FACULTAD DE CIENCIAS ECONÓMICAS Y EMPRESARIALES

MÁSTER EN CIENCIAS ACTUARIALES Y FINANCIERAS

TRABAJO FIN DE MÁSTER

Cálculo de la Prima Pura en un Seguro de Automóvil para la Garantía de Daños Propios, mediante Modelos Lineales Generalizados y Segmentación de Clientes por Conglomerados

AUTOR: Maria Manuela Moura e Moura

TUTOR: María Pérez Martín

CURSO ACADÉMICO: 2016/2017

CONVOCATORIA: Septiembre

Índice

Índice de tablas

Índice de Figuras

Índice de Gráficos

1. Introducción

1.1. Tarificación

El proceso de tarificación juega un papel fundamental en el ámbito de los seguros, pues su objetivo principal consiste en realizar el cálculo de primas de forma equitativa y suficiente, teniendo en cuenta el riesgo que incorpore la póliza. Estos cálculos los efectúa el actuario como profesional responsable, a partir de unas bases técnicas establecidas con información genérica y estadística del riesgo correspondiente.

Desde la perspectiva actuarial, se diferencian dos sistemas de tarificación (Boj, Claramunt, Fortiana y Vegas; 2005):

- *Tarificación a priori o clase rating*: este método establece el importe de la prima sin tener en cuenta el histórico de siniestralidad propiamente dicho de la cartera, sino basándose en características de la misma, prediciendo una siniestralidad esperada y a continuación fijando la respectiva prima. Se recurren a datos de un período anterior, relativos al número de siniestros y sus cuantías, y a las características de los asegurados o tomadores de la póliza. Por otro lado, se parte del supuesto de que las cuantías son independientes y siguen una misma distribución, así como de independencia entre el número de siniestros y el coste por siniestro. Asumiendo esto, se calcula la prima pura como el producto de la esperanza del número de siniestros y la esperanza de las cuantías de los siniestros.

  Se suele recurrir a la creación de grupos de riesgo relativamente homogéneos estableciendo una tarifa acorde a cada uno.

- *Tarificación a posteriori o experience-rating*: este sistema, en cambio, parte de los valores de las primas vigentes y los va actualizando en los siguientes períodos conforme la información que va disponiendo de las pólizas individuales o colectivas. Se basa en que en cada grupo de un tipo de riesgo es heterogéneo pues existen factores tanto conocidos como desconocidos que no se tienen en cuenta e incluso debido a una

incorrecta agrupación de los asegurados. Este método, al tener en cuenta la siniestralidad de cada póliza, recoge dicha heterogeneidad y añade la evolución de los riesgos aplicando bonificaciones o penalizaciones según los resultados obtenidos.

1.2. Modelos Estocásticos de Tarificación: el Seguro de Automóvil

En lo que respecta al seguro de automóviles, actualmente este sigue siendo el más representativo entre los seguros “no vida” con un peso del 32,3% en el año 2016 (Willis Towers Watson, 2017). Las técnicas de tarificación en este ramo son bastante variadas, la elección depende de las preferencias y de la información que disponga la compañía.

En este trabajo, se llevará a cabo un proceso de tarificación a priori de la garantía de daños propios, utilizando los datos del año 1999 procedente de SAS Enterprise-Miner.

En primer lugar, se efectúa un análisis exploratorio de la base de datos, comentando las principales características de las variables cuantitativas y de las cualitativas. A continuación, se crean ocho grupos a través de un análisis clúster bietápico utilizando como variables discriminatorias el nivel de ingresos (INCOME), estudios (EDUCATION) y zona por la que circula el conductor del automóvil (URBANICITY).

Posteriormente, se explican detalladamente los modelos Inflados de Ceros Poisson y Binomial Negativa, y los modelos Hurdle Poisson y Binomial Negativa, modelos contadores utilizados comúnmente en los seguros de no vida. Se aplican dichos modelos a la base de datos y se comprueba para cada clúster, cuál de ellos ajusta mejor la frecuencia de siniestros (CLAIM_FREQ). Una vez modelizada la frecuencia de los accidentes, se multiplica por la esperanza de las cuantías (OLDCLAIM) obteniendo así la Siniestralidad Total, y por tanto la prima pura correspondiente a cada grupo.

Por último, se exponen las conclusiones y se indican las posibles líneas de investigación futuras acorde a los resultados obtenidos.

2. Base de Datos

En el estudio se utilizará la base de datos de Seguros de Coches procedente de SAS Enterprise-Miner versión 7.1. Los datos se corresponden con pólizas del seguro de automóvil referentes al año 1999 en Estados Unidos. Proporcionan información sobre las características de los asegurados, así como el coste de los siniestros y las cuantías referentes a los mismos en los anteriores cinco años.

Se dispone de 10.302 observaciones y 27 variables relativas a las mismas. La información referida a cada póliza se presenta tanto de forma cuantitativa como cualitativa:

Tabla 1: Descripción de las variables de la base de datos

	Variable	Descripción
1	ID	Identificación
2	KIDSDRIV	Número de niños transportados
3	BIRTH	Fecha Nacimiento
4	AGE	Edad
5	HOMEKIDS	Número de Hijos
6	YOJ	Número de años en el trabajo
7	INCOME	Ingresos
8	PARENT1	Padres Solteros
9	HOME_VAL	Ingresos Hogar
10	MSTATUS	Estado Civil
11	GENDER	Sexo
12	EDUCATION	Nivel Máximo de Educación
13	OCCUPATION	Empleo conductor
14	TRAVTIME	Distancia al trabajo (minutos)
15	CAR_USE	Tipo de uso del vehículo
16	BLUEBOOK	Valor del vehículo
17	TIF	Tiempo en vigor
18	CAR_TYPE	Tipo de coche
19	RED_CAR	Coche rojo
20	OLDCLAIM	Cuantía siniestros (últimos 5 años)
21	CLM_FREQ	Número siniestros (últimos 5 años)
22	REVOKED	Licencia revocada (últimos 7 años)
23	MVR_PTS	Puntos registrados del motor del vehículo
24	CLM_AMT	Cuantía siniestros (últimos 5 años)
25	CAR_AGE	Edad vehículo

26	CLAIM_FLAG	Indicador de siniestros
27	URBANICITY	Tipo de zona donde se encuentra la vivienda/trabajo

Fuente: elaboración propia

2.1. Variables Cuantitativas

Cada variable se encuentra expresada en una unidad de medida distinta, por ello, para evaluar su variabilidad es necesario recurrir a medidas relativas como el coeficiente de variación de Pearson. Este, se obtiene mediante el cociente entre la desviación típica y la media aritmética, expresándose de forma habitual en porcentaje (Gorgas, Cardiel y Zamorano, 2009). Al analizar los valores obtenidos, se observa que CLM_AMT seguido de KIDSDRIV y OLDCLAIM, son las variables que presentan mayor dispersión.

La moda de la mayoría de las variables relacionadas con siniestros es cero, lo que indica la existencia de un problema de inflado de ceros, que se comentará más adelante.

Tabla 2: Análisis variables cuantitativas

	N		Media	Mediana	Moda	Desviación estándar	Varianza	Rango	Mínimo	Máximo	Coeficiente de Variación
	Válidos	Perdidos									
AGE	10.295	7	44,84	45	46	8,61	74,07	65	16	81	19%
BLUEBOOK	10.302	0	15.659,92	14.400,00	1.500,00	8.428,77	71.044.083,88	68.240,00	1.500,00	69.740,00	54%
CAR_AGE	9.663	639	8,3	8	1	5,71	32,65	31	-3	28	69%
CLM_AMT	10.302	0	1.511,27	0	0	4.725,25	22.328.014,35	123.247,12	0	123.247,12	313%
CLM_FREQ	10.302	0	0,8	0	0	1,15	1,33	5	0	5	144%
HOME_VAL	9.727	575	154.523,02	160.661,40	0	129.188,44	16.689.653.684,33	885.282,35	0	885.282,35	84%
HOMEKIDS	10.302	0	0,72	0	0	1,12	1,25	5	0	5	155%
INCOME	9.732	570	61.572,08	53.529,28	0	47.457,21	2.252.186.471,05	367.030,26	0	367.030,26	77%
KIDSDRIV	10.302	0	0,17	0	0	0,51	0,26	4	0	4	299%
MVR_PTS	10.302	0	1,71	1	0	2,16	4,66	13	0	13	126%
OLDCLAIM	10.302	0	4.033,98	0	0	8.733,14	76.267.788,73	57.037,00	0	57.037,00	216%
TIF	10.302	0	5,33	4	1	4,11	16,9	24	1	25	77%
TRAVTIME	10.302	0	33,42	32,81	5	15,86	251,69	137,12	5	142,12	47%
YOJ	9.754	548	10,47	11	12	4,11	16,88	23	0	23	39%

Fuente: elaboración propia

2.2. Variables Cualitativas

En lo relativo a las variables categóricas, como se puede observar en las tablas número 3, cabe resaltar que la mayoría de los individuos posee un nivel de estudios de enseñanza obligatoria o carrera universitaria, utilizan su coche para uso privado y en un entorno urbano. Los tipos de coche más comunes son el todoterreno y el monovolumen, siendo la profesión más habitual BLUE COLLAR, trabajadores de jerarquía inferior en las empresas.

Tablas 3: Análisis variables cualitativas

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
CLAIM_FLAG	0	7.556	73	73	73
	1	2.746	27	27	100
	Total	10.302	100	100	
CAR_USE	Commercial	3.789	37	37	37
	Private	6.513	63	63	100
	Total	10.302	100	100	
GENDER	M	4.757	46	46	46
	z_F	5.545	54	54	100
	Total	10.302	100	100	
URBANICITY	Highly Urban/ Urban	8.230	80	80	80
	z_Highly Rural/ Rural	2.072	20	20	100
	Total	10.302	100	100	

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
MSTATUS	z_No	4.114	40	40	100
	Yes	6.188	60	60	60
	Total	10.302	100	100	
PARENT1	No	8.959	87	87	87
	Yes	1.343	13	13	100
	Total	10.302	100	100	
RED_CAR	No	7.326	71	71	71
	Yes	2.976	29	29	100
	Total	10.302	100	100	
REVOKED	No	9.041	88	88	88
	Yes	1.261	12	12	100
	Total	10.302	100	100	

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
OCCUPATION		665	6	6	6
	Clerical	1.590	15	15	22
	Doctor	321	3	3	25
	Home Maker	843	8	8	33
	Lawyer	1.031	10	10	43
	Manager	1.257	12	12	55
	Professional	1.408	14	14	69
	Student	899	9	9	78
	z_Blue Collar	2.288	22	22	100
	Total	10.302	100	100	

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
CAR_TYPE	Minivan	2.694	26	26	26
	Panel Truck	853	8	8	34
	Pickup	1.772	17	17	52
	Sports Car	1.179	11	11	63
	Van	921	9	9	72
	z_SUV	2.883	28	28	100
	Total	10.302	100	100	

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
EDUCATION	<High School	1.515	15	15	15
	Bachelors	2.823	27	27	42
	Masters	2.078	20	20	62
	PhD	934	9	9	71
	z_High School	2.952	29	29	100
	Total	10.302	100	100	

Fuente: elaboración propia

2.3 Adecuación y división de la Base de Datos

Al llevar a cabo un análisis estadístico, la existencia de valores perdidos constituye un inconveniente a la hora de trabajar con bases de datos o muestras. Según Heitjan y Rubin (1991), el conjunto de *missing values* está formado por observaciones con características especiales, como datos agrupados, datos agregados, redondeados, censurados o truncados. Esto, deriva de la ausencia de respuestas en las encuestas (Medina y Galván, 2007), ya sea debido a que los encuestados no están dispuestos a revelar información sensible, no entienden lo que les preguntan, etc., lo que lleva a que contesten de forma parcial. Por ello, el porcentaje de datos omitidos es un factor relevante a tener en cuenta.

Para solventar este problema, se puede optar por la eliminación de casos con datos omitidos o sustituirlos por una estimación, siendo este último el método más habitual. La estimación se puede realizar por métodos de imputación simple como la media o de forma deductiva con procedimientos hot-deck (duplicando un dato existente en la encuesta) o cold-deck (se obtienen de encuestas anteriores u otras fuentes históricas), regresión, máxima verosimilitud; o mediante imputación múltiple, elaborado por Rubin en 1976 y desarrollado por el mismo en 1983 junto a Herzog y en 1986 con Shafer. Este procedimiento consiste en reemplazar cada valor ausente utilizando m>1 simulaciones (Otero, 2011).

En la base de datos se puede observar que las variables AGE, YOJ, INCOME, HOME_VAL y CAR_AGE poseen valores perdidos. Debido a que, representan un 26,2% de los datos, se ha descartado la posibilidad de eliminarlos pues supondría una gran pérdida de información a la hora de realizar el estudio. Para solucionar esta ausencia de valores, se optó por llevar a cabo un método de imputación simple, reemplazando los valores perdidos de forma automática utilizando el método de la mediana, asignando a dichos valores la mediana de los dos puntos cercanos válidos. Se decide utilizar esta medida por ser un procedimiento robusto, aunque lo idóneo sería llevar a cabo una regresión para estimar los datos ausentes a través del valor del resto de variables, sin

embargo, dada la extensión del estudio y puesto que no es el objeto de análisis se elige un procedimiento automático.

Por otro lado, se decide utilizar apena un 80% de la base de datos para el estudio, con el objetivo de evitar un sobreajuste a la hora de aplicar los modelos, pues comúnmente suele ocurrir que un modelo ajuste muy bien una cantidad de datos, sin embargo, si se introducen más casos el modelo deja de ser bueno. A pesar de ello, y debido a la extensión del trabajo, no se ha utilizado el 20% restante.

Para efectuar esta división, se generó una variable aleatoria binaria, es decir, una variable que sigue una distribución Bernoulli y que asigna a cada individuo el valor uno o cero. Definiendo el parámetro p=0,8 para conseguir una división cercana al 80-20%, finalmente, la base de datos se divide en dos, de 8.263 y 2.039 casos respectivamente.

3. Análisis Clúster Bietápico: Introducción

Cuando un cliente desea contratar un seguro, se le asigna una prima o tarifa que deberá pagar de forma periódica durante el período de cobertura del mismo. Esta cuantía se calcula en función de sus características, por ello, teniendo en cuenta la gran diversidad de observaciones, se procederá a su división en diferentes grupos estimando para cada uno el valor de la prima correspondiente a la garantía/seguro.

Para llevar a cabo este procedimiento, se realizará un Análisis Clúster que se puede definir como *"un método estadístico multivariante de clasificación automática que a partir de una tabla de datos (casos-variable), trata de situarlos en grupos homogéneos, conglomerados o clústeres, no conocidos de antemano pero sugeridos por la propia esencia de los datos, de manera que los individuos que puedan ser considerados similares sean asignados a un mismo clúster, mientras que los individuos diferentes (disimilares) se localicen en clústeres distintos"* (Pérez, 2005). Es decir, a partir de la observación de algunas variables se agruparán los individuos que posean características comunes en un mismo grupo, formándose así los diferentes clústeres.

El análisis de conglomerados es un método utilizado tradicionalmente para el estudio de grupos en diversos ámbitos, desde Ciencias Sociales como economía o marketing hasta Ciencias de la Salud, en medicina o enfermería.

La elección de las variables y del tipo de clúster a realizar dependerá del objetivo de cada estudio. En este caso, se pretende formar grupos con el propósito de asignar a cada uno una tarifa de seguro de automóvil acorde a sus características. Partiendo de ello, se seleccionaron las siguientes variables para la formación de los clústeres: EDUCATION, INCOME y URBANICITY. Se considera importante el tipo de zona por donde circula el individuo, pues el entorno urbano o rural afecta no solo a la forma de conducir sino también al grado de exposición al riesgo de siniestralidad. Por otro lado, el nivel de estudios y sus ingresos influyen principalmente en características relacionadas con el tipo de vehículo, sin embargo, de forma indirecta también con la manera de conducir.

Dada la naturaleza de estas tres variables, se optó por realizar un Análisis Clúster en Dos Fases, conocido también como Clúster Bietápico. Tal y como especifica César Pérez (2005), este tipo de análisis posee algunas características que lo diferencia de los métodos más habituales, como son el jerárquico y el no jerárquico:

- *Procedimiento automático del número óptimo de conglomerados*: a través de la comparación de valores de los criterios de selección del modelo para las distintas soluciones de clústeres, este procedimiento determina automáticamente el número óptimo de grupos.
- *Posibilidad de crear modelos de conglomerados con variables categóricas y continuas:* suponiendo que las variables son independientes, permite aplicarles una distribución normal multinomial conjunta.
- *Archivos de datos de gran tamaño*: el algoritmo en dos fases resume los datos construyendo un árbol de características de clústeres, lo que permite analizar archivos de gran tamaño.

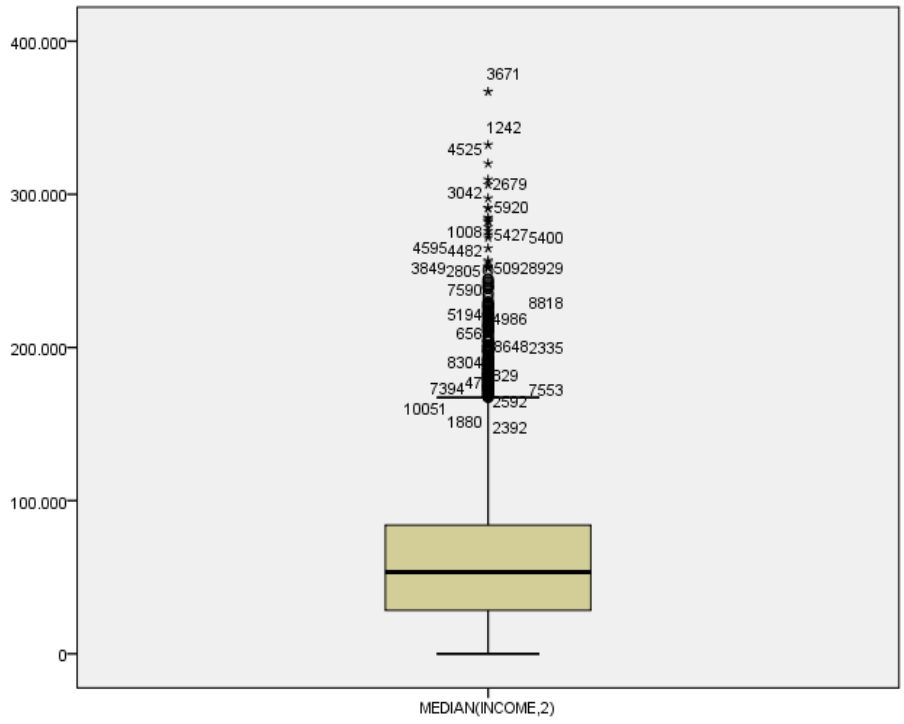
Esta técnica requiere que las variables sean independientes entre sí, que las cuantitativas se distribuían según una normal y las cualitativas como una multinomial. A continuación, se comprobará si se cumplen estas condiciones.

3.1. Requisitos Clúster Bietápico

3.1.1. Distribución de probabilidad de la variable cuantitativa: INCOME

La variable INCOME posee un rango amplio de valores: desde 0 hasta 367.030,26 euros. En el siguiente diagrama de caja, correspondiente a la muestra de 8.263 casos, se observa como la variable ostenta diversos valores extremos, siendo el caso 3671 el que toma el valor más alejado. Según Tukey (1977) y su *Análisis Exploratorio de Datos*, los valores atípicos se corresponden con aquellos casos que toman valores superiores a 1,5 longitudes de caja del percentil 75 y los extremos con valores superiores a 3 longitudes de caja del percentil 75. Debido a la gran cantidad de este tipo de casos (16,62%), se optó por no eliminarlos del análisis. Sin embargo, esto afecta en gran medida a la hora de identificar la distribución teórica de probabilidad que sigue la variable.

Gráfico 1: Diagrama de caja de INCOME



Fuente: elaboración propia

Para contrastar la hipótesis nula de que la variable ingresos se ajusta a una distribución normal, se efectuó la prueba de bondad de ajuste de Kolmogorov-Smirnov-Lilliefors puesto que es la más adecuada al tratarse de una variable cuantitativa y una muestra grande (superior a 30 casos). Este test se fundamenta en la comparación de la función de distribución empírica, F(x), y la función de distribución teórica, $F_0(x)$. Una vez calculadas, la máxima diferencia entre ambas se corresponde con el valor del estadístico (Pardo y Ruiz, 2005).

Realizando la prueba para la variable ingresos, su p-valor es muy pequeño e inferior a 0,05, por ello, a un nivel de confianza del 95% se rechaza la hipótesis nula de que la variable sigue una distribución normal.

Tabla 4: Prueba de normalidad de INCOME

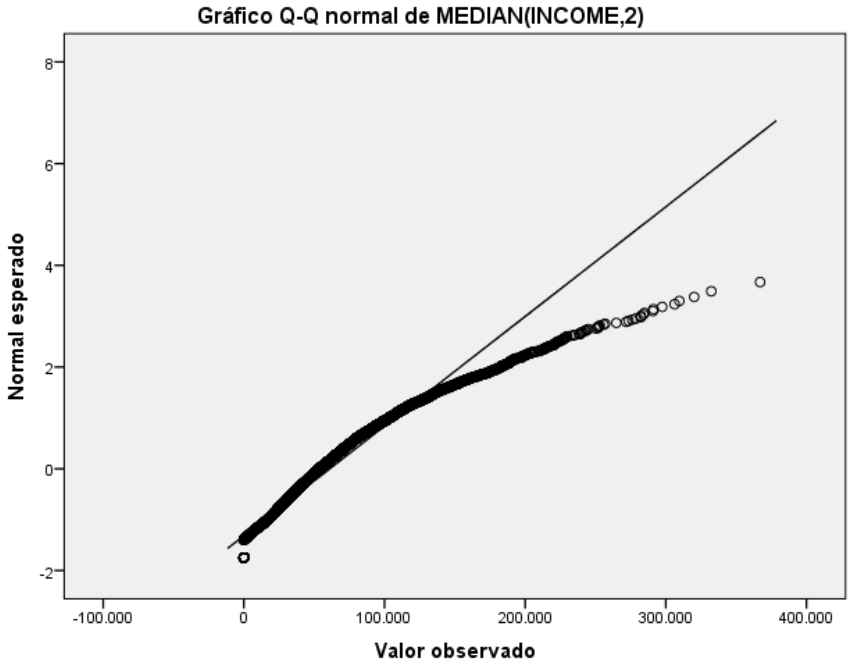
	Kolmogorov-Smirnov[a]		
	Estadístico	gl	Sig.
MEDIAN(INCOME,2)	0,094	8263	0,000

a. Corrección de significación de Lilliefors

Fuente: elaboración propia

Analizando tanto el diagrama de caja como el siguiente gráfico cuantil-cuantil, se observa que la distribución de la variable posee una asimetría positiva, algo habitual en variables relacionadas con ingresos, pues su rango siempre es positivo y habitualmente amplio.

Gráfico 2: Cuantil-Cuantil INCOME



Fuente: elaboración propia

En 1964 Box y Cox establecieron una familia de transformaciones, mediante las cuales se pueden lograr que un conjunto de datos se distribuya según una normal. De este modo, teniendo en cuenta la asimetría existente, se efectuaron las siguientes transformaciones: logarítmica, raíz cuadrada e inversa. Estos tres tipos se utilizan para corregir la asimetría positiva, tal y como establece Tukey (1987) en su conocida "Escalera de transformaciones", donde especifica que tipos de transformaciones utilizar en función del tipo de asimetría existente.

Tabla 5: Pruebas de normalidad de las transformaciones de INCOME

	Kolmogorov-Smirnov[a]		
	Estadístico	gl	Sig.
RaizIncome	0,023	7594	0,000
LnIncome	0,090	7594	0,000
InversaIncome	0,484	7594	0,000

a. Corrección de significación de Lilliefors

Fuente: elaboración propia

Realizadas las transformaciones, se rechaza de nuevo la hipótesis nula de que la variable siga una distribución normal a un nivel de confianza del 95%, dado que el p-valor toma el valor cero y, por lo tanto, es inferior a 0,05.

3.1.2. Distribución de probabilidad de las variables cualitativas: EDUCATION Y URBANICITY

Por definición, un conjunto de observaciones sigue una distribución multinomial cuando existe un conjunto de sucesos (k) que pueden darse con una determinada probabilidad en n ensayos (Sarabia, Gómez y Vázquez, 2007). Siendo:

- A_i los posibles sucesos, i = 1,2, …, k
- $Pr(A_i)= p_i \geq 0$
- X_i= número de sucesos del tipo A_i en los n ensayos
- $X= X_1 + X_2 +…+ X_k$

Se dice que X sigue una distribución multinomial:

$$X \sim Multi(n, p_{1,} p_{2, ..., } p_k)$$

En el caso de que k tome el valor 2, se obtiene la distribución binomial, siendo esta, por lo tanto, un caso particular de la distribución multinomial.

Respecto a las variables EDUCATION y URBANICITY, poseen cinco y dos categorías respectivamente. Si se realizaran n ensayos sobre los datos disponibles, cada categoría tendría una distribución de frecuencias acorde a su probabilidad p_i de ser seleccionada. Por lo que, por definición, las variables siguen una distribución multinomial.

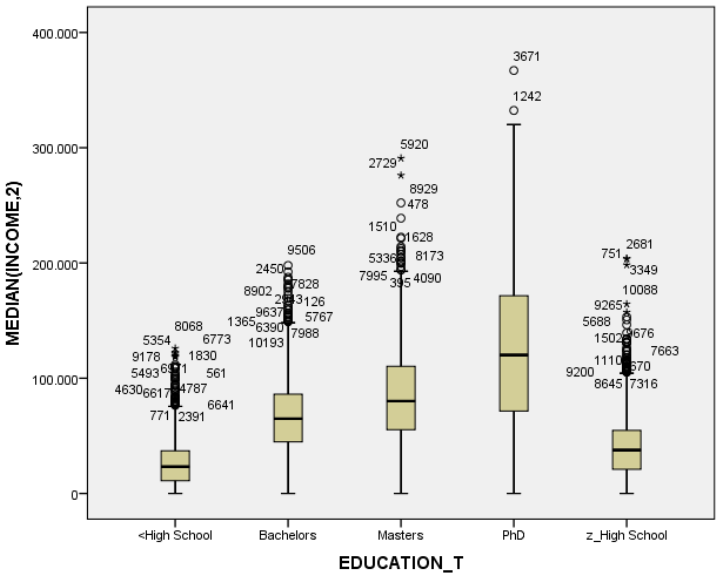
3.1.3. Variables independientes

Puesto que se utilizarán tres variables, se comprobará dos a dos si son independientes entre sí.

INCOME – EDUCATION

Antes de proceder a la comprobación de independencia en sentido estadístico, veamos cómo se distribuye el salario en cada una de las categorías del nivel de educación:

Gráfico 3: Diagrama de caja INCOME-EDUCATION



Fuente: elaboración propia

Se observa que la dispersión de los ingresos es muy distinta en cada nivel de educación, siendo el grupo de doctorados el que presenta mayor dispersión. Por otro lado, al igual que en la distribución de INCOME, en cada categoría existe asimetría positiva, y además, en cada una existen varios valores atípicos y extremos, con la excepción del nivel de doctorado que presenta dos.

Desde un punto de vista general es lógico pensar que, a mayor grado de estudios, se percibirá una cantidad superior de ingresos. En la gráfica, se puede comprobar como la media de ingresos se incrementa a lo largo de los cinco niveles de educación, siendo los individuos que con estudios inferiores a

instituto los que perciben un salario menor y los que poseen un doctorado los que obtienen más ingresos.

Tabla 6: Pruebas de normalidad INCOME-EDUCATION

EDUCATION_T		Kolmogorov-Smirnov[a]			Shapiro-Wilk		
		Estadístico	gl	Sig.	Estadístico	gl	Sig.
MEDIAN(INCOME,2)	<High School	0,116	1.199	0,000	0,904	1.199	0,000
	Bachelors	0,036	2.269	0,000	0,982	2.269	0,000
	Masters	0,040	1.665	0,000	0,984	1.665	0,000
	PhD	0,036	749	0,025	0,987	749	0,000
	z_High School	0,079	2.381	0,000	0,947	2.381	0,000

a. Corrección de significación de Lilliefors

Fuente: elaboración propia

Realizando la prueba de bondad de ajuste de Kolmogorov-Smirnov-Lilliefors, se comprueba que los ingresos no se distribuyen según una normal en los diferentes niveles de educación (p-valor inferior a 0,05, se rechaza la hipótesis nula de normalidad). Como consecuencia, no se puede llevar a cabo el contraste ANOVA clásico para verificar la independencia de las variables.

Teniendo en cuenta lo anterior, se procede a realizar la prueba H de Kruskal y Wallis (1952), procedimiento no paramétrico y similar al método ANOVA, pero con las ventajas de que no necesita que los datos cumplan los supuestos de normalidad y homocedasticidad, y, además, permite trabajar con datos ordinales (Pardo y Ruiz 2005).

En esta prueba se parte de j categorías (5 en el estudio). Definiendo π_j como el promedio de cada clase, la hipótesis a contrastar es:

H_0: $\pi_1 = \pi_2 = \ldots = \pi$

H_1: $\pi_j \neq \pi$

Se procede ordenando las observaciones (n) de menor a mayor y estableciendo para cada una su rango, siendo el 1 para la menor, 2 para la siguiente, y así sucesivamente. A continuación, se calculan los Rj, suma de los

rangos de las observaciones de la muestra j, y se calcula el valor del estadístico:

$$H = \frac{12}{n(n+1)} \sum_{j=1}^{J} \frac{R_j^2}{n_j} - 3(n+1)$$

Tablas 7: Prueba H de Kruskal-Wallis

EDUCATION_T		N	Rango promedio
MEDIAN(INCOME,2)	<High School	1.199	2.127,80
	Bachelors	2.269	4.639,78
	Masters	1.665	5.517,14
	PhD	749	6.358,19
	z_High School	2.381	2.988,45
	Total	8.263	

Estadísticos de prueba[a,b]

	MEDIAN(INCOME,2)
Chi-cuadrado	2.711,480
gl	4
Sig. asintótica	0,000

a. Prueba de Kruskal Wallis

b. Variable de agrupación: EDUCATION_T

Fuente: elaboración propia

Obtenidos los resultados, se rechaza la hipótesis nula a un nivel de confianza del 95% puesto que el p-valor es igual a cero. Los promedios del nivel de ingresos son distintos en cada nivel de estudios, tal y como es de esperar, pues un mayor nivel de estudios equivale a una profesión con mayor salario y por lo tanto ingresos superiores. Por ello, las variables INCOME y EDUCATION se encuentran relacionadas, una influencia el comportamiento de la otra.

INCOME – URBANICITY

Siendo URBANICITY una variable cualitativa con dos categorías, para comprobar su independencia de la variable ingresos se efectuó un contraste de diferencias de medias: Prueba T para muestras independientes. Este método permite comparar dos grupos distintos contrastando la independencia de sus medias (Pardo y Ruiz, 2005). La hipótesis nula a contrastar es ahora:

H_0= igualdad de medias de los ingresos en el entorno urbano y rural.

El estadístico de contraste se calcula como la tipificación de la diferencia de las dos medias muestrales:

$$T = \frac{(\overline{Y_1} - \overline{Y_2}) - (\mu_1 - \mu_2)}{\hat{\sigma}_{\overline{Y_1} - \overline{Y_2}}}$$

El cálculo de la desviación típica de la diferencia dependerá de que las varianzas poblacionales sean iguales o no. En este caso, se puede contemplar en la siguiente tabla que el p-valor de la Prueba de Levene (contraste de homogeneidad de varianzas) es inferior a 0,05 por lo que se rechaza la igualdad de varianzas. En cuanto al p-valor del estadístico T, este es también inferior a 0,05 por lo que se rechaza, a un nivel de confianza del 95%, la hipótesis nula de igualdad de medias de los ingresos en las categorías urbano y rural.

Tabla 8: Contraste de diferencias de medias: Prueba T para muestras independientes

		Prueba de Levene de igualdad de varianzas		Prueba t para la igualdad de medias						
		F	Sig.	t	gl	Sig. (bilateral)	Diferencia de medias	Diferencia de error estándar	95% de intervalo de confianza de la diferencia	
									Inferior	Superior
MEDIAN (INCOME,2)	Se asumen varianzas iguales	122,839	0,000	18,596	8.261	0,000	23.238,552	1.249,633	20.788,958	25.688,147
	No se asumen varianzas iguales			22,258	3.351,084	0,000	23.238,552	1.044,069	21.191,476	25.285,628

Fuente: elaboración propia

Analizando los estadísticos de la relación de estas dos variables, se comprueba que la media de ingresos de conductores en un entorno urbano es superior en relación con la de un entorno rural.

Tabla 9: Estadísticos INCOME – URBANICITY

URBANICITY_T		N	Media	Desviación estándar	Media de error estándar
INCOME_1	Highly Urban/ Urban	6.609	65.714,66	47.686,076	586,575
	z_Highly Rural/ Rural	1.654	42.476,11	35.126,889	863,718

Fuente: elaboración propia

EDUCATION – URBANICITY

Para averiguar la relación existente entre estas dos variables categóricas, es necesario recurrir a una medida de asociación y su correspondiente prueba de significación. Al tratarse de dos variables cualitativas se utilizará el estadístico Chi-Cuadrado, que permite comprobar la independencia entre ambas. Tal y como explican Pardo y Ruiz (2005), consiste en la comparación de las frecuencias observadas y las frecuencias esperadas. Las frecuencias esperadas (m_{ij}) son aquellas que existirían si los criterios de clasificación fueran independientes, y se calculan como el producto de las frecuencias marginales entre el número total de casos. En cambio, las frecuencias observadas (n_{ij}) son las realmente obtenidas. El estadístico se deduce mediante la siguiente fórmula:

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

El valor del estadístico será cero si las variables son totalmente independientes.

Tabla 10: Pruebas de Chi-Cuadrado

Pruebas de chi-cuadrado

	Valor	df	Significación asintótica (bilateral)
Chi-cuadrado de Pearson	467,260[a]	4	0,000
Razón de verosimilitud	520,672	4	0,000
Asociación lineal por lineal	0,871	1	0,351
N de casos válidos	8.263		

a. 0 casillas (,0%) han esperado un recuento menor que 5. El recuento mínimo esperado es 149,93.

Fuente: elaboración propia

Realizada la prueba, se observa que el estadístico no toma el valor cero y que su p-valor es inferior a 0,05, por lo que se puede afirmar a un nivel de confianza del 95% que las variables no son independientes, EDUCATION y URBANICITY están relacionadas.

3.2. Robustez del procedimiento

Tras el análisis realizado, se concluye que la variable INCOME no se distribuye según una normal, sin embargo, esto se debe a la gran cantidad de valores atípicos que posee y que se opta por su no exclusión pues distorsionaría el análisis. En cuanto a las variables categóricas, por definición, si se distribuyen según una multinomial. Por otro lado, los ingresos no son independientes del nivel de estudios y del tipo de zona por la que circula el conductor.

A pesar de que no se cumplen todos los requisitos para la realización de este tipo de clúster, tal y como comentan Rubio-Hurtado y Vilà-Baños, la evidencia empírica ha demostrado que este tipo de procedimiento de formación de

conglomerados es suficientemente robusto. Por este motivo, se proseguirá con el análisis bietápico.

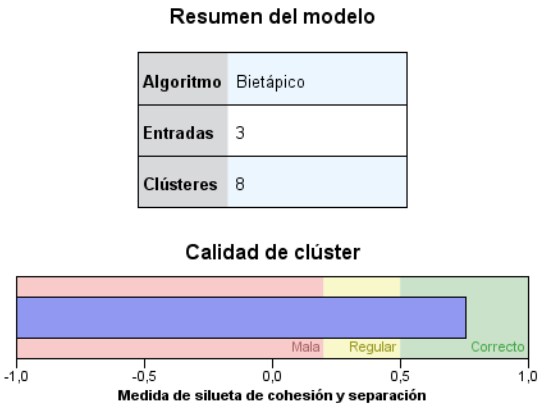
3.3. Análisis Clúster Bietápico: Resultados

Al efectuar un análisis con variables continuas y categóricas se elige la medida de distancia de Log-verosimilitud, que genera una distribución de probabilidad entre las variables partiendo del supuesto de que se cumplen los requisitos comentados anteriormente. La distancia entre los clústeres cambiará en función del decremento en el log-verosimilitud.

Respecto al criterio de agrupación de los clústeres, el programa SPSS permite utilizar el criterio de información bayesiano (BIC) y el criterio de información de Akaike (AIC). Ambos evalúan la calidad de los algoritmos partiendo de que emplear una mayor cantidad de parámetros a la hora de determinar el número óptimo de grupos puede conllevar a un sobreajuste. El BIC penaliza en mayor medida el sobreajuste que el AIC, por lo que suele ser el más adecuado (Rubio-Hurtado y Vilà-Baños, 2016).

Se llevó a cabo el análisis por los dos métodos y estos proporcionaron el mismo resultado. Eligiendo la solución obtenida mediante el criterio BIC, se forman 8 clústeres:

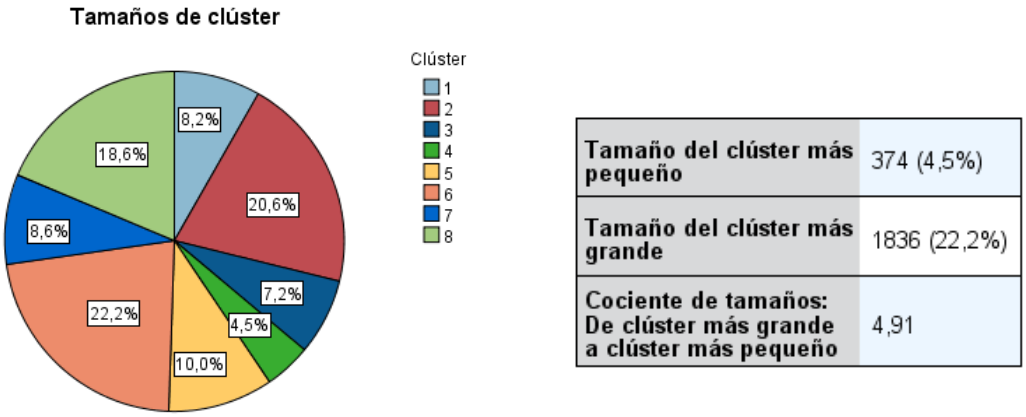
Figura 1: Clúster Bietápico



Fuente: elaboración propia

Según Kaufman y Rouseeuw (1990) y su análisis de estructuras de conglomerados, la calidad de los clústeres será correcta si existe evidencia de dicha estructura en la población; regular si esa evidencia es débil; y mala si no existe tal evidencia. Calidad igual a 1 indicaría que cada caso se encontraría en el centro de su conglomerado y -1 revelaría lo contrario, mientras que el valor 0 significaría que los casos se ubicarían de forma equidistante entre el conglomerado al que se asignó y el siguiente más próximo. En este caso, la calidad es correcta (0,8) y por tanto se puede afirmar que la base de datos se podría separar en estos ocho grupos.

Figura 2: Estructura de los conglomerados



Tamaño del clúster más pequeño	374 (4,5%)
Tamaño del clúster más grande	1836 (22,2%)
Cociente de tamaños: De clúster más grande a clúster más pequeño	4,91

Fuente: elaboración propia

Los clústeres 2, 6 y 8 son los que poseen un mayor número de casos, acaparando el 61,4% de los datos. Asimismo, se observa que la relación de tamaño entre el clúster más y menos numeroso es de 4,91 a 1.

Figura 3: Características de los clústeres obtenidos

Clústeres

Importancia de entrada (predictor)
1,0 0,8 0,6 0,4 0,2 0,0

Clúster	1	2	3	4	5	6	7	8
Etiqueta								
Descripción								
Tamaño	8,2% (681)	20,6% (1700)	7,2% (599)	4,5% (374)	10,0% (825)	22,2% (1836)	8,6% (712)	18,6% (1536)
Entradas	EDUCATION_T	EDUCATION_T	EDUCATION_T Bachelors (72,3%)	EDUCATION_T	EDUCATION_T	EDUCATION_T Bachelors (100,0%)	EDUCATION_T PhD (100,0%)	EDUCATION_T Masters (100,0%)
	MEDIAN(INCOME,2) 35.073,18	MEDIAN(INCOME,2) 41.154,37	MEDIAN(INCOME,2) 60.918,54	MEDIAN(INCOME,2) 26.418,30	MEDIAN(INCOME,2) 27.508,02	MEDIAN(INCOME,2) 66.498,17	MEDIAN(INCOME,2) 124.626,62	MEDIAN(INCOME,2) 85.173,74
	URBANICITY_T	URBANICITY_T	URBANICITY_T	URBANICITY_T	URBANICITY_T	URBANICITY_T	URBANICITY_T	URBANICITY_T

Fuente: elaboración propia

Analizando la tabla resumen, se puede afirmar que los grupos 6, 7 y 8 se definen por personas con un nivel de estudios igual o superior a una carrera universitaria y por poseer los importes de ingresos más elevados en comparación con los demás conglomerados. En cuanto a la importancia predictora de las variables, tanto EDUCATION como URBANICITY e INCOME son igual de relevantes con importancia de 1.

4. Inflado de ceros

Según lo indicado en el apartado anterior, la variable número de siniestros (CLM_FREQ) presenta un exceso de ceros, algo común en el ámbito del seguro de automóvil. La gran cantidad de asegurados con ceros siniestros no indica en realidad que no hayan sufrido ningún accidente, estos ceros son de dos tipos:

- Siniestros ocurridos y no declarados (ceros muestrales): los asegurados optan frecuentemente por no declarar los siniestros que han sufrido, ya sea porque han sido de pequeña importancia o debido a que ello puede llevar a un incremento de la prima en los siguientes períodos.
- Ausencia de siniestros (ceros estructurales): se corresponden con aquellos asegurados que en realidad no han sufrido ningún accidente.

En la elaboración de modelos predictivos es importante diferenciar los ceros muestrales (Mackenzie *et al*. 2002) de los estructurales (Welsh *et al*. 1996, 2000, Barry and Welsh 2002, Podlich *et al*. 2002), pues en caso contrario puede llevar a que se estimen los parámetros incorrectamente, lo que produciría estimaciones erróneas.

A continuación, se explicarán los métodos más comunes que se utilizan para modelar datos con inflado de ceros.

4.1. Modelos Lineales Generalizados

Los modelos lineales generalizados se utilizan como una herramienta que permite valorar y cuantificar la relación existente entre la variable respuesta y las variables explicativas. Fueron propuestos por J. A. Nelder y R. W. N. Wedderburn en 1972, y se diferencian de los modelos de regresión lineal en tres aspectos de la variable respuesta:

- Su distribución pertenece a la familia exponencial, conteniendo a la normal como un caso particular, sin embargo, ya no es necesario que siga una distribución normal como es el caso de la regresión lineal clásica.
- Su esperanza se relaciona linealmente con las variables explicativas, aunque no directamente, sino a través de una función de enlace.
- Su varianza no es necesariamente constante, es una función de su esperanza. Esto es debido a que la variable respuesta sigue una distribución exponencial, siendo habitualmente heterocedástica, lo que hace que su varianza cambie en función de la media.

Estas características hacen que los modelos lineales generalizados sean de gran utilidad en los seguros, pues los datos difícilmente siguen una distribución normal y tampoco suelen presentar homecedasticidad.

Estructura de un GLM

Como indican Piet de Jong y Z. Heller (2008), la función de probabilidad que define las distribuciones pertenecientes a la familia exponencial adquieren la siguiente forma:

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

- Las funciones c() y a() determinan que tipo de distribución seguirá la variable respuesta (Binomial, Normal, Gamma,…).
- θ y φ se corresponden con el parámetro canónico y de dispersión respectivamente. El primer está relacionado con el parámetro de

localización, por lo tanto, asimismo con la media. El segundo, en cambio, tiene relación con el parámetro de escala y consecuentemente con la varianza. Estos dos parámetros son los que caracterizan las distribuciones de la familia exponencial.

A partir de la primera ecuación se puede definir la esperanza y la varianza de la variable respuesta (y):

$$E(y) = \dot{a}(\theta) \quad , \quad Var(y) = \varphi \, \ddot{a}(\theta)$$

Siendo $\dot{a}(\theta)$ y $\ddot{a}(\theta)$,respectivamente, la primera y segunda derivada de $a(\theta)$ respecto de θ.

Función Enlace

Tal y como se comentó al inicio de este apartado, la media de la variable respuesta no se relaciona directamente con las variables explicativas, sino mediante una transformación de su esperanza (μ):

$$g(\mu) = x' \beta$$

La transformación viene determinada por la función de enlace, g(), siendo esta monótona y diferenciable. Las funciones enlace sirven para obligar al modelo a calcular de forma coherente el valor de los parámetros a estimar, es decir, por ejemplo, en el caso de la distribución de Poisson sería erróneo obtener un valor negativo al predecir el número de siniestros.

Cada función de distribución, de acuerdo a sus características, tiene asociada una función de enlace:

Tabla 11: Funciones de enlace características

Función de Distribución	g(*μ*)	Función de Enlace
Normal	μ	Identidad
Poisson	ln μ	Log
Gamma (p= -1)	μ^p	Potencia
Inversa Gaussiana (p= -2)	$\sqrt{\mu}$	Raíz cuadrada
Binomial	ln $\frac{\mu}{1-\mu}$	Logit

Fuente: Piet de Jong y Z. Heller, G. (2008)

Exposición

A la hora de realizar predicciones acerca del número de siniestros o sus cuantías, por ejemplo, es fundamental considerar la exposición al riesgo de los asegurados. Esta se puede definir como el período de tiempo que estuvo un asegurado expuesto al riesgo.

Esta información se introduce en un modelo lineal generalizado no como una variable explicativa adicional, sino como una variable "infiltrada" que corrige la variable respuesta. Siendo "y" la variable respuesta y "n" la exposición, para una función de enlace logarítmica:

$$g(\frac{\mu}{n}) = x' \beta \rightarrow \ln \mu = \ln n + x' \beta$$

El ln n recibe el nombre de *offset*, adoptando la forma de una variable explicativa en la regresión, pero con un coeficiente β igual a uno. De este modo, el valor esperado de "y" es directamente proporcional a la exposición al riesgo:

$$\mu = n\, e^{x'\beta}$$

4.2. Poisson Inflado de Ceros

Jonhson, Kotz y Kemp desarrollaron por primera vez en 1969 modelos Poisson y Binomial Negativa para el tratamiento de datos con inflado de ceros, pero sin tener en cuenta covariables. Fue Diane Lambert (1992), quién generalizó estos modelos introduciendo covariables.

Lambert define el modelo Poisson Inflado de Ceros (Zero Inflated Poisson, en adelante ZIP) como un "*modelo de regresión para datos de conteo con excesos de zeros. El cual asume que se puede observar un cero con propabilidad "p", y con probabilidad "1-p" una distribución de Poisson de parámetro lambda".* En su artículo demuestra que se puede ajustar los datos a una Poisson, aplicando una distribución Binomial para corregir el exceso de ceros y una Poisson a las restantes observaciones.

Partiendo del caso de estudio, en el que se dispone de datos de número de siniestros con una cantidad elevada de ceros, estos pueden corresponderse a asegurados que han sufrido accidentes pero que han decidido no declararlos o a asegurados que en realidad no han tenido ningún siniestro. Partiendo de esta diferenciación y del método elaborado por Lambert, siendo Y el vector de la variable respuesta de un modelo de regresión ZIP:

$$\begin{cases} Y_i = 0 \text{ con probabilidad } p_i \\ Y_i \sim \text{Poisson } (\lambda_i) \text{ con probabilidad } 1\text{-}p_i \end{cases}$$

La función de probabilidad se puede definir como:

$$P(Y_i=y) = \begin{cases} p_i + (1\text{-}p_i)\, e^{-\lambda i} & \text{, si } y = 0 \\ (1\text{-}p_i)\, \frac{e^{-\lambda i} * \lambda^y}{y!} & \text{, si } y > 0 \end{cases}$$

La probabilidad "p_i" se corresponde con aquella de que se observen ceros muestrales, procedentes de asegurados que han sufrido siniestros, pero no los

han declarado, y con probabilidad "1-p_i" el resto de valores, incluidos los ceros estructurales (Velasco, 2008).

Según Lambert, los ceros muestrales se explicarían mediante una distribución Binomial y el resto de valores se ajustarían a una distribución de Poisson.

El modelo ZIP actúa como un Modelo Lineal Generalizado, en concreto, se estima en primer lugar el modelo Binomial utilizando como función de enlace la función logit. A continuación, a partir de lo obtenido, se estima un Modelo Lineal Generalizado con función de enlace logarítmica.

4.3. Binomial Negativa Inflado de Ceros

En el año 2000, Welsh et al proponen que la distribución Binomial Negativa es la más adecuada para el tratamiento de datos con inflado de ceros y, además, presenten sobredispersión, surgiendo así los Modelos Binomial Negativa de Inflado de Ceros (en adelante ZINB).

Según lo analizado en el primer apartado de este estudio, la variable CLM_FREQ presenta sobredispersión, en concreto, su coeficiente de variación toma el valor 1,44.

Teniendo en cuenta la presencia de excesos de ceros, según Welsh, los Modelos ZINB serían los más apropiados para ajustar la frecuencia del número de siniestros.

En este caso, de la misma forma que en los ZIP, existen dos posibilidades: los ceros muestrales y el resto de valores, que incluyen los ceros estructurales. Estos casos se producirían con las siguientes probabilidades:

$$\begin{cases} Y_i = 0 \text{ con probabilidad } \pi_i \\ Y_i \sim \text{Binomial Negativa } (\mu_i, \alpha) \text{ con probabilidad } 1-\pi_i \end{cases}$$

La función de probabilidad adopta la siguiente forma:

$$P(Y_i=y) = \begin{cases} \pi_i + (1-\pi_i)\, g(y_i = 0) & , \text{si } y = 0 \\ (1-\pi_i)\, g(y_i) & , \text{si } y > 0 \end{cases}$$

De forma similar a los ZIP, los ZINB también funcionan como Modelos Lineales Generalizados. Los ceros muestrales se distribuyen según una Binomial con probabilidad π_i y función de enlace logit, mientras que con probabilidad $1-\pi_i$, los demás datos se distribuyen según una Binomial Negativa con función de enlace habitual la logarítmica y función de densidad:

$$g(y_i) = P(Y= y_i \mid \mu_i, \alpha) = \frac{\Gamma(y_i+ \alpha^{-1})}{\Gamma(\alpha^{-1})\Gamma(y_i+ 1)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\alpha^{-1}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i}$$

Cuando α tiende a cero, el modelo resultante es el ZIP, mientras que, para valores positivos del parámetro de dispersión, el modelo sería ZINB.

5. Modelo Hurdle

El Modelo o Distribución de Hurdle fue introducido por Cragg (1971) y revisado más tarde por Mullahy (1986). Este modelo se denomina comúnmente Modelo Valla pues se basa en un proceso antes y despúes de un “obstáculo”, siendo este el valor cero en la mayoría de los casos (Boucher y Guillén, 2009).

En contrapartida con los modelos de Inflado de Ceros, este modelo no diferencia los dos tipos de ceros, sino que se efectúa una división entre ceros y los restantes valores. Los ceros se distribuyen según un Binomial y para los valores distintos de cero, el modelo “cruza la valla” y se rige por una distribución truncada en cero que puede ser una Poisson, una Binomial Negativa o una Geométrica.

5.1. Modelo Hurdle: Poisson Truncada

La distribución de Poisson no puede tomar el valor cero, sino que siempre tomará valores positivos. Por lo que para poder truncarla en ese valor y lograr eliminarlo, se puede utilizar el método de calcular la probabilidad de que tome el valor cero y dividirla proporcionalmente entre todos los valores que toma la distribución (Bueno, 2015).

La función de probabilidad adopta la siguiente forma:

$$P(Y_i=y) = \begin{cases} h(o; z; y) & \text{, si } y = 0 \\ (1- h(0; z; y)) \frac{f(y;x;\beta)}{1-f(0;x;\beta)} & \text{, si } y > 0 \end{cases}$$

Siendo h la función de distribución Binomial que ajusta los valores cero, y f la función de distribución de Poisson según la que se distribuyen los restantes valores que toman los datos. Se divide entre $1 - f(0; x; \beta)$ para truncar la Poisson.

La estimación de la Binomial se realiza mediante Modelo Lineal Generalizado con función de enlace logit y la de Poisson truncada de igual manera, pero con función de enlace logarítmica.

5.2. Modelo Hurdle: Binomial Negativa Truncada

Siguiendo el procedimiento del apartado anterior, se trunca en cero la distribución Binomial Negativa para modelar las observaciones que no toman el valor cero. Su distribución de probabilidad es en este caso:

$$P(Y_i=y) = \begin{cases} h(o; z; y) & \text{, si } y = 0 \\ (1- h(0; z; y)) \frac{g(y;x;\beta)}{1-g(0;x;\beta)} & \text{, si } y > 0 \end{cases}$$

Siendo g() la función de distribución Binomial Negativa, se divide también entre $1 - g(0; x; \beta)$, truncando así la distribución.

La distribución Binomial y la Binomial Negativa truncada se estiman mediante un Modelo Lineal Generalizado con función de enlace logit y logarítmica respectivamente.

6. Análisis de la frecuencia de siniestros

Dada la extensión del trabajo, se optó por seleccionar cuatro de los ocho clústeres obtenidos para el análisis de la frecuencia de siniestros: los tres más numerosos (el 2,6 y el 8) y el clúster que posee menos miembros (el número 4).

6.1. Características de los clústeres elegidos

A continuación, se indica para cada clúster, las características de las tres variables utilizadas para la formación de los grupos:

Tabla 12: Características de los clústeres seleccionados

	Clúster 2	Clúster 4	Clúster 6	Clúster 8
Media Ingresos	41.154	22.869	66.498	85.174
Educación	z_High School	<High School	Bachelors	Masters
Zona	Highly Urban/ Urban	z_Highly Rural/ Rural	Highly Urban/ Urban	Highly Urban/ Urban
Nº Miembros	1.710	374	1.836	1.536

Fuente: elaboración propia

Se observa que los grupos 6 y 8 son los que perciben ingresos más elevados, algo lógico pues poseen un nivel educativo elevado, carrera universitaria y máster respectivamente. Los miembros de estos grupos circulan en un entorno urbano, zona donde trabajan.

Por otro lado, el clúster 2 posee un nivel de ingresos intermedio, estudios hasta instituto y se mueven en una zona urbana. En cuanto al clúster 4, sus componentes son los que perciben un salario menor y tienen un nivel de estudios inferior a todos los demás, transitando en un entorno rural.

6.2. Estimación de los modelos

Mediante el programa R y utilizando el paquete “pscl” creado por Jackman en 2008, se han estimado los modelos explicados anteriormente a través de las funciones *zeroinfl()* y *hurdle()* pertenecientes a dicho paquete.

Tras varias pruebas, las variables seleccionadas (cuya descripción se encuentra en la tabla 1) para la estimación de los cuatro modelos han sido:

Clúster 2: AGE_1, CLAIM_FLAG, MVR_PTS.

Clúster 4: PARENT1, MVR_PTS.

Clúster 6: MSTATUS, GENDER, CAR_USE, CLAIM_FLAG, MVR_PTS.

Clúster 8: MSTATUS, CAR_USE, CLAIM_FLAG, MVR_PTS, TRAVTIME.

Los modelos se han estimado incluyendo la exposición, siendo esta de 5 años, período al cual se corresponden los datos.

6.3. Resultados obtenidos

En las siguientes tablas se puede comprobar que los valores teóricos obtenidos son bastante similares en los cuatro modelos:

Tablas 13: Valores teóricos de la frecuencia de siniestros

Clúster 2

Nº de siniestros		0	1	2	3	4	5
Valores Observados		814	280	340	217	46	3
Valores Teóricos	ZIP	647,7263	566,1482	299,9008	124,6807	43,7152	13,2531
	ZIBN	647,7275	566,1416	299,9005	124,6831	43,7172	13,2540
	HURDLEP	644,5565	570,9570	299,8143	123,3997	43,2840	13,2783
	HURDLEBN	644,5586	570,9564	299,8134	123,3993	43,2838	13,2783

Clúster 4

Nº de siniestros		0	1	2	3	4
Valores Observados		319	23	21	8	3
Valores Teóricos	ZIP	296,1367	60,1406	13,4925	3,3004	0,7490
	ZIBN	296,1365	60,1400	13,4929	3,3007	0,7492
	HURDLEP	296,1908	60,3653	13,1402	3,2921	0,7976
	HURDLEBN	296,1909	60,3652	13,1401	3,2921	0,7976

Clúster 6

Nº de siniestros		0	1	2	3	4	5
Valores Observados		983	288	317	192	49	7
Valores Teóricos	ZIP	810,5579	572,3272	281,0532	115,3632	40,3990	12,1735
	ZIBN	810,5465	572,3299	281,0568	115,3656	40,4005	12,1742
	HURDLEP	802,8876	584,2448	280,0692	112,6834	39,6017	12,1970
	HURDLEBN	802,8879	584,2443	280,0690	112,6835	39,6018	12,1971

Clúster 8

Nº de siniestros		0	1	2	3	4	5
Valores Observados		948	184	204	163	32	5
Valores Teóricos	ZIP	764,9559	463,2959	199,1379	74,5172	24,6019	7,1397
	ZIBN	764,9653	463,2825	199,1339	74,5193	24,6048	7,1415
	HURDLEP	762,0791	468,5845	197,8552	73,2597	24,4346	7,2770
	HURDLEBN	762,0784	468,5830	197,8551	73,2604	24,4353	7,2774

Fuente: elaboración propia

6.4. Análisis de las diferencias entre valores observados y teóricos

En cuanto a la diferencia entre los valores reales y los estimados, se observa que existen importantes diferencias entre ambos.

Por este motivo, se llevó a cabo una prueba de bondad de ajuste Chi-cuadrado que permite comprobar si la distribución empírica de una variable se ajusta a una distribución teórica, siendo esta la hipótesis nula a contrastar. Definiendo n_{ij} como las frecuencias observadas y m_{ij} las frecuencias teóricas, el estadístico de contraste es el siguiente:

$$X^2 = \sum_i \sum_j \frac{(n_{ij} - m_{ij})^2}{m_{ij}}$$

No se rechaza la hipótesis nula si el valor del estadístico es menor que el valor del percentil 1-α de la distribución Chi-cuadrado con v grados de libertad. Para estos casos toma los siguientes valores:

Tabla 14: Valor del percentil 1-α de la distribución Chi-cuadrado con v grados de libertad

	$X^2_{(1-\alpha),\ v}$
Clúster 2	$X^2_{(1-0,05),\ 4} = 9,49$
Clúster 4	$X^2_{(1-0,05),\ 3} = 7,82$
Clúster 6	$X^2_{(1-0,05),\ 4} = 9,49$
Clúster 8	$X^2_{(1-0,05),\ 4} = 9,49$

Fuente: elaboración propia

Por otro lado, se han calculado las diferencias absolutas entre los valores estimados y los reales.

Tablas 15: Estadístico Chi-cuadrado y diferencias absolutas

Clúster 2	Diferencias Chi	Diferencias Absolutas	Diferencias Absolutas (%)
ZIP	269,0815	597,38	35,1399%
ZIBN	269,0717	597,37	35,1393%
HURDLEP	277,3242	607,18	35,7165%
HURDLEBN	277,3236	607,18	35,7164%

Clúster 4	Estadístico Chi-Cuadrado	Diferencias Absolutas	Diferencias Absolutas (%)
ZIP	42,3356	74,4619	19,9096%
ZIBN	42,3312	74,4607	19,9093%
HURDLEP	42,4001	74,9446	20,0387%
HURDLEBN	42,3997	74,9444	20,0386%

Clúster 6	Estadístico Chi-Cuadrado	Diferencias Absolutas	Diferencias Absolutas (%)
ZIP	237,4754	583,1274	31,7607%
ZIBN	237,4773	583,1348	31,7612%
HURDLEP	255,7621	607,1999	33,0719%
HURDLEBN	255,7615	607,1992	33,0719%

Clúster 8	Estadístico Chi-Cuadrado	Diferencias Absolutas	Diferencias Absolutas (%)
ZIP	320,2231	565,2228	36,7984%
ZIBN	320,1979	565,2008	36,7969%
HURDLEP	331,3686	576,2329	37,5152%
HURDLEBN	331,3645	576,2312	37,5151%

Fuente: elaboración propia

Se observa que para cada uno de los modelos de cada clúster se rechaza la hipótesis nula, los valores observados y teóricos son muy distintos. Sin embargo, según comenta Vegas (1998), frecuentemente las distribuciones de Poisson y Binomial Negativa no superan la prueba del chi-cuadrado en el Seguro del Automóvil.

Asimismo, a través de las diferencias absolutas se comprueba que las diferencias son muy elevadas.

6.5. Elección del modelo: Criterio de Información de Akaike

A pesar de la gran diferencia entre los valores estimados y los reales, a la hora de elegir qué modelo ajustaría mejor la frecuencia de siniestros se recurre al Criterio de Información de Akaike (AIC). Este criterio, elaborado por Akaike en 1974, se basa en la penalización del exceso de parámetros ajustados, siendo en sí mismo un estimador muestral de E [ln f(X | θ)], es decir, la esperanza de la log-verosimilitud:

$$AIC\ (k) = -\ 2\ \ln\ \ L[\ \hat{\theta}\ (k)\ ] + 2k$$

L[$\theta\ (k)$] se corresponde con la función de verosimilitud de las observaciones, $\hat{\theta}\ (k)$ es la estimación de máxima verosimilitud del vector de parámetros θ , y k es el número de parámetros independientes estimados dentro del modelo (Caballero, 2011).

El AIC estima la pérdida de información entre el modelo real y el modelo estimado, es decir, la diferencia entre la distribución de probabilidad de las observaciones y la distribución de probabilidad asociada a la estimación.

Tabla 16: Valores AIC

AIC	Clúster 2	Clúster 4	Clúster 6	Clúster 8
ZIP	4.339,358	379,028	4.275,219	3.266,099
ZIBN	4.340,503	381,028	4.277,226	3.268,114
HURDLEP	4.311,359	373,784	4.253,214	3.251,393
HURDLEBN	4.313,359	375,784	4.255,215	3.253,397

Fuente: elaboración propia

De acuerdo a este criterio, se selecciona el modelo que posea un menor valor del AIC, y por lo tanto una menor pérdida de información. En los cuatro clústeres, se comprueba que el modelo Hurdle de Poisson es el que mejor ajusta la distribución de frecuencia de los siniestros.

6.6. Probabilidad de ocurrencia de siniestros para un asegurado

Para ver el funcionamiento del Modelo Hurdle Poisson, se procedió a calcular la probabilidad que tiene un asegurado de declarar siniestros, eligiendo el clúster número 4 a modo de ejemplo.

Según lo explicado el Modelo Hurdle Poisson posee dos funciones de enlace, una para el modelo binomial y otra para el de poisson, siendo estas logit y logarítmica respectivamente:

$$\text{Logit: } \mu = \frac{e^z}{1+e^z} \qquad \text{Logarítmica: } \lambda = e^z$$

Considerando los coeficientes estimados para el clúster 4:

Imagen 4: Modelo Hurdle Poisson, Clúster 4

```
Count model coefficients (truncated poisson with log link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -1.01386    0.21877  -4.634 3.58e-06 ***
datosc4$PARENT1Yes -0.10028    0.29636  -0.338    0.735
datosc4$MVR_PTS    -0.09438    0.06181  -1.527    0.127
Zero hurdle model coefficients (binomial with logit link):
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)        -3.18795    0.28753 -11.088  < 2e-16 ***
datosc4$PARENT1Yes  1.17256    0.39212   2.990  0.00279 **
datosc4$MVR_PTS     0.70566    0.09411   7.498 6.47e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Fuente: elaboración propia

Partiendo de un asegurado que sea padre/madre soltera y que tenga una puntuación de 10 en el motor del vehículo, se calculan la media de la distribución binomial y la lambda de la de poisson:

$$\mu = \frac{e^z}{1+e^z} = \frac{e^{-3,18795+1,17256+0,70566*10}}{1+e^{-3,18795+1,17256+0,70566*10}} = 0,9936$$

$$\lambda = e^{-1,01386-0,10028-0,09438*10} = 0,1277$$

Estos datos, indican que el asegurado tiene una probabilidad de 99,36% no declarar ningún siniestro (se recuerda que vive en un entorno rural y posee el

nivel de ingresos inferior de todos los clústeres), y que se estima que sufra 0,1277 siniestros en cinco años.

A partir de la función de densidad del Modelo Hurdle, se calcula las probabilidades del número de siniestros declarados:

$h(0; z; y)$ = 0,9936

(1- $h(0; z; y)$) $\frac{f(y;x;\beta)}{1-f(0;x;\beta)}$ = (1-0,9936)* $\frac{1}{1-0}$ * e^{-1} * $\frac{0{,}1277^1}{1!}$ = 0,00030066

Tabla 17: Probabilidades de ocurrencia de siniestros

Nº de siniestros	0	1	2	3	4	5
Probabilidades	0,9936	$3{,}0066 \times 10^{-4}$	$1{,}4124 \times 10^{-5}$	$6{,}6354 \times 10^{-7}$	$3{,}1172 \times 10^{-8}$	$1{,}4644 \times 10^{-9}$

Fuente: elaboración propia

Teniendo en cuenta que la probabilidad de que declare cero siniestros es muy elevada, cercana al 1, el resto de probabilidades son bastantes reducidas, y por lo tanto, es muy poco probable que declare algún siniestro en el período de 5 años.

6.7. Prima Pura

La prima total de un seguro se debe de basar en el riesgo que asume la compañía aseguradora, que le transfiere el asegurado. La prima pura es el principal componente de la prima total, y se define como la esperanza matemática de la siniestralidad total.

Una vez elegido el modelo que mejor ajuste la frecuencia de siniestros en cada uno de los clústeres, se calcula la esperanza de dicha variable en cada grupo.

Por otro lado, se toma la variable OLDCLAIM, que se corresponde con las cuantías pagadas de siniestros en los últimos cinco años, y se calcula su esperanza para cada clúster.

Mediante el producto de las esperanzas de la frecuencia de siniestros y de sus cuantías, se obtiene así la esperanza matemática de la siniestralidad total para cada grupo de asegurados:

E (Siniestralidad Total) = E(CLAIM_FREQ) X E(OLDCLAIM)

Por la Ley de los Grandes Números:

lim $_{n\to\infty}$ Prob [| $\overline{S_n} - E(S)$ | > ε] = 0, $\forall$ ε >0

Esto viene a decir, que la prima pura toma un valor muy cercano al valor esperado de la siniestralidad total.

Aplicando este método a los datos del estudio y utilizando el Modelo Hurdle Poisson, se obtienen los siguientes resultados relativos a la frecuencia y cuantías de siniestros, y las primas puras correspondientes a cada clúster:

Tablas 18: Primas Puras y sus componentes

	Clúster 2	Clúster 4	Clúster 6	Clúster 8
E(CLAIM_FREQ)	1,050152	0,2667637	0,9291042	0,7943951
E(OLDCLAIM)	5.395,53	1.671,77	4.790,73	3.995,47
Prima Pura 5 años	5.666,13	445,97	4.451,08	3.173,98
Prima Pura Anual	1.133,23	89,182	890,22	634,80

Fuente: elaboración propia

Se observa que los miembros del clúster número 2 son los que tendrían que pagar una prima pura superior, pues se corresponden con aquellos que poseen valores más elevados de frecuencia y cuantías de siniestros. Estos asegurados, según la clasificación de clúster, circulan en un ambiente urbano y poseen niveles de ingresos intermedios y educación media baja. Se podrían afirmar, que están expuestos a un mayor nivel de riesgo al circular en ciudades, poseer posiblemente coches de gama media y ser más temerarios.

Por otro lado, los asegurados pertenecientes al clúster 4 pagarían la menor prima pura. Podría deberse a que a pesar de conducir coches de menor calidad y por tanto menos seguros, al circular por entornos rurales el riesgo de siniestro es mucho menor.

Por último, los miembros de los clústeres 6 y 8 pagarían primas puras intermedias. Estos individuos circulan en ciudades, por lo que el riesgo es mayor al igual que los del clúster 2, sin embargo, sus niveles de educación son superiores, por lo tanto, se puede pensar que conducen de forma más cuidadosa y menos temeraria.

7. Prima Pura: Caso particular

Como caso particular, se ha calculado la prima pura modelizando la frecuencia de siniestros para la totalidad de la base de datos. Utilizando también el Modelo Hurdle Poisson, sin embargo, se han incluido un mayor número de variables explicativas: KIDSDRIV, AGE, HOMEKIDS, YOJ_1, PARENT1, MSTATUS, GENDER, EDUCATION, TRAVTIME, CAR_USE, CAR_TYPE, RED_CAR, REVOKED, MVR_PTS, CAR_AGE_1, CLAIM_FLAG, TIF y URBANICITY.

De la misma forma que en el apartado anterior, se estimó la frecuencia de siniestros:

Tabla 19: Frecuencia de siniestros estimada para toda la base de datos

Nº de siniestros	0	1	2	3	4	5
Valores Observados	5057	1046	1178	783	182	17
Valores Estimados (HURDLE POISSON)	3.938,80	2.692,92	1.158,27	362,9598	88,7455	17,7773

Fuente: elaboración propia

La diferencia entre los valores reales y los estimados sigue siendo elevada, al igual que ocurría cuando se estimaron para cada clúster, aunque para 2 y 5 siniestros se ajusta bastante bien.

La esperanza de la frecuencia de siniestros en cinco años asciende a 0,7920856 para toda la base de datos.

Posteriormente, partiendo de los valores de la esperanza matemática de las cuantías de los siniestros para cada clúster (utilizados previamente) se multiplican por la esperanza de la frecuencia de siniestros total y se calculan las primas puras correspondientes:

Tabla 20: Primas Puras, caso particular

	Clúster 2	Clúster 4	Clúster 6	Clúster 8
Prima Pura 5 años	4.273,72	1.324,18	3.794,67	3.164,75
Prima Pura Anual	854,74	264,84	758,93	632,95

Fuente: elaboración propia

De esta forma, y debido a que no se han considerado grupos a la hora de estimar la frecuencia de siniestros, las primas puras obtenidas son inferiores a las anteriores para los grupos que pagan primas más elevadas, y superiores para el caso del clúster 4. Esto ocurre debido a la gran heterogeneidad de la muestra. Tal y como se comentó inicialmente, la formación de grupos es fundamental en tarificación. Si esta división no se produce, conllevaría a un problema de Selección Adversa, haciendo que los asegurados que paguen una prima demasiado elevada, teniendo en cuenta el riesgo al que están expuestos, abandonen la cartera, y los demás que pagan una inferior a la que les corresponden, puedan originar problemas de solvencia a la compañía.

8. Conclusiones y posibles líneas de investigación

Este estudio presenta de forma simplificada la importancia de la formación de clústeres a la hora de tarificar en seguros de no vida, como es el caso del seguro automóvil.

Partiendo de una base de datos muy heterogénea, es importante seleccionar variables que sean significativamente discriminatorias. En este caso, utilizando el nivel de ingresos, educación y el tipo de zona por donde circulan los conductores para la formación de los clústeres, a pesar de que no se cumplieran todos los requisitos del clúster bietápico, los grupos son suficientemente homogéneos.

Por otro lado, un problema que presentan los datos del estudio es el inflado de ceros. Al estimar la frecuencia de siniestros, la diferenciación entre declarar cero siniestros o más es muy relevante. En este análisis, dicha estimación presenta diferencias significativas con los valores reales en los cuatro tipos de modelos utilizados. Esto se debe posiblemente a que se tendría que haber realizado un estudio más exhaustivo aplicando una distribución de probabilidad para la frecuencia de ceros, otra para los valores intermedios, y por último, una para los valores más elevados, obteniendo así un modelo compuesto.

En lo que respecta a la prima pura de los clústeres estudiados, los resultados se ajustan correctamente a la realidad, debido en gran medida a que se utilizan valores observados y no los estimados de las cuantías pagadas por los siniestros. Una posible línea de investigación sería estimar, a su vez, las cuantías que han pagado los asegurados en los últimos cinco años por los siniestros que hayan sufrido.

Por último, cabe mencionar que este tipo de procedimiento para el cálculo de la prima pura de la garantía de daños propios, puede ajustarse como un método de tarificación adecuado para una compañía aseguradora pequeña, pues una gran empresa de seguros necesitaría modelos mucho más elaborados.

9. Bibliografía

- Boj del Val, E.; Claramunt Bielsa, M. M.; Fortiana Gregori, J.; Vegas Montaner, A. (2005). *Base de datos y estadísticas del seguro de automóviles en España: influencia en el cálculo de primas.*

- Boucher, J. P.; Guillén, M. (2009). *Una revisión de los modelos para paneles de datos de enumeración con aplicaciones a seguros*. Matemática Aplicada. RACSAM.

- Bueno Blanco, M. (2015). *Modelos Lineales Generalizados, Modelos Inflados de ceros y Modelo Hurdle.* Trabajo fin de máster. Universidad Complutense de Madrid. Dirigido por Dr. Antonio Heras.

- Caballero Díaz, F. (2011). *Selección de modelos mediante criterios de información* en análisis factorial. Aspectos teóricos y computacionales. Tesis Doctoral. España: Universidad de Granada. Disponible en: https://hera.ugr.es/tesisugr/19964808.pdf

- Gorgas García, F.J.; Cardiel López, N.; Zamorano Calvo, J. (2009). *Estadística básica para estudiantes de ciencias.* España: Universidad Complutense de Madrid. Disponible en: http://pendientedemigracion.ucm.es/info/Astrof/users/jaz/ESTADISTICA/libro_GCZ2009.pdf

- Lambert, D. (1992). Zero-Inflated Regresion, with an Application to Defects in Manufacturing*. Technometrics*. Disponible en: https://www.jstor.org/stable/1269547?seq=1

- Medina, F.; Galván, M. (2007). *Imputación de datos: teoría y práctica.* Naciones Unidas. Disponible en: http://repositorio.cepal.org/bitstream/handle/11362/4755/S0700590_es.pdf

- NCSS Statistical Software. *The Zero-Inflated Negative Binomial Regression Model*.
  Disponible en:
  https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Zero-Inflated_Negative_Binomial_Regression.pdf

- Otero García, D. (2011). *Imputación de datos faltantes en un Sistema de Información sobre Conductas de Riesgo.* Trabajo Fin de Máster. Universidad de Santiago de Compostela. Disponible en:
  http://eio.usc.es/pub/mte/descargas/ProyectosFinMaster/Proyecto_616.pdf

- Pardo, A.; Ruiz, M.A. (2005). *Análisis de Datos con SPSS 13 Base*. España: McGraw-Hill España.

- Peña, D. 1 (2002). *Análisis de Datos Multivariantes.* Madrid: McGraw-Hill.

- Pérez Cuellos, F. (2017). *Seguro de autos. Evolución y perspectivas para el futuro*. Willis Towers Watson. Disponible en:
  https://semanadelseguro.inese.es/2017/wp-content/uploads/2016/12/Seguro-de-Autos_2017.pdf

- Pérez López, C. (2005). *Métodos estadísticos avanzados con SPSS*. Madrid: Thomson Paraninfo.

- Piet de Jong; Z. Heller, G. (2008). *Generalized Linear Models for Insurance Data*. Nueva York: Cambridge University Press.

- Rubin, D. B. (1976). *Inference and missing data*. Biometrika Trust. Disponible en: http://www.stat.cmu.edu/~fienberg/Statistics36-756/Rubin-Biometrika-1976.pdf

- Rubio-Hurtado, M.J., y Vilà-Baños, R. (2017). *El análisis de conglomerados bietápico o en dos fases con SPSS*. REIRE. Revista d'Innovació i Recerca en Educació, 10(1), 118-126. Disponible en: http://revistes.ub.edu/index.php/REIRE/article/viewFile/reire2017.10.11017/20151

- Sánchez Meca, J.; Ato García, M.; López Pina, J.A. et al (1989). *Estadística Exploratoria y Confirmatoria con el Paquete Systat.* Universidad de Murcia. Disponible en: https://books.google.es/books?id=skZNyUe4giQC&pg=PA93&lpg=PA93&dq=transformaciones+tukey&source=bl&ots=ar6iw1XyFv&sig=3AbPeeEcaM9wmLPJBrB2HYRx56s&hl=es&sa=X&ved=0ahUKEwirhsmRgobVAhVDExoKHahGCQc4ChDoAQghMAA#v=onepage&q=transformaciones%20tukey&f=false

- Sarabia Alegría, J. M.; Gómez Déniz, Emilio; Vázquez Polo, F. J. (2007). *Estadística Actuarial. Teoría y Aplicaciones.* Pearson Education, SA.

- Velasco Vázquez, M. L. (2008). *Un Modelo de Regresión Poisson Inflado con Ceros para Analizar datos de un Experimento de Fungicidas en Jitomate*. Tesis Doctoral. Universidad Verocruzana. Disponible en:

  https://www.uv.mx/mapli/files/2012/05/Un-Modelo-de-Regresion-Poisson-Inflado-con-Ceros-para-Analizar-Datos-de-un-experimento-de-Fungicidas-en-Jitomate.pdf

- Vegas Asencio, J. (1998). *Algunos aspectos actuariales que surgen en las aplicaciones del reglamento de ordenación y supervisión de los seguros privados.* Disponible en: https://www.fundacionmapfre.org/documentacion/publico/es/catalogo_imagenes/grupo.cmd?path=1054573

- Zhang, H.; Gutiérrez, H. (2010). *Teoría Estadística: Aplicaciones y Métodos.* Universidad Santo Tomás. Disponible en:

  https://books.google.es/books?id=62u0U46_QLsC&pg=PA407&lpg=PA407&dq=box+cox+teoria+transformaciones&source=bl&ots=ZPyF_IyBNl&sig=fPUZMHCONskuvRLm4-U_x5vlx3s&hl=es&sa=X&ved=0ahUKEwiCkOOj_YXVAhUDcBoKHbqwAHkQ6AEIVDAI#v=onepage&q=box%20cox%20teoria%20transformaciones&f=false

10. Anexos

A continuación, se puede observar el código de programación R utilizado en la formulación del cálculo de las primas puras.

Se enseña el código para el clúster 2, siendo el procedimiento similar para los otros tres grupos. Asimismo, para salvaguardar la autoría del proyecto se encuentra omitida alguna parte del mismo.

```
##################Base de Datos
library(xlsx)
library(car)
library(MASS)
library(pscl)
BD80
#####################Composición Clusteres
table(BD80$CLUSTER)
############################Cluster 2
datosc2=subset(BD80,CLUSTER>1&CLUSTER<3)
summary(datosc2)
str(datosc2)
summary(datosc2$INCOME_1)
table(datosc2$EDUCATION)
table(datosc2$URBANICITY)
summary(datosc2$CLM_FREQ)
CLM_FREQ2<-factor(datosc2$CLM_FREQ,labels=c(0,1,2,3,4,5))
table(datosc2$CLM_FREQ)
FAGE2<-cut(datosc2$AGE_1,breaks=c(16,25,35,45,55,81),include.lowest=T)
```

```
##############################ZIP

ModeloZIP2<-
zeroinfl(datosc2$CLM_FREQ~FAGE2+datosc2$CLAIM_FLAG+datosc2$MVR_
PTS,offset=log(datosc2$DURATION),dist ="poisson")

summary(ModeloZIP2)

AIC(ModeloZIP2)


##############################ZIBN

ModeloZIBN2<-
zeroinfl(datosc2$CLM_FREQ~FAGE2+datosc2$CLAIM_FLAG+datosc2$MVR_
PTS,offset=log(datosc2$DURATION),dist ="negbin")

summary(ModeloZIBN2)

AIC(ModeloZIBN2)


########################HURDLE POISSON

ModeloHurdleP2<-
hurdle(datosc2$CLM_FREQ~FAGE2+datosc2$CLAIM_FLAG+datosc2$MVR_
PTS,offset=log(datosc2$DURATION),dist ="poisson")

summary(ModeloHurdleP2)

AIC(ModeloHurdleP2)

########################HURDLE BN

ModeloHurdleBN2<-
hurdle(datosc2$CLM_FREQ~FAGE2+datosc2$CLAIM_FLAG+datosc2$MVR_
PTS,offset=log(datosc2$DURATION),dist ="negbin")

summary(ModeloHurdleBN2)

AIC(ModeloHurdleBN2)
```

```
################################Claim Frequency
ECF2<-sum(ValoresHurdleP2*c(0,1,2,3,4,5))/sum(ValoresHurdleP2)
ECF4<-sum(ValoresHurdleP4*c(0,1,2,3,4))/sum(ValoresHurdleP4)
ECF6<-sum(ValoresHurdleP6*c(0,1,2,3,4,5))/sum(ValoresHurdleP6)
ECF8<-sum(ValoresHurdleP8*c(0,1,2,3,4,5))/sum(ValoresHurdleP8)

########################################Prima Pura
PrimaPura2<- mean(datosc2$OLDCLAIM)*ECF2
PrimaPura4<- mean(datosc4$OLDCLAIM)*ECF4
PrimaPura6<- mean(datosc6$OLDCLAIM)*ECF6
PrimaPura8<- mean(datosc8$OLDCLAIM)*ECF8
data.frame(PrimaPura2,PrimaPura4,PrimaPura6,PrimaPura8)



################################################Sin Clusteres
FAGE<-cut(BD80$AGE_1,breaks=c(16,25,35,45,55,81),include.lowest=T)
ModeloHurdleP<-
hurdle(BD80$CLM_FREQ~BD80$KIDSDRIV+FAGE+BD80$HOMEKIDS+BD80
$YOJ_1+BD80$PARENT1+BD80$MSTATUS+BD80$GENDER+BD80$EDUCA
TION+BD80$TRAVTIME+BD80$CAR_USE+BD80$CAR_TYPE+BD80$RED_C
AR+BD80$REVOKED+BD80$MVR_PTS+BD80$CAR_AGE_1+BD80$CLAIM_
FLAG+BD80$TIF+BD80$URBANICITY,offset=log(BD80$DURATION),dist
="poisson")


summary(ModeloHurdleP)


AIC(ModeloHurdleP)
```

```
ECF<-sum(ValoresHurdleP*c(0,1,2,3,4,5))/sum(ValoresHurdleP)

PP2<-mean(datosc2$OLDCLAIM)*ECF
PP4<- mean(datosc4$OLDCLAIM)*ECF
PP6<- mean(datosc6$OLDCLAIM)*ECF
PP8<- mean(datosc8$OLDCLAIM)*ECF

data.frame(PP2,PP4,PP6,PP8)
```