Data pp;
      Set tfm.bbdd_tfm;
            Format Superf_m2 $12.;
            If v04 = 1  then Superf_m2 = "Sin informar";
            If v04 = 2  then Superf_m2 = "0-45 m2";
            If v04 = 3  then Superf_m2 = "45-75 m2";
            If v04 = 5  then Superf_m2 = "90-105 m2";
            If v04 = 6  then Superf_m2 = "105-150 m2";
            If v04 = 7  then Superf_m2 = "150-180 m2";
            If v04 = 8  then Superf_m2 = ">180 m2";
            If v04 = 99 then Superf_m2 = "75-90 m2";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Superf_m2 / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;
```

```
Data pp;
      Set tfm.bbdd_tfm;
            Format Superf_m2_agrup $12.;
            If v04_b = 1  then Superf_m2_agrup = "Sin informar";
            If v04_b = 2  then Superf_m2_agrup = "0-45 m2";
            If v04_b = 3  then Superf_m2_agrup = "45-75 m2";
            If v04_b = 5  then Superf_m2_agrup = "90-180 m2";
            If v04_b = 8  then Superf_m2_agrup = ">180 m2";
            If v04_b = 99 then Superf_m2_agrup = "75-90 m2";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Superf_m2_agrup / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

*Año estudio;
PROC FREQ DATA=tfm.BBDD_TFM;
     ODS GRAPHICS ON;
     TABLES an / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

Data pp;
      Set tfm.bbdd_tfm;
            Format ano_estudio $9.;
            If v09 = 1   then ano_estudio = "2010";
            If v09 = 2   then ano_estudio = "2011/2012";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES ano_estudio / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

*Capital asegurado edificio;
Data pp;
      Set tfm.bbdd_tfm;
            Format Cap_edificio $12.;
            If v10_a = 1   then Cap_edificio = "<=18000";
            If v10_a = 2   then Cap_edificio = "<=24000";
            If v10_a = 3   then Cap_edificio = "<=36000";
            If v10_a = 4   then Cap_edificio = "<=42000";
            If v10_a = 5   then Cap_edificio = "<=48000";
            If v10_a = 6   then Cap_edificio = "<=60000";
            If v10_a = 7   then Cap_edificio = "<=66000";
            If v10_a = 8   then Cap_edificio = "<=72000";
            If v10_a = 9   then Cap_edificio = "<=90000";
            If v10_a = 10  then Cap_edificio = "<=120000";
            If v10_a = 11  then Cap_edificio = "<=180000";
            If v10_a = 12  then Cap_edificio = "<=300000";
            If v10_a = 13  then Cap_edificio = ">300000";
Run;
```

```
PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Cap_edificio / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

*Capital asegurado mobiliario;
Data pp;
      Set tfm.bbdd_tfm;
            Format Cap_mobiliario $12.;
            If v10_c = 0  then Cap_mobiliario = "<Sin Capital";
            If v10_c = 1  then Cap_mobiliario = "<=6000";
            If v10_c = 2  then Cap_mobiliario = "<=12000";
            If v10_c = 4  then Cap_mobiliario = "<=24000";
            If v10_c = 7  then Cap_mobiliario = "<=42000";
            If v10_c = 8  then Cap_mobiliario = ">42000";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Cap_mobiliario / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

*Tipo vivienda;
Data pp;
      Set tfm.bbdd_tfm;
            Format Tipo_Inm $20.;
            If v08 = 2  then Tipo_Inm = "Chalet adosado";
            If v08 = 3  then Tipo_Inm = "Chalet independiente";
            If v08 = 4  then Tipo_Inm = "Casa tradicional";
            If v08 = 5  then Tipo_Inm = "Otras viviendas";
            If v08 = 99 then Tipo_Inm = "Piso";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Tipo_Inm / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

*Uso inmueble;
Data pp;
      Set tfm.bbdd_tfm;
            Format Uso_Inm $12.;
            If v06 = 2  then Uso_Inm = "Secundaria";
            If v06 = 3  then Uso_Inm = "Alquilada";
            If v06 = 4  then Uso_Inm = "Desocupada";
            If v06 = 99 then Uso_Inm = "Habitual";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Uso_Inm / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;
```

```
*Ubicación inmueble;
Data pp;
      Set tfm.bbdd_tfm;
            Format Ubicacion_Inm $12.;
            If v03 = 2  then Ubicacion_Inm = "Urbanización";
            If v03 = 3  then Ubicacion_Inm = "Despoblado";
            If v03 = 99 then Ubicacion_Inm = "Casco urbano";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Ubicacion_Inm / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

*Tipo-Uso-Ubicación;
Data pp;
      Set tfm.bbdd_tfm;
            Format Tipo_Uso_Ubic $40.;
            If v08_b = 0  then Tipo_Uso_Ubic = "Resto";
            If v08_b = 1  then Tipo_Uso_Ubic = "Piso-Habitual-Casco
            urbano";
            If v08_b = 2  then Tipo_Uso_Ubic = "Piso-Secundaria-Casco
            urbano";
            If v08_b = 3  then Tipo_Uso_Ubic = "Chalet Adosado-
            Habitual-Casco urbano";
            If v08_b = 4  then Tipo_Uso_Ubic = "Piso-Alquilada-Casco
            urbano";
            If v08_b = 5  then Tipo_Uso_Ubic = "Casa Tradicional-
            Habitual-Casco urbano";
            If v08_b = 6  then Tipo_Uso_Ubic = "Chalet Indep-Habitual-
            Casco urbano";
            If v08_b = 8  then Tipo_Uso_Ubic = "Chalet Ados-Secundaria-
            Casco urbano";
            If v08_b = 9  then Tipo_Uso_Ubic = "Chalet Indep-
            Secundaria-Casco urbano";
            If v08_b = 10 then Tipo_Uso_Ubic = "Otras viviendas-
            Habitual-Casco urbano";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Tipo_Uso_Ubic / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

*Antigüedad póliza;
Data pp;
      Set tfm.bbdd_tfm;
            Format Antiguedad_pol $40.;
            If v22 = 1  then Antiguedad_pol = "NP";
            If v22 = 2  then Antiguedad_pol = "2 Años";
            If v22 = 3  then Antiguedad_pol = "3 Años";
            If v22 = 4  then Antiguedad_pol = "4 Años";
            If v22 = 5  then Antiguedad_pol = "5 Años";
            If v22 = 6  then Antiguedad_pol = "6 Años";
            If v22 = 7  then Antiguedad_pol = "7 Años";
            If v22 = 8  then Antiguedad_pol = "8 Años";
            If v22 = 9  then Antiguedad_pol = "9 Años";
```

```
            If v22 = 10 then Antiguedad_pol = "10 Años";
            If v22 = 11 then Antiguedad_pol = "11 Años";
            If v22 = 12 then Antiguedad_pol = "12 Años";
            If v22 = 13 then Antiguedad_pol = "13 Años";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Antiguedad_pol / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;

Data pp;
      Set tfm.bbdd_tfm;
            Format Antiguedad_pol_agrup $40.;
            If v22_b = 1  then Antiguedad_pol_agrup = "NP";
            If v22_b = 2  then Antiguedad_pol_agrup = "2 Años";
            If v22_b = 3  then Antiguedad_pol_agrup = "3 Años";
            If v22_b = 4  then Antiguedad_pol_agrup = "4-5 Años";
            If v22_b = 6  then Antiguedad_pol_agrup = "6-10 Años";
            If v22_b = 10  then Antiguedad_pol_agrup = "+10 Años";
Run;

PROC FREQ DATA=work.pp;
     ODS GRAPHICS ON;
     TABLES Antiguedad_pol_agrup / missing
     PLOTS = FREQPLOT
     OUT = work.pp;
RUN;
```

4. Cramérs V

```
/*********************************/
/* ASOCIACION ENTRE VARIABLES     */
/*********************************/

/* ESTE PROGRAMA GENERA LA V DE CRAMER PARA CADA DOS COMBINACIONES DE
FACTORES DE RIESGO */
/* Y GENERA UN INFOME Y UN GRAFICO CON TODA LA INFORMACION GENERADA */
/* LA TABLA FINAL SE GRABA EN DESTINO */

%MACRO GENERAR_ASOCIACIONES (FACTORES=, ORIGEN=, DESTINO=);

/*************/
/* CONSTRUIR */
/*************/

      %MACRO CONSTRUIR;
            %LET INDICE=1;
            DATA LISTAVAR;
            FORMAT VARIABLE $20.0;
                  %DO %UNTIL (%SCAN (&FACTORES.,&INDICE.) EQ);
                  %LET MACROVAR=%SCAN (&FACTORES.,&INDICE.);
                  VARIABLE="&MACROVAR.";
                  OUTPUT;
                  %LET INDICE=%EVAL (&INDICE+1);
                  %END;
            RUN;
       %MEND CONSTRUIR;

      %CONSTRUIR

      /*************/
      /* CRUCEVAR */
      /*************/

      PROC SQL; CREATE TABLE CRUCEVAR AS
            SELECT *
            FROM LISTAVAR (RENAME= (VARIABLE=VARA))
            CROSS JOIN LISTAVAR (RENAME= (VARIABLE=VARB))
            WHERE VARA<VARB;
      QUIT;

       DATA CRUCEVAR;
            SET CRUCEVAR;
                  NAME=COMPRESS ('TABLE'||_N_);
                  FORM=COMPRESS ('FORM'||_N_);
                  FORMULA=COMPRESS (VARA||'*'||VARB);
                  CALL SYMPUT (NAME, NAME);
                  CALL SYMPUT (FORM, FORMULA);
      RUN;

      PROC SQL NOPRINT; SELECT COUNT (*) INTO: MAXVAR FROM CRUCEVAR;
      QUIT;
```

```
/**********/
/* CRAMER */
/**********/

%MACRO CRAMER (TABLA=, FORMULA=, NUMERO=);
      PROC FREQ DATA=&ORIGEN. NOPRINT;
            TABLES &FORMULA. / CHISQ;
            OUTPUT OUT=&TABLA. (RENAME= (_CRAMV_=CRAMV)) CRAMV;
      RUN;
      DATA &TABLA.;
            SET &TABLA.;
            /* PONER %GLOBAL PARA QUE CALL SYMPUT GENERE
            MACROVARIABLES GLOBALES */
            /* CONSTRUIR LA LLAMADA AL CALL SYMPUT USANDO
            VARIABLES DE TABLA */
            /* UN %PUT DENTRO DEL PASO DATA NUNCA VA A
            FUNCIONARNOS */
                  %GLOBAL CRAMV&NUMERO;
                  NOMBRE=COMPRESS ('CRAMV'||&NUMERO.);
                  CALL SYMPUT (NOMBRE, CRAMV);
      RUN;
      PROC SQL; DROP TABLE &TABLA.;
%MEND CRAMER;

%MACRO LANZAR;
       %DO I=1 %TO &MAXVAR.;
            %CRAMER (TABLA=&&TABLE&I., FORMULA=&&FORM&I.,
            NUMERO=&I.)
       %END;
%MEND LANZAR;

%LANZAR

/***************/
/* TABLA FINAL */
/***************/

%MACRO FINAL;
      DATA &DESTINO.;
            %DO I=1 %TO &MAXVAR.;
                  TABLA=SYMGET ("TABLE&I");
                  FORMULA=SYMGET ("FORM&I");
                  VARUNO=SCAN (FORMULA,1,'*');
                  VARDOS=SCAN (FORMULA,2,'*');
                  CRAMV=INPUT (SYMGET("CRAMV&I"),20.0);
                  OUTPUT;
            %END;
            FORMAT CRAMV COMMAX12.4;
      RUN;
%MEND FINAL;

%FINAL
```

```
      /************/
      /* INFORMES */
      /************/

      TITLE "ASOCIACION ENTRE VARIABLES";
      OPTIONS MISSING=-;

      PROC REPORT DATA=&DESTINO.;
                  COLUMN VARDOS VARUNO, CRAMV;
                  DEFINE VARDOS / GROUP 'INICIAL';
                  DEFINE VARUNO / ACROSS 'FINAL';
                  DEFINE CRAMV  / ANALYSIS FORMAT=COMMAX12.2 CENTER '';
      RUN;

      OPTIONS MISSING=.;
      TITLE "GRAFICO GENERAL DE ASOCIACIONES";

      PROC GTILE DATA=&DESTINO.;
            FLOW CRAMV TILEBY=(VARUNO,VARDOS) /     MINLEGENDVALUE=0
            MAXLEGENDVALUE=1;
            RUN;
      QUIT;

%MEND GENERAR_ASOCIACIONES;

/*MODELO DE FRECUENCIA TFM: FACTORES_EXPO */

%GENERAR_ASOCIACIONES
      (FACTORES = v01_b v04_b v09 v10_a v10_c v08_b v22_b,
       ORIGEN =tfm.BBDD_TFM,
       DESTINO =tfm.cramer)
```

5. Construcción del GLM

```
ODS listing exclude obstats parameterestimates;
ODS output
      ObStats=TFM.obstats
      Modelfit=TFM.modfit
      Modelinfo=TFM.modinfo
      Type1=TFM.type1
      Type3=TFM.type3
      Parameterestimates=TFM.parmest;
title1 "MODELO GLM - TFM";
Ods html path  = "C:\TFM\BBDD\"
File  = "Modelo.htm";

*Modelizamos la frecuencia con un Modelo GLM;
Proc GENMOD data=TFM.Clases_riesgo ORDER=formatted;
       Class v01_b v04_b   v09 v10_a v10_c  v08_b  v22_b;
       Model numsin_100=v01_b v04_b v09 v10_a v10_c v08_b v22_b
       /dist=poisson
       Link=log
       Offset=lnduree
       Obstats
       Dscale
       type1
       type3
       Maxit=500
       Residuals;
 Run;

ODS output close;

*Estimación de las relatividades;
Data TFM.parmest;
      Set TFM.parmest;
            If stderr <> 0 then tval=estimate/stderr;
            Rel=round(exp(estimate),0.0001);
            Pval=probchisq;
            Format pval percent7.1;
Run;

Option nodate;
Proc print data=TFM.parmest noobs uniform;
      By parameter notsorted;
      Id parameter;
Run;

Data TFM.grafi_fq;
      Set TFM.Clases_riesgo (keep=lnduree durac_100);
      Set TFM.obstats;
            Est=pred/durac_100;           /* frecuencia estimada  */
            Fr=numsin_100/exp(lnduree);   /* frecuencia observada */
Run;


Proc Summary data=tfm.grafi_fq nway missing;
      Class v01_b v04_b   v09 v10_a v10_c  v08_b  v22_b;
      Var    durac_100  pred numsin_100 est fr;
      Output out=tfm.ver sum=;
Run;
```

```
%global KAO IMP;
%let font=swiss;
%let KAO=WIN;
%let IMP=PSLA4;

%macro GRAFICOS;
    Title '';
    Footnote '';
      Goptions reset=global ctext=black ftext=swiss reset=(all)
      norotate
      Hpos=0 vpos=0 gsfmode=replace target=&IMP device = &KAO
      Htitle=4 pct htext=1.5 pct nodisplay;
      title1   height=3 pct 'Deviance Contra Frecuencia Estimada';
      Footnote height=3 pct 'La deviance depende de la forma
      funcional';
      axis1 label=('Frecuencia Estimada');
      axis2 label=(angle=-90 rotate=90 'Deviance Residual
      Estandarizada');
   Proc gplot data=TFM.Grafi_fq gout=grafcat;
          Plot stresdev*est/
               Haxis=axis1
               Vaxis=axis2
               Vref=0
               Frame;
   Run;

   Data junk;
      Set TFM.Grafi_fq;
            If stresdev=. Then delete;
   Run;
   Proc sort data=work.junk;
       By stresdev;
   Run;
   Proc univariate noprint data=work.junk;
      Var stresdev;
      Output out=work.stats n=nobs median=median qrange=hspr
      mean=mean std=std;
   Run;

   Data work.quantil;
      Set work.junk;
            If _N_=1 then set stats;
            I+1;
            p=(i-.5)/nobs;
            z=Probit(p);
            Sigma=hspr/1.349;
            Normal=median+z*sigma;
            se=(sigma/((1/sqrt(2*3.1415926))*exp(-
            (z**2)/2)))*sqrt(p*(1-p)/nobs);
            Lower=normal-2*se;
            Upper=normal+2*se;
            Resid=stresdev-normal;
            Label z='Quantile Normale' resid='Grafico de la Normal';
    Run;
```

```
   Goption reset=(footnote);
   axis2 label=('');
   Proc gplot data=quantil GOUT=grafcat;
      Plot stresdev*z=1
      Normal*z=2
      Lower*z=3
      Upper*z=3
      /overlay frame hminor=1 vminor=1 vaxis=axis1 haxis=axis2
      Name='grqq';
      symbol1 i=none h=1.1 v=- color=black;
      symbol2 i=join l=3   v=none color=blue;
      symbol3 i=join l=20  v=none color=green;
      axis1 label=(h=1.5 a=90 r=0) value=(h=1.3);
      axis1 label=('') value=(h=1.3);
      Title 'Quantili Normale e Dev. Std. ';
    Run;

   title1 height=3 pct 'Distribución de la Deviance';
   Footnote '';
   axis1 label=('Deviance Residual Estandarizada');
   axis2 label=('Frecuencia');
   Proc gchart data=TFM.Grafi_fq gout=grafcat;
        Format stresdev comma5.1;
        Vbar stresdev /
        Maxis=axis1
        Raxis=axis2;
   Run;

   Goptions target=&IMP device = &KAO display;
   Proc greplay igout=grafcat
      Tc=tempcat
      Nofs;
      /* Define plantilla */
      Tdef newtwo des='Dos cuadrados de igual tamaño'
      /* Define los dos paneles */
      1/llx=0  lly=0  ulx=0  uly=50  urx=100  ury=50  lrx=100  lry=0
      color=black
      2/llx=0  lly=50 ulx=0  uly=100 urx=100  ury=100 lrx=100  lry=50
      color=black;
      /* Asigna la plantilla definida */
      Template newtwo;
      /* Pone los cuatro gráficos en la plantilla definida */
      Treplay 1:grqq
              2:gchart;
      Treplay 1:gplot
              2:gchart;
      Run;
    Quit;
    Goptions reset=(all);

%mend GRAFICOS;
%graficos;
```

6. Construcción del GAM

```
Data Base (keep= NACTU v01_b v04_b v09 v10_a v10_c  v08_b v22_b
DPRIESGO numsin_100 durac_100 lnduree);
      Set tfm.bbdd_tfm;
Run;

Data Obsgam (Keep= Observation v01_b v04_b v09 v10_a v10_c  v08_b
v22_b Pred Stresdev);
      Set tfm.Obstats;
Run;

Data obsgam_o;
      Set Obsgam;
      Format v01_b2 v04_b2 v092 v10_a2 v10_c2 v08_b2 v22_b2 best12.;
            v01_b2=input(v01_b,2.);
            v04_b2=input(v04_b,2.);
            v092=input(v09,2.);
            v10_a2=input(v10_a,2.);
            v10_c2=input(v10_c,2.);
            v08_b2=input(v08_b,2.);
            v22_b2=input(v22_b,2.);
      Drop v01_b v04_b v09 v10_a v10_c  v08_b v22_b;
      Rename
            v01_b2=v01_b
            v04_b2=v04_b
            v092=v09
            v10_a2=v10_a
            v10_c2=v10_c
            v08_b2=v08_b
            v22_b2=v22_b;
Run;

Proc sort data=Base out=Base_o;
      By v01_b v04_b v09 v10_a v10_c  v08_b v22_b;
Run;

Proc sort data=Obsgam_o nodupkey;
      By v01_b v04_b v09 v10_a v10_c  v08_b v22_b;
Run;

Data GAM;
      Merge Base_o (in=a) Obsgam_o (in=b);
      By v01_b v04_b v09 v10_a v10_c  v08_b v22_b;
      If a;
Run;

Data GAM_BBDD (drop=observation);
      Set GAM;
Run;

Data GAM_BBDD;
      Set GAM_BBDD;
            Postalcode=DPRIESGO;
Run;
```

```
Data coordenadascp (keep=postalcode latitude longitude);
      Set coordenadascp;
Run;

Proc sort data= coordenadascp;
    By postalcode;
Run;

Proc sort data=GAM_BBDD;
    By postalcode;
Run;

Data union1 nounion1 FILE1;
      Merge GAM_BBDD (in=a) tfm.coordenadascp (IN=b);
      By postalcode;
      If a=1 AND b=1 THEN OUTPUT union1;
      ELSE IF a=1 AND b=0 THEN OUTPUT nounion1;
      ELSE IF a=0 AND b=1 THEN OUTPUT FILE1;
Run;

Proc means data=union1 nway;
      Class latitude longitude;
      Var Stresdev;
      Output out=wiltfr mean=;
Run;

Data tfm.wiltfr2;
      Set wiltfr;
            Where _freq_>10;
Run;

Ods listing close;
Ods html path  = "/CARGAS/TMP/EBT0176/"
Gpath = "/CARGAS/TMP/EBT0176/HTML/PNG" (url="png/")
File  = "wiltfr2.htm";
Ods graphics on;

Proc gam data=tfm.wiltfr2     PLOTS=(ALL) plots=components(additive);
      Model Stresdev=SPLINE2(longitude,latitude,DF=20)/ dist=gaussian;
      Freq _freq_;
      Output out=out2    ALL;
      Run;
Quit;

Ods graphics off;
Ods html close;
Ods listing;

Proc SQL;
      CREATE TABLE  postal1 AS
      SELECT max (postalcode) as postalcode, latitude as latitude,
      longitude as longitude
      FROM   coordenadascp
      GROUP BY latitude, longitude;
Quit;
```

```
Proc sort data=out2;
    By latitude longitude;
Run;

Proc sort data=postal1;
    By latitude longitude;
Run;

Data union1 nounion1 FILE1;
      Merge out2 (in=a) postal1 (IN=b);
      By latitude longitude;
      If a=1 AND b=1 THEN OUTPUT union1;
      ELSE IF a=1 AND b=0 THEN OUTPUT nounion1;
      ELSE IF a=0 AND b=1 THEN OUTPUT FILE1;
Run;

*proc gam--> y = Intercept + spline(x1, x2);
Data out2;
      Set union1 nounion1;
            Relatividades = exp (P_Stresdev-(-0.3094));*hay que sacar
            el estimador del output "Intercept", este valor es el que
            buscamos, que nos dice cuanto hay que cargar o recargar la
            tarifa por código postal;
Run;
```

Referencias bibliográficas

[1] Yiu-Kuen Tse. *Nonlife Actuarial Models. Theory, Methods and Evaluation*. Cambridge. International Series On Actuarial Science.

[2] Francisco Javier Martín-Pliego y Luis Ruiz-Maya Pérez. *Fundamentos de Probabilidad*. Paraninfo. Tercera edición.

[3] Miguel Usábel. *Advanced Nonlife Contingencies. The Basic Variables. The Number of Claims*. Universidad Carlos III de Madrid. Actuarial Science & Finance program.

[4] Luis Cayuela. *Modelos Lineales Generalizados (GLM)*. Universidad de Granada.

[5] Montserrat Guillén y Catalina Bolancé. *Modelos Lineales Generalizados en seguros*. Universidad de Barcelona.

[6] P.McCullagh y J.A.Nelder. *Generalized Linear Models*. Second Edition (Chapman & Hall/CRC Monographs on Statistics & Applied Probability).

[7] Nelder, J.A. and R.W.M. Wedderburn. 1972. *Generalized Linear Models*. Journal of the Royal Statistical Society, Series A 135:370--84.

[8] Hastie, Trevor J. y Daryl Pregibon. 1992. *Generalized Linear Models*. Statistical Models in S, ed. John M. Chambers and Trevor J. Hastie. Pacific Grove, California: Wadsworth and Brooks/Cole.

[9] Lindsey, James K. 1997. *Applying Generalized Linear Models*. New York: Springer.

[10] Huimin Liu. *Generalized Additive Model*. Department of Mathematics and Statistics University of Minnesota Duluth, Duluth, MN 55812. December 2008.

[11] Wood, Simon N. *Generalized Additive Model: an introduction with R*, Chapman and Hall/CRC

[12] Hasties, T.J. and Tibshirani, R.J. (1990), *Generalized Additive Models*, New York: Chapman and Hall.

[13] *SAS Customer Support Knowledge Base and Community*