MODELIZACIÓN DE LA PROBABILIDAD DE SINIESTROS Y ACCIDENTES DE AUTOMÓVIL

AUTOR: GIMÉNEZ BERTRAN, ORIOL

TUTORA: VAYÀ VALCARCE, ESTHER

GRADO EN ECONOMÍA 2020-2021

RESUMEN

El trabajo consiste en la modelización econométrica de siniestros y accidentes de automóvil. Las bases de datos usadas para su elaboración hacen referencia a un tiempo y espacio distintos, pero debido a la disposición de elevados números de observaciones podemos obtener conclusiones capaces de englobar cualquier población a nivel mundial. Al acabar el trabajo, el lector será capaz de tener una clara idea sobre los perfiles de conductor/a que son potencialmente más propensos a tener un siniestro o accidente grave en las carreteras, en función de características tanto internas como externas de la persona o el vehículo.

PALABRAS CLAVE

Siniestro de automóvil, accidente automovilístico, observación, variable estadísticamente significativa, variable cuantitativa, variable cualitativa, categoría base.

ABSTRACT

The work consists of the econometric modeling of car claims and accidents. The databases used for its elaboration refer to a different time and space, but due to the availability of high numbers of observations we can obtain conclusions capable of encompassing any population worldwide. At the end of the work, the reader will be able to have a clear idea about the driver profiles that are potentially more likely to have a loss or serious accident on the roads, depending on both internal and external characteristics of the person or the vehicle.

KEYWORDS

Car claim, car accident, observation, statistically significant variable, quantitative variable, qualitative variable, base category.

ÍNDICE

I. INTRODUCCIÓN

La elección, como trabajo de final de grado, de una temática relacionada con los siniestros y accidentes de automóvil debe su origen a la incorporación, en modalidad de prácticas universitarias, del autor a una empresa aseguradora nacional desde finales del año 2020 y en vigor actualmente. Al realizar la asignatura de prácticas externas a la vez que el Trabajo de Final de Grado, tanto la tutora como el autor decidieron que podría ser de interés tratar un tema relacionado con las mencionadas prácticas.

Los objetivos generales de este trabajo, planteados inicialmente, son modelizar la probabilidad de siniestro y accidente automovilístico con la intención de establecer y detectar qué perfil de conductor/a es más propenso a los hechos descritos, así como indagar en las características externas al individuo que condicionan esta probabilidad.

En un primer momento, el autor intentó solicitar a su empresa información a nivel micro de la cartera de clientes, pero la petición fue denegada por secreto administrativo y confidencialidad de los datos corporativos. Siguiendo con la búsqueda de bases de datos, nos dirigimos a otras entidades relacionadas con el sector en cuestión, nuevamente negando su participación argumentado lo mismo. Debido a dichas negativas, tuvimos que buscar otras bases de datos de acceso al público, encontrando dos que estaban estrechamente relacionadas con el tema. La primera hace referencia a un conjunto de personas tomadoras de una póliza de seguro de automóvil entre los años 2004-2005 en Australia. Esta base de datos es bastante parecida a la que se hubiera obtenido de la empresa donde trabaja el autor porque da información del tipo de vehículo y del tomador/a de la póliza. La segunda muestra fue elegida porque integra un amplio conjunto de accidentes automovilísticos en España durante el 2018, de esta manera tendremos información más cercana y reciente. Se optó por trabajar con ella ya que, a diferencia de la otra, esta nos brinda información para un periodo más actualizado y nos detalla especialmente la hora, el mes, junto con la tipología de accidente, carretera y vehículo donde el accidente fue dado. Para ambos objetivos mencionados, se tendrán en consideración perfiles de conductores muy diferentes debido al gran número de observaciones que tienen las muestras analizadas. Por lo tanto, trataremos con dos bases de datos, cada una protagonista de un capítulo del trabajo.

La metodología del trabajo se sostiene en el tratamiento de datos mediante la utilización del programa econométrico ‘Gretl’, estudiado a lo largo del grado universitario. La información necesaria para exponer los resultados que este nos imprimía, se obtuvo a partir de conocimientos propios del autor, explicaciones en internet y muy especialmente, documentos realizados por el departamento de Econometría, Estadística y Economía Aplicada para impartir las clases universitarias proporcionados por la tutora del trabajo.

Finalmente, para acabar con este capítulo, me gustaría mandarle un mensaje de gratitud a mi tutora Esther Vayà. Desde el momento en que contacté con ella para transmitirle

mi voluntad de que me dirigiera el trabajo hasta el día de la entrega, ha estado corrigiendo, ayudando y animándome durante el proceso, siempre disponible para cualquier duda. Sin embargo, esto no se limita aquí, ya que como profesora de las distintas asignaturas universitarias de econometría respondió de la misma buena manera.

II. SINIESTRO DE AUTOMÓVIL

Tal y como se ha comentado en el capítulo introductorio, la primera base de datos seleccionada hace referencia a pólizas de seguro de vehículos durante los años 2004-2005 en Australia. En primer lugar, iniciamos el capítulo con una descripción de la base de datos utilizada, para modelizar la probabilidad de siniestro de automóvil, haciendo una explicación de lo que refleja cada una de las variables disponibles de la muestra tratada. Seguidamente, realizamos un proceso de análisis estadístico-descriptivo. De esta manera, visualizamos cómo se comportan tanto las variables dependientes como potenciales variables explicativas.

Una vez hecho todo esto, plantearemos dos modelizaciones diferentes. En primer lugar, construiremos un Modelo de Regresión Lineal Múltiple (MRLM) con el objetivo de conocer los determinantes del número de siniestros que sufre el tomador del seguro en el espacio y tiempo analizados. En segundo lugar, construiremos un Modelo de Probabilidad logístico que nos permitirá conocer los factores que afectan a la probabilidad de sufrir un siniestro (y no el número de siniestros como en el caso del MRLM).

2.1 Descripción de la base de datos

La base de datos usada para la realización de este capítulo hace referencia a pólizas de seguro de vehículos durante los años 2004-2005 en Australia. En concreto, hay 67.856 pólizas o, lo que es lo mismo, observaciones. Las variables del modelo que disponemos son las descritas a continuación:

1. ClaimNb: Número de siniestros que el tomador del seguro ha tenido durante el periodo de tiempo estudiado.
2. ClaimOcc: Variable dicotómica que toma valor 1 si el tomador observado ha sufrido al menos un siniestro y 0 en caso contrario.
3. Exposure: Tiempo que la póliza de seguro tuvo efecto. Los datos están comprendidos entre el valor 0 (la póliza no tuvo ningún día de efecto, descartando este valor absoluto porque todas las observaciones tienen un valor para esta variable superior a 0) y 1 (la póliza de seguro tuvo efecto durante todo un año).
4. VehValue: Valor del vehículo en miles de dólares australianos[1], en el momento de tiempo estudiado.
5. VehAge: Edad del vehículo. Definida para las categorías; 'VehMasNuevo', 'VehNuevo ', 'VehViejo' y 'VehMasViejo'.

[1] 1 dólar australiano (AUD) aproximadamente equivale a 0,62 euros (€) en la actualidad.

6. VehBody: Tipo de vehículo. Definida para las categorías; ‘station wagon’, ’sedan’, ’utility’, ’hatchback’, ’minibus’, ’hardtop’, ’coupe’, ’convertible’, ’truck’, ’bus’, ‘motorized caravan’, ’panel van’ y ’roadster’.
7. Gender: Género (Mujer/Hombre) del tomador de la póliza.
8. DrivAge: Edad del tomador de la póliza. Definida para las categorías; ‘PersMasJoven’, ‘PersJoven’, ‘PersTrabaj’, ‘PersTrabajMayor’, ‘PersMayor’, ‘PersMasMayor’.
9. ClaimAmount: Cantidad de dinero indemnizada por la compañía aseguradora al tomador de la póliza (en dólares australianos).

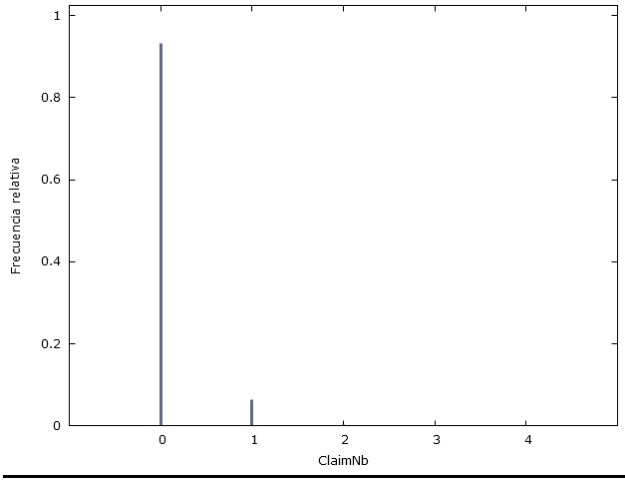
Las variables que serán modelizadas econométricamente son ClaimOcc y ClaimNb. El resto de las variables serán utilizadas como explicativas de las anteriores.

Tal y como puede observarse, hay ciertas variables explicativas que son cualitativas, es decir, no están formadas por valores numéricos sino por categorías. Para poder tratar con ellas, tuvimos que transformar dichas variables para convertirlas en variables ficticias. Tomamos de ejemplo Gender. Esta variable dispone de dos categorías (Mujer y Hombre) a partir de las cuales creamos primero una variable ficticia llamada ‘Mujer’ donde aparece el valor 1 para cada observación que sea una mujer y 0 si es un hombre. De forma similar, creamos su complementaria, la variable ficticia ‘Hombre’ donde la observación toma valor 1 si es un hombre y 0 si es mujer. Este mismo procedimiento lo hacemos para cada categoría de cada variable ficticia del modelo.

2.2 Análisis descriptivo

En primer lugar, elaboramos un histograma y la distribución de frecuencias para las dos variables endógenas del modelo: ClaimNb y ClaimOcc (gráficos 2.1 y 2.2 respectivamente).

Gráfico 2.1. Histograma de la variable ClaimNb



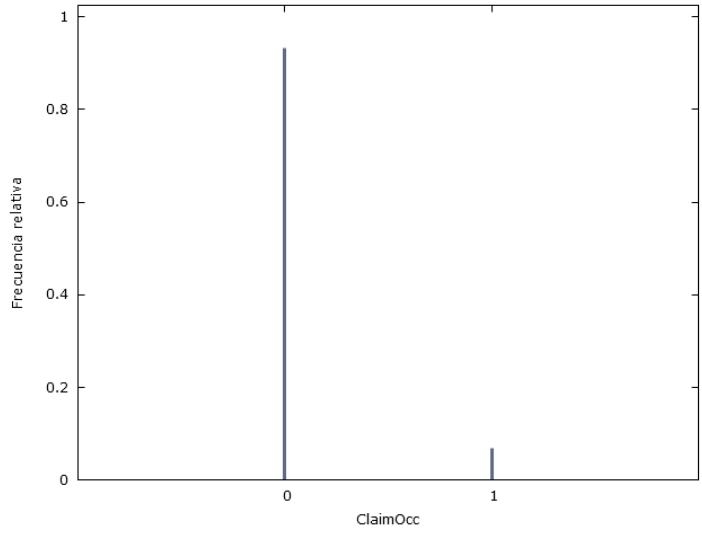
	frecuencia	rel.	acum.	
0	63232	93.19%	93.19%	**********
1	4333	6.39%	99.57%	**
2	271	0.40%	99.97%	
3	18	0.03%	100.00%	
4	2	0.00%	100.00%	

Fuente: Elaboración propia

Revisando los resultados de ClaimNb y fijándonos en la frecuencia relativa, concluimos que el 93,19% de los individuos del modelo no tuvieron ningún siniestro a lo largo del año. Seguidamente, el número de personas que tuvieron únicamente un siniestro fue de 4.333, representando un 6,39% del total de los datos. Para terminar, sufrieron dos, tres o cuatro siniestros, 271, 18 y 2 personas respectivamente, esto hace que la frecuencia relativa sea muy cercana a 0%.

En el caso de la variable ClaimOcc (variable dicotómica), los resultados son similares: el 93,19% de los individuos no tuvieron siniestro alguno (exactamente 63.232 personas), mientras que el 6,81% restante sí se vieron implicados en uno o más.

Gráfico 2.2. Histograma de la variable ClaimOcc



	frecuencia	rel.	acum.	
0	63232	93.19%	93.19%	**********
1	4624	6.81%	100.00%	*
2				

Fuente: Elaboración propia

A continuación, presentamos los estadísticos principales de las principales variables cuantitativas del modelo.

En primer lugar, se observa que la media de siniestros que tuvo el grupo de individuos fue de 0,0728. Podemos decir entonces que la gran mayoría de observaciones no

tuvieron ningún siniestro, al ser prácticamente un valor nulo. En cambio, para aquellos que sí han sufrido algún tipo de siniestro, el número máximo fue cuatro.

Con relación a la variable anterior, aquellas personas que efectivamente sí habían tenido un siniestro o más fueron indemnizadas por una cantidad media de 137,3 AUD (87€) y una cantidad máxima de 55.922 AUD (35.450€).

Para la variable Exposure, apreciamos que la media tiene un valor de 0,469, es decir, el conjunto de observaciones tuvo de promedio una póliza de seguro activa durante unos 171 días aproximadamente.
Destacamos de la variable VehValue que el precio medio del conjunto de vehículos de los individuos en nuestra base de datos fue de 1.780 AUD, unos 1.130 €. Asimismo, puede resultar interesante analizar el valor máximo de entre todos los vehículos observados, siendo éste de 34.600 AUD (21.935€).

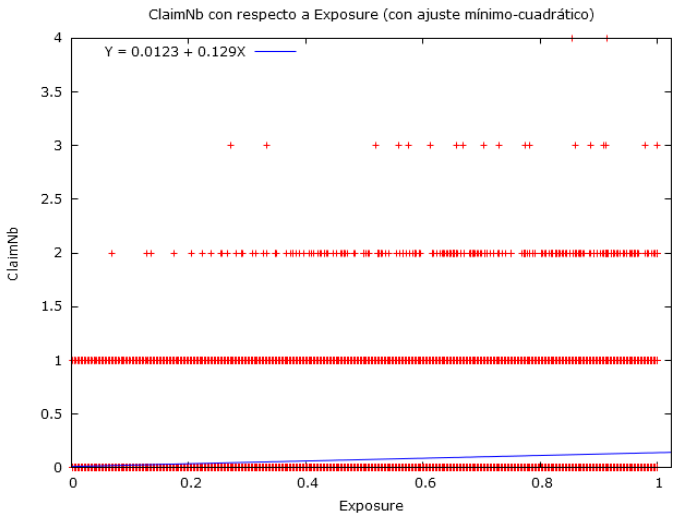
Tabla 2.1. Estadísticos principales de las variables cuantitativas (N=67856)

Variable	Media	Mediana	D. T.	Mín	Máx
ClaimNb	0.0728	0.00	0.278	0.00	4.00
Exposure	0.469	0.446	0.290	0.00274	0.9993
VehValue	1.78	1.50	1.21	0.00	34.6
ClaimAmount	137.3	0.00	1056	0.00	55922

Fuente: Elaboración propia

Seguidamente, procedemos a estudiar la posible relación entre las variables cuantitativas anteriores y nuestra variable endógena. Así, como se observa en el gráfico 2.3, parece existir una relación lineal significativa entre las variables ClaimNB y Exposure, implicando que, a medida que la observación pasa más tiempo asegurada y por lo tanto más tiempo analizada, el número de siniestros aumenta.

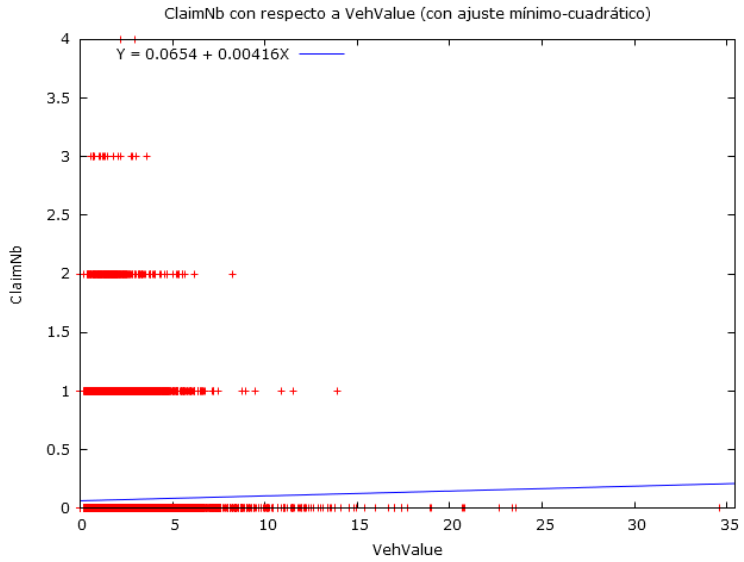
Gráfico 2.3. Gráfico de dispersión entre ClaimNB y Exposure.



Fuente: Elaboración propia

Estudiando el gráfico de dispersión entre las ClaimNb y VehValue (gráfico 2.4), obtenemos conclusiones muy parecidas a las descritas anteriormente. Nuevamente existe una relación lineal significativa entre las variables en cuestión, con pendiente positiva. Por lo tanto, podemos decir que, al incrementar el valor del vehículo, también lo hace el número de siniestros observados.

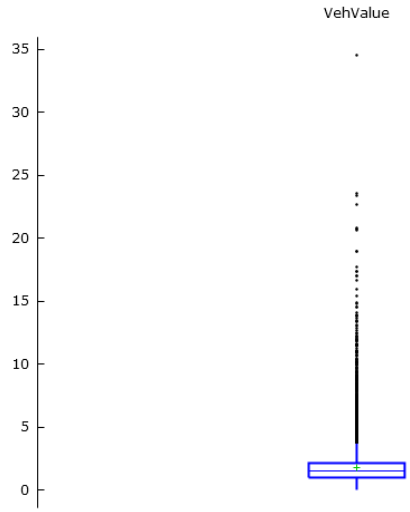
Gráfico 2.4. Gráfico de dispersión entre ClaimNB y VehValue



Fuente: Elaboración propia

Relacionado con el gráfico anterior, dado que se han observado algunos valores muy elevados para la variable VehValue, se ha procedido a realizar un gráfico de caja o box plot para esta variable (gráfico 2.5).

Gráfico 2.5. Gráfico de caja (box-plot) de la variable VehValue



Fuente: Elaboración propia

Como se puede observar, el valor del primer cuartil es 1,01; el segundo cuartil o mediana 1,50; el tercero 2,15 y el último 3,85. En este caso, el gráfico nos permite ver un conjunto de observaciones fuera de la caja y muy alejadas de ella, indicando la posible existencia de observaciones atípicas que puedan llegar a tener una influencia real en el modelo posterior (por ello, se procederá posteriormente a chequear la posible existencia de observaciones atípicas o con influencia real en la estimación).

A su vez, se ha analizado la matriz de correlaciones entre las variables cuantitativas anteriores. Como se puede observar, la variable ClaimAmount presenta una correlación positiva y significativa con las variables VehValue y Exposure (correlación superior al valor crítico) si bien no es elevada (una correlación muy elevada entre variables podría llevarnos a un problema de multicolinealidad grave en el modelo posterior).

Tabla 2.2. Matriz de correlación entre variables independientes (n=67856)

Coeficientes de correlación, usando las observaciones 1 - 67856
Valor crítico al 5% (a dos colas) = 0.0075 para n = 67856

Exposure	VehValue	ClaimAmount	
1.0000	-0.0006	0.0365	Exposure
	1.0000	0.0098	VehValue
		1.0000	ClaimAmount

Fuente: Elaboración propia

Por último, procedemos a analizar las variables ficticias construidas con la intención de contrastar si, especialmente el número de siniestros (ClaimNb) presenta medias significativamente diferentes para las diferentes categorías de dichas variables ficticias. Trabajaremos a partir de las cuatro variables cualitativas que disponemos: VehAge, DrivAge, Gender y VehBody. Los resultados se presentan en los cuadros comprendidos entre el 2.1 y 2.4.

Empezamos haciendo el contraste para la variable Gender (Cuadro 2.1), estimando una regresión en la que la variable ClaimNb se regresa frente a una constante y la variable Hombre (variable que tiene un 1 si el individuo es hombre y 0 en caso contrario).

Cuadro 2.1. Contraste de medias de la variable ClaimNb frente a la variable Gender

MCO, usando las observaciones 1-67856
Variable dependiente: ClaimNb

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.0733622	0.00141597	51.81	<0.0001	***
Hombre	−0.00140374	0.00215657	−0.6509	0.5151	

Fuente: Elaboración propia

Si nos fijamos en el coeficiente asociado a la constante, este refleja la media de siniestros para las mujeres, porque estas actúan como categoría base. Por su parte, el coeficiente asociado a Hombre indica que la media de siniestros para los hombres es 0,00140374 menor que la media de las mujeres. Adicionalmente, decimos que la media de los individuos que son hombres no es estadísticamente distinta respecto a la media de las mujeres porque el valor p de 0,5151 supera claramente a 0,005. Asumiendo que el género de los conductores tiene un mismo peso en el conjunto de la muestra, se puede apreciar qué este no es condicionante de siniestros automovilísticos.

Replicamos el análisis para la variable cualitativa VehAge (Cuadro 2.2). Ahora, la categoría base son los vehículos más nuevos. Vemos que la media de siniestros que tuvieron dichos vehículos fue de 0,0714694. Analizando la significación de los coeficientes, no se observan diferencias significativas en la media de los siniestros en función de la edad de los vehículos, si bien para que los vehículos jóvenes tienen un número de siniestros superior a la de los más jóvenes (no ocurre lo mismo con los vehículos antiguos o muy antiguos, los cuales no parecen tener una media de siniestro significativamente diferentes).

Cuadro 2.2. Contraste de medias de la variable ClaimNb frente a la variable VehAge

MCO, usando las observaciones 1-67856

Variable dependiente: ClaimNb

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.0714694	0.00251244	28.45	<0.0001	***
VehNuevo	0.0101608	0.00331313	3.067	0.0022	***
VehViejo	0.000600014	0.00318881	0.1882	0.8508	
VehMasViejo	−0.00491881	0.00322423	−1.526	0.1271	

Fuente: Elaboración propia

El cuadro 2.3. muestra los resultados para la variable edad del conductor. Obtenemos unos resultados bastante cautivadores. La media de siniestros para los conductores más jóvenes de la muestra es de 0,0914316. A partir de aquí, se aprecia como a medida que aumenta la edad de los conductores, disminuye el número de siniestros. La afirmación se fundamenta en los coeficientes asociados a cada categoría, que son todos negativos, y estos en valor absoluto más grandes que la categoría anterior. Es lógico pensar que personas con más tiempo de vida tengan mayor experiencia al volante, viéndose implicados en menos siniestros.

Cuadro 2.3. Contraste de medias de la variable ClaimNb frente a la variable DriveAge

MCO, usando las observaciones 1-67856
Variable dependiente: ClaimNb

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.0914316	0.00366971	24.92	<0.0001	***
PersJoven	−0.0137617	0.00441279	−3.119	0.0018	***
PersTrabaj	−0.0160209	0.00428615	−3.738	0.0002	***
PersTrabajMayor	−0.0182337	0.00427121	−4.269	<0.0001	***
PersMayor	−0.0310739	0.00454635	−6.835	<0.0001	***
PersMasMayor	−0.0318623	0.00502769	−6.337	<0.0001	***

Fuente: Elaboración propia

Para terminar con este apartado, replicamos el análisis, pero para la variable tipo de vehículo. De esta manera, los vehículos con una media significativamente menor a la categoría base 'Sedan' son los pertenecientes al grupo 'Utility', de lo contrario los significativamente mayores son 'Station Wagon', 'Hardtop', 'Coupe', 'Bus', 'Motorized Caravan' y 'Panel van'. En este caso, hemos querido poner a propósito a 'Sedan' como categoría base ya que este tipo de automóvil es el más frecuente, con una composición no tan particular como las otras categorías.

Cuadro 2.4. Contraste de medias de la variable ClaimNb frente a la variable VehBody

MCO, usando las observaciones 1-67856
Variable dependiente: ClaimNb

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.0718751	0.00186535	38.53	<0.0001	***
StationWagon	0.00487291	0.00287001	1.698	0.0895	*
Utility	−0.0116920	0.00451092	−2.592	0.0095	***
HatchBack	−0.00156058	0.00275126	−0.5672	0.5706	
MiniBus	−0.00911363	0.0105534	−0.8636	0.3878	
Hardtop	0.0142553	0.00724382	1.968	0.0491	**
Coupe	0.0242787	0.0101321	2.396	0.0166	**
Convertible	−0.0348381	0.0309604	−1.125	0.2605	
Truck	0.00241057	0.00690548	0.3491	0.7270	
Bus	0.136458	0.0401890	3.395	0.0007	***
MotorizedCaravan	0.0462351	0.0247511	1.868	0.0618	*
PanelVan	0.0185504	0.0103127	1.799	0.0721	*
Roadster	0.0392360	0.0535601	0.7326	0.4638	

Fuente: Elaboración Propia

2.3. Modelización del Número de Siniestros

Tal y como hemos introducido antes, en esta parte del trabajo pretendemos modelizar qué variables son realmente relevantes para explicar el número de siniestros sufrido para un individuo de la muestra. El modelo inicialmente propuesto es el siguiente:

$$ClaimNb_i = f(Exposure_i, VehValue_i, VehAge_i, VehBody_i, DrivAge_i, Gender_i)$$

En este sentido, cabe decir que las variables cualitativas se han incluido en el modelo de regresión por medio de las variables ficticias construidas a partir de ellas.

Tendremos en consideración todas las observaciones de nuestra base de datos, concretamente 67.856 individuos. La estimación de los parámetros poblacionales en nuestro MRLM se hará a través del método de Mínimos Cuadrados Ordinarios (MCO). Los resultados analíticos se presentan en el Cuadro 2.5 a continuación.

Cuadro 2.5. Modelización de la variable ClaimNb

MCO, usando las observaciones 1-67856
Variable dependiente: ClaimNb
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.0199112	0.00824944	2.414	0.0158	**
Exposure	0.129474	0.00380109	34.06	<0.0001	***
VehValue	0.00777651	0.00235264	3.305	0.0009	***
sq_VehValue	−0.00057387	0.000178738	−3.211	0.0013	***
VehNuevo	0.00651009	0.00343816	1.893	0.0583	*
VehViejo	0.000245435	0.00370909	0.06617	0.9472	
VehMasViejo	−0.00068425	0.00466423	−0.1467	0.8834	
Sedan	0.00270004	0.00349056	0.7735	0.4392	
Utility	−0.0128526	0.00443860	−2.896	0.0038	***
Hatchback	0.000556769	0.00394795	0.1410	0.8878	
MiniBus	−0.00520767	0.00964008	−0.5402	0.5891	
HardTop	0.00824561	0.00764998	1.078	0.2811	
Coupe	0.0322751	0.0117099	2.756	0.0058	***
Convertible	−0.0250328	0.0232101	−1.079	0.2808	
Truck	−0.00255593	0.00718803	−0.3556	0.7222	
Bus	0.126958	0.0644670	1.969	0.0489	**
MotorizedCaravan	0.0485665	0.0303259	1.601	0.1093	
PanelVan	0.00941546	0.0118083	0.7974	0.4252	
Roadster	0.0301035	0.0785682	0.3832	0.7016	
PersJoven	−0.0143338	0.00474049	−3.024	0.0025	***
PersTrabaj	−0.0183772	0.00461056	−3.986	<0.0001	***
PersTrabajMayor	−0.0204694	0.00456178	−4.487	<0.0001	***
PersMayor	−0.0350196	0.00469326	−7.462	<0.0001	***
PersMasMayor	−0.0335664	0.00511210	−6.566	<0.0001	***
Hombre	−0.00163823	0.00223255	−0.7338	0.4631	

Media de la vble. dep.	0.072757	D.T. de la vble. dep.	0.278204
Suma de cuad. residuos	5144.286	D.T. de la regresión	0.275390
R-cuadrado	0.020472	R-cuadrado corregido	0.020125
F(24, 67831)	53.11977	Valor p (de F)	8.0e-252
Log-verosimilitud	−8766.178	Criterio de Akaike	17582.36
Criterio de Schwarz	17810.48	Crit. de Hannan-Quinn	17652.82

Fuente: Elaboración propia

Primero de todo, debemos remarcar que se ha seleccionado la estimación por Desviaciones Típicas Robustas con el objetivo de minimizar el problema de heteroscedasticidad detectado a partir del contraste de White.

Asimismo, cabe decir que se ha incorporado el cuadrado de la variable VehValue como regresor (sqVehValue). Esta se tuvo que añadir debido a la existencia de un error conocido como error de especificación de la forma funcional que tiene consecuencias negativas sobre los estimadores MCO (serían sesgados e inconsistentes). Nos encontramos en esta situación cuando establecemos una función de relación entre la variable endógena y las independientes que no es la correcta. Para explicar de manera más detallada esta situación, imagínese que queremos modelizar la relación existente entre la variable Ahorro de una persona y la Renta de este. De esta manera, estimamos el modelo por MCO.

$$Ahorro_i = \beta_0 + \beta_1 Renta_i + \varepsilon_i$$

Podemos ver que estamos delante de un modelo lineal ya que, si diésemos valores distintos a la variable Renta obteniendo así unos valores de Ahorro, para luego plasmar los resultados en un gráfico, veríamos que se puede construir una línea recta. Además, haciendo la derivada $\frac{\partial Ahorro}{\partial Renta}$ el resultado es β_1 . Esto nos indicaría que si se incrementa en una unidad la variable Renta, la variable Ahorro lo hace también, pero en β_1 unidades. Hablamos de un incremento porque económicamente se ve demostrado que la relación entre dichas variables es positiva. Ahora bien, podría suceder que en ciertos individuos esta relación no fuese lineal, siendo quizá, cuadrática, cúbica, exponencial, logarítmica… bien, para comprobarlo bastaría con hacer la misma estimación que hemos realizado anteriormente, pero con la variable independiente elevada al cuadrado, cubo, transformada a logaritmos… y observar si la variable nuevamente añadida es significativa para explicar el comportamiento de la endógena. Un ejemplo sería tal que así:

$$Ahorro_i = \beta_0 + \beta_1 Renta_i + \beta_2 Renta_i^2 + \varepsilon_i$$

Si calculamos la derivada $\frac{\partial Ahorro}{\partial Renta}$ obtenemos resultados distintos. El incremento del Ahorro para un incremento de una unidad de Renta deja de ser β_1 para ser ahora $\beta_1 + 2 * \beta_2 * Renta_{i2}$.

En relación con este posible problema de forma funcional no lineal en nuestro modelo, estimamos nuestro modelo principal añadiendo las variables Exposure y VehValue elevadas al cuadrado para intentar detectar dicho error. Los resultados nos indicaron que sq_Exposure no era significativa para explicar el número de siniestros, no incluyéndola en el modelo final. De lo contrario, sí observamos que sq_VehValue era significativa, debiéndose incluir en la estimación para evitarla inconsistencia de las estimaciones derivadas de un error en la forma funcional.

Dicho todo lo anterior, analizando los resultados obtenidos de la estimación se puede concluir que las variables que han resultado ser relevantes a un nivel de significación del 1% son: Exposure, VehValue, sq_VehValue, los vehículos tipo Utility y Coupe y todas las categorías de la variable edad del conductor exceptuando la categoría asociada a 'old people'. Al 5% de significación aparece relevante el tipo Bus. Si flexibilizamos el nivel de significación hasta el 10%, los coches jóvenes también serían explicativos. Si nos centramos en el impacto que tienen estas variables significativas sobre la dependiente vemos que, por un lado, los dos regresores cuantitativos incrementan el número de siniestros estimado, esto se debe al signo positivo de los coeficientes beta estimados. Por ejemplo, si la póliza del seguro hubiese tomado efecto durante todo un año, el número estimado de siniestros seria 0,13 mayor que si de lo contrario, la póliza no hubiese tomado efecto ni un solo día. En el caso del valor del vehículo, dados los signos del efecto marginal, se observa como a medida que aumenta el valor del vehículo, aumenta también el número de siniestros hasta llegar a un mayor máximo a partir del cual disminuyen los mismos (quizás debido a una mayor seguridad de los coches de más alta gamma). Nuevamente se obtiene la misma conclusión en relación a la edad del conductor: a menor edad, mayor número de siniestros. A su vez, resulta interesante destacar que, tal y como hemos visto en el contraste de medias, el género del conductor no es estadísticamente significativo a la hora de determinar el número de siniestros. De forma similar, la gran mayoría de tipos de vehículo no han resultado ser relevantes para explicar el número de siniestros, exceptuando aquello señalados anteriormente.

Para acabar, fijándonos en el R-cuadrado, decimos que el conjunto de variables explicativas, únicamente permiten explicar el 2,05% del comportamiento y variabilidad de la endógena. Esto nos hace pensar que deberíamos indagar en la búsqueda de otras variables externas que puedan ser influyentes en ClaimNb.

Por último, cabe decir que no se ha detectado ninguna observación atípica ni con influencia real sobre el análisis. Si el lector estuviese interesado en saber cómo se ha llegado a esta conclusión, puede consultarlo en el apartado final de Anexos.

2.4 Modelización de la Probabilidad de Siniestro

Seguidamente intentaremos predecir el comportamiento de la variable dependiente dicotómica ClaimOcc. Recordemos que esta variable toma valor 0 si la observación no ha tenido ningún siniestro y 1 en caso de que haya tenido uno o más. De esta manera, mediante el Modelo Logit predeciríamos si un individuo sufriera algún siniestro o no, en

función de características propias a cada uno. La expresión escrita del modelo analizado es

$$Prob\ (Y_i = 1) = \frac{1}{1 + e^{-Z}} = \frac{e^Z}{1 + e^Z}$$

Donde Z depende de las mismas variables que en modelo de regresión analizado en el apartado anterior.

Los resultados de la estimación del modelo lógit planteado se muestran en el cuadro 2.6.

Cuadro 2.6. Modelización de la probabilidad de siniestro

Logit, usando las observaciones 1-67856
Variable dependiente: ClaimOcc
Desviaciones típicas basadas en el Hessiano

	Coeficiente	*Desv. Típica*	*z*	*Pendiente**
const	−3.32118	0.100528	−33.04	
Exposure	1.86204	0.0549388	33.89	0.104737
VehValue	0.0305749	0.0186973	1.635	0.00171979
VehNuevo	0.0701615	0.0483724	1.450	0.00400972
VehViejo	−0.0337587	0.0520799	−0.6482	−0.00188743
VehMasViejo	−0.105478	0.0613702	−1.719	−0.00581421
Sedan	−0.0272896	0.0472530	−0.5775	−0.00152868
Utility	−0.231756	0.0729062	−3.179	−0.0119470
Hatchback	−0.0487249	0.0531750	−0.9163	−0.00271497
Minibus	−0.0781053	0.162571	−0.4804	−0.00424829
Hardtop	0.113878	0.0982140	1.159	0.00672026
Coupe	0.373881	0.132659	2.818	0.0247364
Convertible	−0.811010	0.604080	−1.343	−0.0323542
Truck	−0.0910300	0.101489	−0.8969	−0.00493000
Bus	1.04274	0.378192	2.757	0.0930413
MotorizedCaravan	0.571966	0.291070	1.965	0.0414564
PanelVan	0.0271754	0.139824	0.1944	0.00154657
RoadSter	−0.0195527	0.743258	−0.02631	−0.00109040
PersJoven	−0.205543	0.0590006	−3.484	−0.0109409
PersTrabaj	−0.257642	0.0575200	−4.479	−0.0136605
PersTrabajMayor	−0.291440	0.0573639	−5.081	−0.0153618
PersMayor	−0.509095	0.0637939	−7.980	−0.0247172
PersMasMayor	−0.511685	0.0730008	−7.009	−0.0241315
Hombre	−0.0151263	0.0325418	−0.4648	−0.00085005

Media de la vble. dep.	0.068144	D.T. de la vble. dep.	0.251995
R-cuadrado de McFadden	0.040169	R-cuadrado corregido	0.038747
Log-verosimilitud	−16205.22	Criterio de Akaike	32458.44
Criterio de Schwarz	32677.44	Crit. de Hannan-Quinn	32526.08

*Evaluado en la media
Número de casos 'correctamente predichos' = 63232 (93.2%)
f(beta'x) en la media de las variables independientes = 0.252
Contraste de razón de verosimilitudes: Chi-cuadrado(23) = 1356.36 [0.0000]

		Predicho	
		0	1
Observado	0	63.232	0
	1	4.624	0

Fuente: Elaboración Propia

Atendiendo a la significación de los coeficientes, concluimos que las variables y categorías, en comparación con sus respectivas categorías base, influyentes en la probabilidad de siniestro, son: Exposure, VehMasViejo (a un nivel de significación del 10%), algunos tipos de vehículos en particular y la edad del conductor. Contrariamente, resultan no relevantes a la hora de explicar la probabilidad de siniestro la variable valor del vehículo (resultado que se repite en modelos siguientes) y todas las edades de los vehículos (a excepción de los vehículos más antiguos, que son significativos para determinar la probabilidad de siniestro, y a la vez la disminuye, rechazando entonces la creencia que vehículos nuevos pueden presentar menos probabilidad de siniestros que los viejos). Para terminar, el género del conductor no afecta a la probabilidad de siniestralidad.

Atendiendo al signo de los coeficientes significativos, se observa que, cogiendo de ejemplo el coeficiente asociado a PersMasMayor, afirmamos que este nos indica que si la persona esta categorizada como de las más grandes de edad, la probabilidad que sufra algún siniestro es menor. Esto se debe al signo negativo del coeficiente. Se puede explicar tal vez porqué las personas mayores tienen, como decíamos al principio del trabajo, más experiencia conduciendo o simplemente es que usan menos el vehículo. Esta explicación también sirve para todos los otros coeficientes obtenidos, pero siempre teniendo en cuenta el signo. Un dato que podemos detectar es que una persona del rango de edad más grande tiene 2,5 puntos porcentuales menos de probabilidad de tener un siniestro que la categoría de conductores más jóvenes.

Las pendientes estimadas (que captan los efectos marginales de las variables explicativas) aparecen significativamente positivas indicando que un incremento de los regresores, o en su defecto en el caso de las cualitativas que esa observación haga que tomar valor 1, implica a la vez un aumento significativo de las posibilidades de ocurrencia de siniestro. Al revés, ocurriría para pendientes de los regresores negativos. De ejemplo, los automóviles tipo ‘Utility’ hacen disminuir la probabilidad de siniestro en 0,0119470, mientras que los ‘Bus’ la aumentan en 0,0930413.

Especialmente queremos destacar la variable referente a la edad del conductor. Como hemos dicho antes, todas las edades resultan estadísticamente significativas en comparación con la categoría base para explicar la probabilidad de sinestro, pero debemos añadir que esta probabilidad disminuye a medida que el conductor tiene más años. Esta afirmación se sostiene en el hecho que los coeficientes estimados para cada categoría presentan signos cada vez más negativos, al igual que sus pendientes estimadas.

En los MRLM para evaluar la calidad del modelo construido, lo hacíamos mediante el Coeficiente de Determinación R^2. Debido a que estamos delante de un modelo de regresión logístico de probabilidad, esta evaluación se hace mediante la comparación de valores predichos respecto los observados y el R^2 de McFadden. Nuestro R^2 de McFadden es 0,04, esto implica que el modelo estimado presenta una muy baja capacidad predictiva para la variable dependiente, probablemente debido a que la muestra está muy descompensada. Esta descompensación en la muestra (donde el valor de 0 siniestros es mayoritario) puede explicar que, a pesar de un porcentaje de aciertos muy elevado, no se prediga correctamente ningún caso en el que sí que se ha producido algún siniestro.

2.5. Modelización de la cantidad de dinero indemnizada en caso de siniestro

Hasta ahora hemos trabajado con el conjunto total de observaciones que nos mostraba la base de datos. De lo contrario, en este apartado, tendremos en cuenta únicamente aquellos individuos que, a lo largo del tiempo, padecieron al menos un siniestro. Una vez restringida la población, el número de observaciones pasa a ser de 4.624. Para no hacer algo repetitivo, dejaremos de un lado la estadística descriptiva del modelo, centrándonos en resultados más concluyentes.

Ahora que estamos teniendo en cuenta tan solo esos individuos que tuvieron siniestro, es interesante saber qué variables influyen en ClaimAmount, es decir, la cantidad de dinero que, en caso de siniestro, la persona recibió de la compañía aseguradora (cuadro 2.7). Analizando los coeficientes asociados a las variables regresoras, decimos que resultan significativos sobre el comportamiento de la endógena, tanto el coeficiente asociado a la variable Exposure, que ya ha aparecido significativo en diversos casos anteriormente, como el hecho que el vehículo sea de los más viejos y el conductor sea un hombre, debemos añadir que todas las edades de la persona también se incluyen en este grupo de variables. Profundizando en estos coeficientes influyentes decimos que, si el tiempo que la observación tuvo una póliza de seguro activa fue de un año, es decir 'Exposure' toma valor igual a uno, la cantidad de dinero indemnizada por la compañía aseguradora es de 1779.27 AUD menos que si el tiempo fue de cero días, donde la 'Exposure' tomaría valor de cero y al multiplicar el coeficiente asociado beta por un valor nulo, esta quedaría anulado. Analizando los demás regresores significativos, nos encontramos que, si el vehículo observado es de los más antiguos de la muestra, entonces, VehMasViejo al ser una ficticia toma valor 1 conllevando a una cantidad de dinero indemnizada de 421,21 AUD mayor respecto si el vehículo no cumple esta

característica. Resultados contrarios sacamos si analizamos la edad del conductor. El hecho que el conductor tenga una avanzada edad asocia una indemnización de 595,14 AUD menor que un conductor que pertenece al grupo de edad de los más jóvenes. Para terminar, como decíamos el género del conductor también resulta ser significativo para explicar ClaimAmount. Entonces, si el conductor es un hombre, habrá más dinero indemnizado en caso de accidente por parte de la empresa en comparación a si el conductor es una mujer, concretamente de 346,39 AUD más.

Cuadro 2.7. Modelización de la variable ClaimAmount

MCO, usando las observaciones 1-4624
Variable dependiente: ClaimAmount

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	2950.96	360.371	8.189	<0.0001	***
Exposure	−1779.27	198.303	−8.972	<0.0001	***
VehValue	107.704	71.1126	1.515	0.1300	
VehNuevo	177.818	161.237	1.103	0.2702	
VehViejo	254.199	175.584	1.448	0.1478	
VehMasViejo	421.210	209.840	2.007	0.0448	**
Sedan	−17.1494	164.304	−0.1044	0.9169	
Utility	234.862	246.059	0.9545	0.3399	
HatchBack	223.991	186.442	1.201	0.2297	
MiniBUs	772.160	548.372	1.408	0.1592	
HardTop	292.563	326.242	0.8968	0.3699	
Coupe	642.456	441.805	1.454	0.1460	
Convertible	106.206	2036.15	0.05216	0.9584	
Truck	638.777	340.056	1.878	0.0604	*
Bus	−468.537	1177.58	−0.3979	0.6907	
MotorizedCaravan	−1484.51	945.428	−1.570	0.1164	
PanelVan	257.725	464.268	0.5551	0.5788	
Roadster	−1036.59	2485.48	−0.4171	0.6767	
PersJoven	−430.723	196.375	−2.193	0.0283	**
PersTrabaj	−631.900	191.174	−3.305	0.0010	***
PersTrabajMayor	−606.123	191.265	−3.169	0.0015	***
PersMayor	−792.706	213.677	−3.710	0.0002	***
PersMasMayor	−595.139	244.409	−2.435	0.0149	**
Hombre	346.393	108.923	3.180	0.0015	***

Media de la vble. dep.	2014.404	D.T. de la vble. dep.	3548.907
Suma de cuad. residuos	5.66e+10	D.T. de la regresión	3506.980
R-cuadrado	0.028346	R-cuadrado corregido	0.023488
F(23, 4600)	5.834677	Valor p (de F)	2.16e-17
Log-verosimilitud	−44292.59	Criterio de Akaike	88633.18
Criterio de Schwarz	88787.72	Crit. de Hannan-Quinn	88687.56

Fuente: Elaboración Propia

III. ACCDIDENTE GRAVE DE TRÁFICO

Descripción del capítulo

Al igual que hemos hecho en el Capítulo 2, para este también antes de empezar explicaremos lo realizado, los objetivos y la estructura.

Primero de todo, detallaremos la base de datos usada y luego haremos un análisis de estadística-descriptivo de ciertas variables dependientes e independientes del modelo a través de la elaboración de gráficos de distribución de frecuencias y análisis de los estadísticos principales. Esto nos permitirá ver cómo están compuestas todas ellas y los valores que toman. Nuevamente, presentamos diversos contrastes de medias, pero en este caso, al ser la variable dependiente del tipo binaria obtendremos una especie de media de probabilidad en función de características propias de cada observación.

Cerrando el trabajo y como objetivo, primero de todo, queremos construir un modelo de regresión lineal múltiple a partir de la muestra disponible. Con esto, sabremos qué variables son estadísticamente significativas para determinar el número de personas que resultan gravemente heridas en un accidente de automóvil, implicando su hospitalización o en su defecto, víctimas mortales. Terminaremos con la elaboración del modelo logit que nos permitirá saber qué características (externas al conductor) influyen en la probabilidad del suceso descrito.

3.1 Análisis de la Base de datos

La base de datos utilizada para la construcción de este capítulo fue obtenida de la web pública de la DGT, en el apartado de Estadísticas e Indicadores. De allí nos descargamos la información en formato Excel para luego tratarla con el Gretl. Para la realización de esta parte del trabajo, hemos tomado un total de 27.244 observaciones. Cada observación hace referencia a un accidente de automóvil en España el año 2018, además, en todas hay como mínimo una víctima moral o herido hospitalizado. Los datos incluyen todos los accidentes habidos en las provincias españolas excepto las pertenecientes a Catalunya y el País Vasco.

Las variables que aparecen en el estudio son las que mencionamos a continuación:

Hora: Hora del accidente. Definida para las categorías: mañana, tarde y noche

EstacionAño: Estación del año que se produjo el accidente. Definida para las categorías: Invierno, primavera, verano y otoño.

LocProvincia: Localización geográfica de la provincia del accidente. Definida para las categorías: norte, sur, centro e islas.

Zona: Variable cualitativa que muestra si el accidente fue en una vía interurbana o travesía.

Sentido: Sentido de la carretera. Definida para las categorías: ascendente, descendente, ambas a la vez, o se desconoce.

Tipo_Via: Variable que explica el tipo de vía del accidente. Definida para las categorías: autopista, autovía, camino vecinal, vía convencional, vía de servicio, ramal de enlace, travesía y otro tipo.

Trazado: Trazado de la vía. Definida para las categorías: fuera de Intersección, en intersección, o se desconoce.

Tipo_Accidente: Variable que nos muestra el tipo de accidente. Definida para las categorías: por alcance, fronto-Lateral, atropello a personas, salida de la vía por la derecha con vuelco, salida de la vía por la derecha otro tipo, salida de la vía por la derecha con colisión, salida de la vía por la derecha con despeñamiento, salida de la vía por la izquierda con vuelco, salida de la vía por la izquierda otro tipo, salida de la vía por la izquierda con colisión, salida de la vía por la izquierda por despeñamiento, lateral, múltiple o en caravana, frontal, caída, atropello a animales, colisión contra obstáculo o elemento de la vía, vuelco y otro tipo.

Total_Muertos: Número total de fallecidos 30 días habidos en el accidente.

Acc_Muerte_o_Hosp: Variable que indica el número total de víctimas mortales y heridos Hospitalizados que hubo en el accidente.

Acc_Grave: Variable dicotómica que toma valor 1 si en el accidente hubo víctimas y/o heridos hospitalizados y 0 en caso contrario.

Queremos detallar que, las variables dependientes de los modelos resultantes serán el Acc_Muerte_o_Hosp, y Acc_Grave, en función del apartado en el que nos encontremos. Por otro lado, las variables independientes que adoptarán un papel de regresores serán todas las variables restantes descritas arriba. En este caso, todas estas últimas variables son del tipo cualitativas y como bien decíamos en el Capítulo 2, añadimos ficticias con categorías binarias para su tratamiento.

En un principio se planteó la idea que la variable Tipo_Accidente podría presentar categorías no excluyentes. Es decir, que una misma observación, tuviera asociado dos tipos de accidentes distintos a la vez. Tras un análisis para comprobarlo, se concluyó que las categorías eran excluyentes.

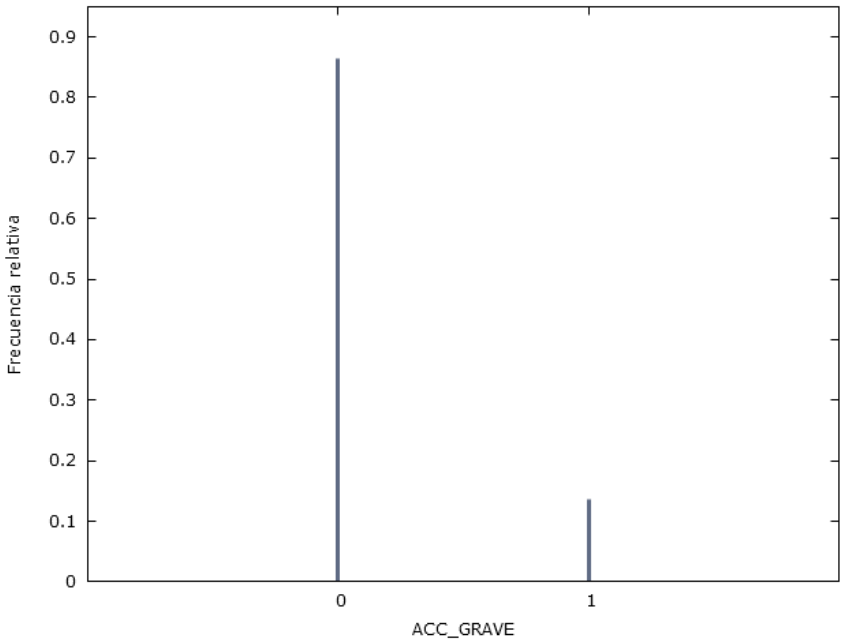
3.2 Estadística descriptiva

Al igual que en el capítulo anterior, nos disponemos a estudiar cómo se distribuyen tanto algunas de las variables independientes y dependientes. Empezamos mostrando la distribución de frecuencias para la variable regresada Acc_Grave. Decíamos anteriormente que es una variable binaria que toma valor 1 si en el accidente hubo alguna víctima mortal y/o heridos graves hospitalizados y 0 en caso de que los heridos

sean únicamente leves. Apreciamos que del total de accidentes que hubo, 3.722 fueron accidentes con daños graves a las personas involucradas, esta cifra representa un 13,66% del conjunto de la muestra. De lo contrario, las otras 23.522 observaciones restantes, fueron accidentes leves sin mayor gravedad, implicando un 86,34%.

Profundizando un poco más la variable, debido a que Acc_Grave se obtiene a partir de la suma de muertos y heridos hospitalizados, para luego transformarla en una variable binaria, podemos ver que se produjeron un total de 1.091 muertos. Este número es el resultante de sumar 928 observaciones con 1 muerto, 60 con 2, 10 con 3, 2 con 4 y 1 con 5, respectivamente. El resultado se calcula mediante la apreciación de la distribución de frecuencias de la variable Total_Muertos, que para hacer una explicación más fluida no lo representaremos.

Gráfico 3.1. Histograma de la variable Acc_Grave



Distribución de frecuencias para ACC_GRAVE, observaciones 1-27244

	frecuencia	rel.	acum.	
0	23522	86.34%	86.34%	**********
1	3722	13.66%	100.00%	****

Fuente: Elaboración propia

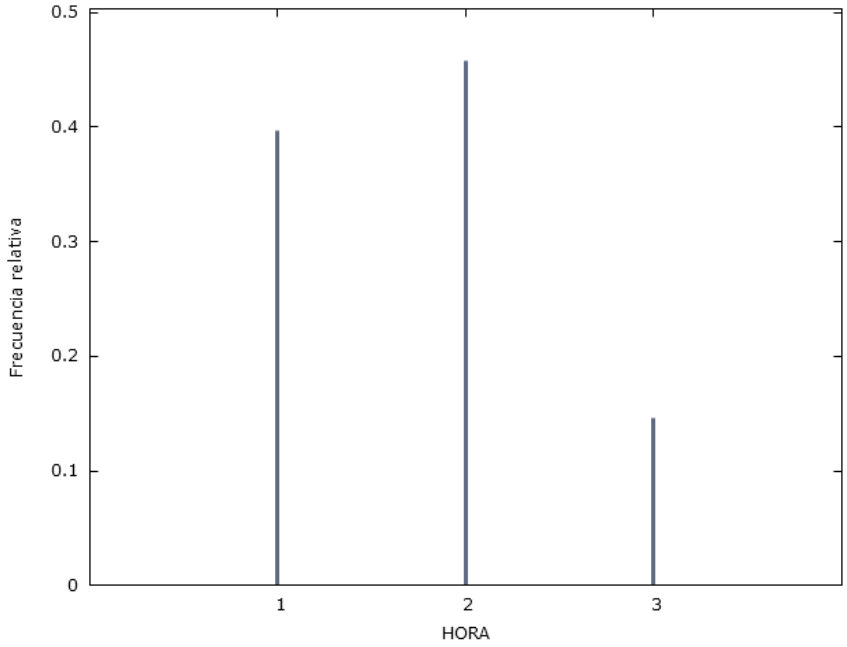
Ahora que se ha analizado la formación de la variable endógena, análogamente repetimos, pero esta vez para algunas de las variables explicativas que el autor considera que detallan más información.

Mostramos el análisis de distribución de frecuencias para las variables cualitativas Hora, EstacionAño y Localizacion.

De esta manera si nos centramos primero en la variable que refleja la hora en la que el accidente sucedió podemos apreciar que predominan accidentes ocurridos durante la mañana, en concreto, el 45,75% del total. La cantidad de accidentes habidos durante la tarde se asemeja a los de la mañana, difiriendo únicamente en un 6% aproximadamente. Finalmente, el 14.57% de observaciones se dieron durante la noche, porcentaje claramente menor a los otros dos, sin embargo, no menos importante.

Debido a que claramente durante la mañana es cuando el tráfico de automóviles tiene mayor peso, es lógico pensar que haya más accidentes. A medida que transcurre el día la movilidad de vehículos disminuye y con ello las víctimas y hospitalizados.

Gráfico 3.2. Histograma de la variable Hora



Distribución de frecuencias para HORA, observaciones 1-27244

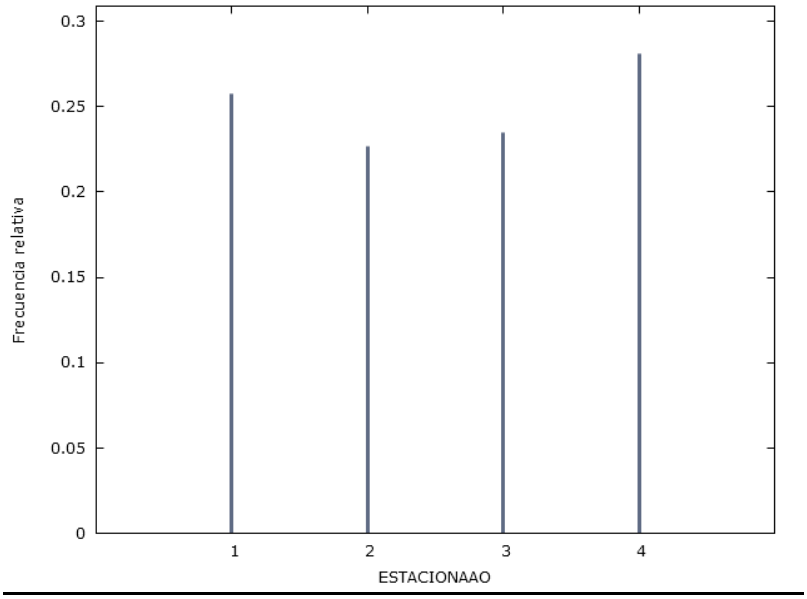
	frecuencia	rel.	acum.	
Tarde	10812	39.69%	39.69%	**********
Mañana	12463	45.75%	85.43%	**********
Noche	3969	14.57%	100.00%	*****

Fuente: Elaboración Propia

Para la variable que muestra en qué estación del año tuvo lugar la observación concluimos que es durante los meses del verano cuando aparecen más accidentes. Esta variable se comporta de una manera más uniforme, es decir, las cuatro estaciones del año tuvieron un número de sucesos bastante igual; contrariamente a lo que vimos para HORA. Este pequeño protagonismo de verano se debe a factores que la gran parte de la población conoce, meses de vacaciones donde la movilidad se alza sumado a,

posiblemente, más fiestas y reuniones sociales incrementando la presencia de substancias perjudicantes a la hora de conducir.

Gráfico 3.3. Histograma de la variable EstacionAño



Distribución de frecuencias para ESTACIONAÑO, observaciones 1-27244

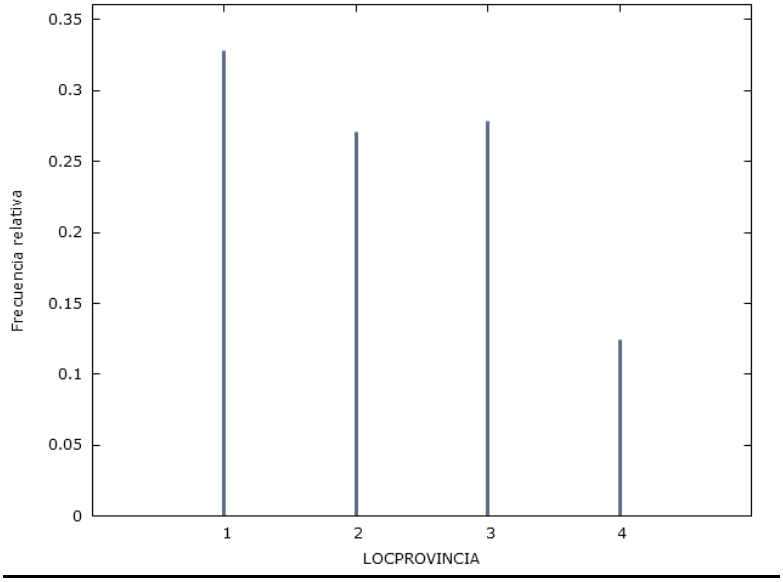
	frecuencia	rel.	acum.	
Otoño	7013	25.74%	25.74%	*********
Invierno	6188	22.71%	48.45%	********
Primavera	6394	23.47%	71.92%	********
Verano	7649	28.08%	100.00%	**********

Fuente: Elaboración propia

Una vez comentado el tiempo de los accidentes, realizamos lo mismo haciendo referencia al espacio, es decir el lugar en que se produjo. Primero de todo, debemos destacar que estos datos pueden estar un poco sesgados debido a que la agrupación geográfica de cada provincia fue a criterio del autor; quizá, por ejemplo, una provincia asignada al norte, a efectos geográficos es del centro del país. De esta manera, el número de observaciones reales en cada categoría puede ser un poco ambiguo.

Pasando por alto lo dicho y centrándonos en la parte descriptiva, vemos que es mayor el número de accidentes en las provincias del centro, específicamente, del total de las observaciones 8912 tuvieron lugar en estas. A la cola seguidamente encontramos a SUR y NORTE. Hay que destacar que ISLAS, aparece como última porque obviamente tiene menor densidad de circulación en comparación con las provincias peninsulares.

Gráfico 3.4. Histograma de la variable LocProvincia



Distribución de frecuencias para LOCPROVINCIA, observaciones 1-27244

	frecuencia	rel.	acum.	
CENTRO	8912	32.76%	32.76%	**********
NORTE	7354	27.03%	59.79%	*********
SUR	7558	27.78%	87.58%	**********
ISLAS	3379	12.42%	100.00%	****

Observaciones ausentes = 41 (0.15%)

Fuente: Elaboración propia

Para concluir con los procesos de estadística descriptiva del capítulo tratado, diremos que el número máximo de personas involucradas en un mismo accidente grave fue de 12 (se desconoce cuántas de ellas fueron víctimas y cuántas heridas graves).

En el segundo capítulo a partir del contraste de medias para las variables cualitativas, pudimos observar aquellas categorías con una media de siniestros más elevada y si estas eran o no estadísticamente significativas. Ahora haremos el mismo procedimiento, pero al tener como variable dependiente una del tipo dicotómica (Acc_Grave), mostraremos la media de probabilidad de tener accidente grave en función de las ficticias.

Para empezar, queremos ver para la variable Hora, aquella franja horaria que presenta una mayor probabilidad de accidente grave. Si ponemos entonces, como endógena Acc_Grave y explicativas todas sus las categorías excluyendo Tarde, obtenemos que la Noche, implica una mayor probabilidad media de accidente grave, en comparación con la Mañana y Tarde. A esta conclusión llegamos debido a que, el coeficiente asociado a las horas nocturnas es el mayor de los tres obtenidos. En concreto, la probabilidad de

que el accidente sea en la Tarde (categoría base) es el valor de la constante, es decir 0,14. El coeficiente asociado a Mañana, nos indica que esta categoría presenta una probabilidad de 0,01 menor respecto a la base. Contrariamente, los accidentes durante la noche tienen 0,04 más probabilidades de ser graves. Además, afirmamos que todas las categorías son estadísticamente significativas para explicar la probabilidad analizada.

Cuadro 3.1. Contraste de medias de la variable Acc_Grave frente a la variable Hora

MCO, usando las observaciones 1-27244
Variable dependiente: ACC_GRAVE

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.136700	0.00329923	41.43	<0.0001	***
Mañana	−0.0122516	0.00450865	−2.717	0.0066	***
Noche	0.0379032	0.00636685	5.953	<0.0001	***

Fuente: Elaboración Propia

Paralelamente, para la variable estación del año, concluimos que es durante el verano, que el riesgo de sufrir un accidente con víctimas y/o hospitalizados graves es más grande. Como todos los regresores aparecen con signo positivo, remarcamos que, durante el otoño, la media de probabilidad es la menor. Mirando los valores p, detectamos que únicamente los accidentes en verano son significativos para explicar la probabilidad media de la variable dependiente.

Cuadro 3.2. Contraste de medias de la variable Acc_Grave frente a la variable EstacionAño

MCO, usando las observaciones 1-27244
Variable dependiente: ACC_GRAVE

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.126907	0.00410003	30.95	<0.0001	***
Invierno	0.00722340	0.00598846	1.206	0.2277	
Primavera	0.00649915	0.00593699	1.095	0.2737	
Verano	0.0233085	0.00567651	4.106	<0.0001	***

Fuente: Elaboración Propia

Si hacemos procesos análogos a los descritos, pero tomando en consideración el lugar del accidente en territorio español, decimos que es en el Norte donde los accidentes graves pueden tener mayor protagonismo. Al igual que explicábamos antes, esta variable no tiene en consideración las provincias catalanas ni vascas, así que esta conclusión puede presentar alguna incerteza a nivel empírico.

Cuadro 3.3. Contraste de medias de la variable Acc_Grave frente a la variable LocProvincia

MCO, usando las observaciones 1-27244 (n = 27203)
Se han quitado las observaciones ausentes o incompletas: 41
Variable dependiente: ACC_GRAVE

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.121185	0.00363426	33.35	<0.0001	***
Norte	0.0402238	0.00540498	7.442	<0.0001	***
Sur	0.0120514	0.00536487	2.246	0.0247	**
Islas	0.00962301	0.00693131	1.388	0.1650	

Fuente: Elaboración Propia

Siguiendo con el análisis, obtenemos que hay una clara diferencia entre la probabilidad de accidentes graves en zonas de travesías y vías interurbanas, siendo las primeras, altamente más peligrosas, porque tiene un coeficiente estimado de cinco centésimas superior aproximadamente.

Cuadro 3.4. Contraste de medias de la variable Acc_Grave frente a la variable Zona

MCO, usando las observaciones 1-27244
Variable dependiente: ACC_GRAVE

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.135805	0.00209839	64.72	<0.0001	***
Travesias	0.0477801	0.0160965	2.968	0.0030	***

Fuente: Elaboración Propia

Por lo que se refiere al sentido de la vía, aquellas que tienen sentido ascendente y descendente a la vez, son las que implican una mayor probabilidad media.

Cuadro 3.5. Contraste de medias de la variable Acc_Grave frente a la variable Sentido

MCO, usando las observaciones 1-27244
Variable dependiente: ACC_GRAVE

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.130573	0.00314032	41.58	<0.0001	***
Ascendente	0.00151384	0.00434899	0.3481	0.7278	
Ambos	0.0636015	0.00786063	8.091	<0.0001	***
Se_Desconoce	0.0183629	0.0501343	0.3663	0.7142	

Fuente: Elaboración Propia

Finalmente, afirmamos que el atropello a personas es el tipo de accidente que imprime un mayor porcentaje de víctimas y/o heridos graves porque tiene el mayor coeficiente

positivo. Estos resultados obtenidos, realizan la validez de la base de datos, ya que a priori resulta lógico pensar que una persona como peatona, al estar totalmente descubierta, puede sufrir más daños si lo comparamos con todos los otros tipos de accidentes, donde las personas se encuentran protegidas mínimamente por la seguridad del vehículo.

Cuadro 3.6. Contraste de medias de la variable Acc_Grave frente a la variable TipoAccidente

MCO, usando las observaciones 1-27244
Variable dependiente: ACC_GRAVE

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.0570608	0.00445248	12.82	<0.0001	***
Fronto_Lateral	0.0644726	0.00706139	9.130	<0.0001	***
Atropello_Pers	0.336360	0.0128529	26.17	<0.0001	***
Salida_via_der_vuelco	0.0759580	0.00848168	8.956	<0.0001	***
Lateral	0.0366349	0.00915008	4.004	<0.0001	***
Salida_via_der_otrotipo	0.0435603	0.00940463	4.632	<0.0001	***
Mulitple_o_caravana	−0.00338522	0.00918059	−0.3687	0.7123	
Frontal	0.323061	0.0101809	31.73	<0.0001	***
Salida_via_der_colision	0.146977	0.00886167	16.59	<0.0001	***
Caida	0.0676052	0.0108723	6.218	<0.0001	***
Salida_via_izq_colision	0.140503	0.00972501	14.45	<0.0001	***
Salida_via_izq_otrotipo	0.0644807	0.0113577	5.677	<0.0001	***
Salida_via_der_despeñamiento	0.169502	0.0175368	9.665	<0.0001	***
Salida_via_izq_despeñamiento	0.173065	0.0219566	7.882	<0.0001	***
Salida_via_izq_vuelco	0.0879004	0.0102698	8.559	<0.0001	***
Atropello_animal	0.00855597	0.0176013	0.4861	0.6269	
Colision_obstaculo_elementovia	0.0784890	0.0173893	4.514	<0.0001	***
Vuelco	0.125421	0.0205680	6.098	<0.0001	***
Otro_tipo_acc	0.161689	0.0200859	8.050	<0.0001	***

Fuente: Elaboración Propia

A modo de resumen, obtenemos que los accidentes que tienen una media de probabilidad más elevada de ser graves son los que se producen durante la noche, en verano, en las provincias del norte del país y en travesías, donde la seguridad es menor.

Además, cuando el sentido de la vía es doble produciéndose el accidente por atropello a personas. Para concluir, debemos destacar que este contraste probabilístico no se ha hecho para el conjunto de variables cualitativas porque para el tipo y trazado de vía el resultado mayor es 'Otro tipo' y 'Se desconoce', respectivamente, dando lugar a conclusiones no específicas.

3.3 Modelización del Número de personas muertas y/o heridas graves en accidentes

Siguiendo con la estructura del trabajo, presentamos seguidamente el MRLM que obtenemos a partir de la información disponible. En él, incluimos cómo endógena dependiente la variable Acc_Muerte o Hosp y como posibles explicadores de su comportamiento, todo el conjunto de regresores apreciables debajo. Intentaremos ver qué fue lo que influyó en el hecho que el accidente automovilístico implicara personas tanto muertas como heridas graves. La especificación del modelo es la siguiente.

$$AccMuerteoHosp_i = f(Hora_i, EstacionAño_i, LocProvincia_i, Zona_i, Sentido_i, TipoVia_i, Trazado_i, TipoAccidente_i)$$

Nuevamente, seleccionamos la estimación mediante Desviaciones típicas robustas para solucionar el problema de heteroscedasticidad detectado previamente al hacer el contraste de White.

Siguiendo con lo anterior, incluimos como variable explicativa nuestra endógena al cuadrado y al resultar que su beta estimado es significativo, rechazaremos la hipótesis nula que el modelo lineal es el correcto, para afirmar que será correcta un modelo cuadrático, como mínimo.

Si recordamos anteriormente, en el capítulo 2, en el MRLM que construimos, obtuvimos un R-cuadrado significativamente bajo. En este modelo, apreciamos un valor del 59,9%, esto implicaba, como comentamos, que los regresores podían explicar alrededor de un 60% de la variabilidad total de la endógena. Seguramente este número se puede incrementar para tener unos resultados más precisos, buscando variables relevantes no disponibles, pero es destacable positivamente debido el gran salto comparando los dos modelos presentados.

Resultan variables independientes significativas para explicar el comportamiento de la endógena la Hora del accidente, el hecho que se produjese en verano, en una travesía, si la vía es del tipo camino vecinal, vía convencional u otro no descrito en la muestra. También si el trazado se encuentra dentro de una intersección. Finalmente, resultan influyentes todos los tipos de accidentes mostrados, exceptuando el tipo múltiple o en caravana, a la vez que los producidos por salida de la vía por la izquierda con despeñamiento. Profundizando más en el análisis, mediante la observación de los coeficientes beta asociados a los regresores significativos podremos saber qué factores, tanto propios como externos al vehículo hacen, conllevan un aumento o disminución del número de personas muertas y/o heridas graves en los accidentes de la muestra. Si cogemos primero la variable ficticia hora, en concreto, la categoría referente a la noche se ve claramente que el hecho que el accidente se dé en esta franja horaria incrementa

el número de personas involucradas en un accidente grave. Esta afirmación se sostiene debido a que el coeficiente beta es con signo positivo 0,0242996. Analíticamente, un accidente ocurrido en la noche hará que la categoría Noche tome valor 1, implicando que una de las partes de la ecuación del modelo sea, el valor unitario multiplicado por la beta obtenido, dando un resultado positivo, por lo tanto, un número de accidente grave estimado mayor en comparación a un accidente no producido durante la noche. Para no causar posibles incertezas, mostramos la parte de los resultados para un accidente dado durante las horas estudiadas.

$$Acc\_Muerte\_o\_Hosp_i = \beta_o + \beta_1 DHora\_2_{i1} + \beta_2 DHora\_3_{i2} + (\dots) + \varepsilon_i =$$

$$Acc\_Muerte\_o\_Hosp_i = 0.0422367 + (-\cancel{0.00813019 * 0}) + 0.0242996 * 1 + (\dots) + \varepsilon_i$$

Siguiendo con el mismo procedimiento, resulta interesante ver que los accidentes que tuvieron lugar en los meses de verano son significativos para explicar el comportamiento de la variable dependiente y alzando el valor de esta debido también, al coeficiente positivo resultante en la estimación. Finalmente afirmamos que todos los tipos de accidentes que hemos descrito ser significativos tienen asociado un aumento de la variable dependiente, exceptuando los que son referente a atropello a animales, que disminuyen el número de personas estimado.

Cuadro 3.7. Modelización de la variable Acc_Muertey/oHosp

MCO, usando las observaciones 1-27244 (n = 27203)
Se han quitado las observaciones ausentes o incompletas: 41
Variable dependiente: ACC_MUERTEoHOSP
Desviaciones típicas robustas ante heterocedasticidad, variante HC1

	Coeficiente	*Desv. Típica*	*Estadístico t*	*valor p*	
const	0.0422367	0.00656736	6.431	<0.0001	***
Mañana	−0.00813019	0.00379626	−2.142	0.0322	**
Noche	0.0242996	0.00705795	3.443	0.0006	***
Invierno	−0.00205868	0.00521947	−0.3944	0.6933	
Primavera	0.00443515	0.00497564	0.8914	0.3727	
Verano	0.0137935	0.00518688	2.659	0.0078	***
Norte	0.00761468	0.00475454	1.602	0.1093	
Sur	0.000893874	0.00431256	0.2073	0.8358	
Islas	−0.00094675	0.00594019	−0.1594	0.8734	
Travesias	0.0259984	0.0152004	1.710	0.0872	*
Ascendente	−0.00209915	0.00385552	−0.5445	0.5861	
Ambos	−0.00343547	0.00806932	−0.4257	0.6703	
Se_Desconoce	0.0309808	0.0558424	0.5548	0.5790	
Camino_Vecinal	0.0365014	0.0113855	3.206	0.0013	***
Via_Convencional	0.0276507	0.00656059	4.215	<0.0001	***
Autopista	0.000641657	0.00675112	0.09504	0.9243	
Via_Servicio	−0.00369944	0.0219394	−0.1686	0.8661	
Ramal_Enlace	0.0323647	0.0226290	1.430	0.1527	

Otro_Tipo	0.0926513	0.0318625	2.908	0.0036	***
En_Interseccion	−0.0396672	0.00762390	−5.203	<0.0001	***
Se_Desconoce	−0.0315251	0.144762	−0.2178	0.8276	
Fronto_Lateral	0.0547117	0.00930594	5.879	<0.0001	***
Atropello_Pers	0.249215	0.0190566	13.08	<0.0001	***
Salida_via_der_vuelco	0.0409778	0.00717336	5.712	<0.0001	***
Lateral	0.0281713	0.00668212	4.216	<0.0001	***
Salida_via_der_otroTipo	0.0142467	0.00690807	2.062	0.0392	**
Multiple_Caravana	0.00148228	0.00567958	0.2610	0.7941	
Frontal	0.252520	0.0470466	5.367	<0.0001	***
Salida_via_der_colision	0.102529	0.0125366	8.178	<0.0001	***
Caida	0.0400658	0.00829725	4.829	<0.0001	***
Salida_cia_izq_colision	0.0975631	0.0140665	6.936	<0.0001	***
Salida_via_izq_otroTipo	0.0331752	0.00872350	3.803	0.0001	***
Salida_via_der_despeñamiento	0.112794	0.0211224	5.340	<0.0001	***
Salida_via_izq_despeñamiento	0.0276502	0.0572236	0.4832	0.6290	
Salida_via_izq_vuelco	0.0522294	0.00902688	5.786	<0.0001	***
Atropello_Animal	−0.0234853	0.0120953	−1.942	0.0522	*
Colision_ostaculo_elemetoVia	0.0549570	0.0157044	3.499	0.0005	***
Vuelco	0.0899237	0.0190513	4.720	<0.0001	***
Otro_tipo_acc	0.116412	0.0235419	4.945	<0.0001	***
sq_ACC_MUERTEoHOSP	0.219215	0.0409571	5.352	<0.0001	***

Media de la vble. dep.	0.168584	D.T. de la vble. dep.	0.487581
Suma de cuad. residuos	2592.962	D.T. de la regresión	0.308965
R-cuadrado	0.599039	R-cuadrado corregido	0.598464
F(39, 27163)	92.09687	Valor p (de F)	0.000000
Log-verosimilitud	−6628.700	Criterio de Akaike	13337.40
Criterio de Schwarz	13665.84	Crit. de Hannan-Quinn	13443.28

Fuente: Elaboración Propia

3.4 Modelización de la probabilidad de Accidente Grave

Para terminar con el trabajo presentamos el modelo de probabilidad construido. Para ello, ponemos como variable dependiente Acc_Grave que, tal y como decíamos en la descripción de la base de datos, se trata de una variable binaria con valor de la observación unitario si el accidente ha tenido tanto muertos como heridos graves y cero cuando los heridos son únicamente leves. La expresión del modelo en este caso queda tal que así.

$$Prob\ (Y_i = 1) = \frac{1}{1 + e^{-Z}} = \frac{e^Z}{1 + e^Z}$$

Donde Z depende de las mismas variables que en el modelo de regresión analizado en el apartado anterior.

Ahora nos centraremos en los regresores que han resultado ser estadísticamente significativos a la hora del cálculo de probabilidad de accidente grave. Vemos que la probabilidad de tener un accidente grave durante las horas de la noche y la mañana es significativamente distinta en comparación a las horas de la tarde, al actuar esta como categoría base. Para la variable EstacionAño, se aprecia que es durante el varano tan solo, que la probabilidad es significativamente superior respecto a otoño, todas las demás estaciones no presentan significación. Para la variable donde se localiza la provincia del accidente diremos que, únicamente las observaciones en el norte presentan una probabilidad de accidente grave superior a la categoría de referencia, en este caso, 'Centro'. Siguiendo con el análisis, observamos que existen diferencias de probabilidad entre travesías y vías interurbanas. Tomando como categoría base las autovías, podemos añadir que solo la probabilidad de tener un accidente grave es significativamente distinta en vías y caminos vecinales, de las demás categorías no podemos decir lo mismo. Para terminar con la observación de los valores p asociados a cada categoría, afirmamos que todos los tipos de accidente, exceptuando los múltiple o en caravana y los atropellos a animales, tienen una probabilidad significativamente distinta en comparación a los accidentes tipo 'Por alance'.

El hecho que el accidente se produzca en una travesía también incrementa posibilidades significativamente, a la práctica, es coherente que así sea ya que debemos tener en cuenta que las travesías acostumbran a ser tramos de zona poblada, donde la seguridad vial, tanto para los pasajeros del vehículo como los peatones, es reducida. Si el accidente tuvo lugar en vías del tipo convencional o camino vecinal también actúa como significante. La realidad nos corrobora este último dato, pues este tipo de vías parecen ser las más rudimentarias comparadas con las otras disponibles. Continuando con el estudio, obtenemos que, el trazado de la vía implica siendo esta una intersección también es significativa a la hora de determinar la probabilidad, entendiendo la intersección como el cruce de dos o más caminos. La gran mayoría de regresores usados en el modelo han aparecido estadísticamente significativo a la hora de explicar la probabilidad de accidente grave, nuevamente aclarando que se compara con la categoría base, sin embargo, detallaremos algunos que no han resultado serlo. Para

empezar, el invierno. En un principio podemos pensar que, debido al frío y lo que conlleva, como, por ejemplo, nieve, heladas, viento… sería obvio pensar que influyen en la probabilidad, pero el modelo nos dice lo contrario. El sentido de la vía, en especial aquellas a la vez ascendentes y descendentes tampoco es elocuente, cosa que cualquier lector podría pensar antes de ver los resultados para nuestra muestra.

Todas estas conclusiones las hemos podido explicar mediante la observación del 'valor p' asociado a cada regresor. Bien, si ahora nos centramos en los valores de las pendientes, podremos saber si la probabilidad de accidente grave se ve aumentada o disminuida en función de las características únicas de éste. Siguiendo la estructura de los otros modelos logit, centramos el estudio en los pendientes de las variables explicativas significativas descritas justo antes. De esta manera, pendientes con signo positivo harán incrementar la probabilidad y aquellos negativos la decrecerán. Para agilizar el análisis, primero miraremos los pendientes positivos para acabar con los negativos. En el primer grupo, aparecen los accidentes habidos durante la noche y en los meses que comprende verano y en las provincias norteñas. En consonancia a lo explicado anterior, las travesías hacen subir la probabilidad de accidente grave al igual que las vías convencionales y caminos vecinales. Para terminar, todos los tipos de accidente significativos estadísticamente hacen aumentar la probabilidad de ser grave. En el bloque de pendientes negativos significativos, aparecen los accidentes por la mañana, en intersecciones de vehículos.

Cuadro 3.8. Modelización de la probabilidad de Accidente Grave

Logit, usando las observaciones 1-27244 (n = 27203)
Se han quitado las observaciones ausentes o incompletas: 41
Variable dependiente: ACC_GRAVE
Desviaciones típicas basadas en el Hessiano

	Coeficiente	*Desv. Típica*	*z*	*Pendiente**
const	−2.87267	0.0819506	−35.05	
Mañana	−0.0873577	0.0405918	−2.152	−0.00894189
Noche	0.204490	0.0532822	3.838	0.0221909
Invierno	0.0220816	0.0538560	0.4100	0.00227692
Primavera	0.0445191	0.0533411	0.8346	0.00461103
Verano	0.140872	0.0500475	2.815	0.0148091
Norte	0.108890	0.0479062	2.273	0.0113942
Sur	0.0260647	0.0489922	0.5320	0.00268716
Islas	−0.00128033	0.0645474	−0.01984	−0.00013136
Travesias	0.322892	0.145539	2.219	0.0373089
Ascendente	−0.0129307	0.0389429	−0.3320	−0.00132687
Ambos	−0.0746061	0.0699950	−1.066	−0.00747669
Se_desconoce	0.108661	0.481493	0.2257	0.0116246
Camino_Vecinal	0.356112	0.0963136	3.697	0.0414261
Via_Convencional	0.293377	0.0512961	5.719	0.0294220
Autopista	−0.0331735	0.0941048	−0.3525	−0.00336763
Via_Servicio	−0.0307065	0.276620	−0.1110	−0.00311505
Ramal_Enlace	0.380892	0.246864	1.543	0.0450821
Otro_Tipo	0.808089	0.213817	3.779	0.111250

En_interseccion	−0.496029	0.0496021	−10.00	−0.0469443
Se_desconoce	−0.0439931	1.25241	−0.03513	−0.00443968
Fronto_Lateral	0.890554	0.0825236	10.79	0.116754
Atropello_Pers	2.23123	0.0979371	22.78	0.425008
Salida_via_der_vuelco	0.700767	0.0884167	7.926	0.0898105
Lateral	0.546455	0.101563	5.380	0.0671631
Sal_via_der_otroTipo	0.398793	0.102982	3.872	0.0467820
Multile_Caravana	−0.0531486	0.122349	−0.4344	−0.00535857
Frontal	2.12934	0.0884587	24.07	0.393179
Sal_via_Der_colision	1.25941	0.0830062	15.17	0.190414
Caida	0.733342	0.109583	6.692	0.0966717
Sal_via_izq_colision	1.28722	0.0884650	14.55	0.198257
Sal_via_izq_otroTio	0.637615	0.113859	5.600	0.0816140
Sal_via_der_despeñamiento	1.32033	0.137226	9.622	0.211414
Sal_via_izq_despeñamiento	1.31226	0.166952	7.860	0.210481
Sal_via_izq_vuelco	0.825394	0.0998727	8.264	0.111676
Atropello_Animal	−0.175371	0.216970	−0.8083	−0.0168564
Colision_obstaculo_elementoVia	0.828849	0.159898	5.184	0.114311
Vuelco	1.22209	0.168553	7.250	0.190926
Otro_tipo_acc	1.42541	0.155601	9.161	0.235351

Media de la vble. dep.	0.136603	D.T. de la vble. dep.	0.343434
R-cuadrado de McFadden	0.079874	R-cuadrado corregido	0.076279
Log-verosimilitud	−9980.736	Criterio de Akaike	20039.47
Criterio de Schwarz	20359.70	Crit. de Hannan-Quinn	20142.70

*Evaluado en la media

Número de casos 'correctamente predichos' = 23482 (86.3%)

f(beta'x) en la media de las variables independientes = 0.343

Contraste de razón de verosimilitudes: Chi-cuadrado(38) = 1732.81 [0.0000]

		Predicho	
		0	1
Observado	0	23.461	26
	1	3.695	21

Fuente: Elaboración Propia

Terminando el trabajo, nuevamente explicaremos la bondad del ajuste del modelo resultante. Decimos que, al ser un modelo con datos de corte transversal, obtenemos un R^2 de McFadden notablemente bajo, muy cerca del 0, implicando una capacidad predictiva de probabilidad de automóvil bastante pobre. Para seguir con este análisis tendremos en cuenta las observaciones correctamente predichas. Para ello hay que sumar los valores que aparecen en las celdas con el mismo valor. Es decir, sumaremos los casos que el modelo predice que no tendrían accidente grave y efectivamente son apoyados por la observación, en nuestro caso 23.482, más los casos que tuvieron accidente y efectivamente en la realidad se produjo, concretamente 21. En cambio, si miramos cuando el modelo falla, tenemos que 3.695 casos el modelo predijo no tener accidente gravo cuando realmente sí tuvieron. Nuevamente sumamos los casos que el modelo predijo accidente grave y la no se produjeron con certeza, precisamente 26, obteniendo finalmente 3.721 resultados erróneos.

IV. CONCLUSIONES

Una vez construidos los modelos, tenemos información suficiente para detallar y lograr los objetivos propuestos para cada capítulo. Mediante la redacción de los resultados obtenidos, elaboramos las conclusiones que el trabajo nos brinda.

Siguiendo con la estructura, primeramente, haremos referencia al capítulo uno, sobre los siniestros de automóviles. Al calcular el contraste de medias de la variable dependiente número de siniestros frente la variable género del conductor, apreciamos que, contrario a lo que popularmente se puede pensar, la media de siniestros de los individuos que son hombres no es estadísticamente distinta respecto a la media de las mujeres. Al principio del trabajo, creíamos que los vehículos más antiguos de la muestra serían los que presentarían mayor valor para la variable dependiente tomada, pero al hacer el contraste apreciamos que no había diferencias significativas entre esta categoría y los vehículos más nuevos. Resultó interesante ver, que a medida que la edad del conductor incrementaba, la media de siniestros decrecía. Creer entonces que personas mayores debido a la pérdida de facultades físicas se ven protagonistas de más siniestros es totalmente erróneo. Una posible explicación a esto es el hecho que el grupo de individuos de más avanzada edad tiende a usar cada vez menos el vehículo, siendo sus familiares los responsables de su movilidad. Calculando lo mismo, pero para la variable tipo de vehículo, resultó interesante ver que, entre otros, los vehículos tipo Coupe y Hardtop, presentaban una media de siniestros significativamente mayor a la categoría base Sedan. Muy probablemente esto se debe a que, si el lector analiza estos dos tipos de coche, verá que ambos son de estilo deportivo, y al tener una mayor potencia implicaran seguramente más siniestros.

Dicho esto, planteamos el modelo de regresión lineal múltiple para cumplir con los objetivos y detallar aquellas variables que resultaban ser significativas para determinar el número de siniestros que un individuo con unas características dadas podría sufrir. Por un lado, la variable independiente Exposure (el tiempo que la póliza de seguro había tenido efecto) era estadísticamente significativa para explicar el comportamiento de la endógena y tenía asociado un coeficiente estimado positivo. Es lógico pensar que, si el conductor había estado más tiempo asegurado, también había estado más tiempo tomado en consideración en la muestra, con lo cual esto hace aumentar el número de siniestros que es capaz de sufrir. Paralelamente, el valor del vehículo presentaba resultados parecidos a la variable tratada justo antes, siendo nuevamente significativos. Concluimos que vehículos más caros implicaban un valor más grande de la variable dependiente. Asociado a lo que hemos visto en el contraste de medias, los coches más caros tienden a ser los más potentes y de mayor cilindrada, con un manejo más complejo haciendo aumentar el número de siniestros. Debemos destacar que, los vehículos más antiguos no resultaron ser significativos para la variable dependiente número de siniestros, posiblemente porque estos son los que presentan mayor tiempo de desuso. Es lógico pensar que vehículos recientemente adquiridos se usen más, resultando mayor siniestralidad. Para todas las categorías de la edad del conductor/a se obtenían unos

coeficientes estadísticamente significativos para explicar el comportamiento de nuestra variable dependiente y con valor absoluto en crecimiento. Por último, el género del conductor/a resultó no ser significativo para el número de siniestros.

El capítulo terminaba con la construcción del modelo de probabilidad logit para poder establecer un perfil de conductor/a más propenso a tener siniestro. En este caso, tomamos como endógena ClaimOcc y regresores los mismos que en el MRLM. Las variables independientes que resultaron ser estadísticamente significativas para explicar la ocurrencia de siniestro (siempre en comparación con sus respectivas categorías base, en caso de variables cualitativas) en caso de variables cualitativas, de sus respectivas categorías) fueron las siguientes: Exposure, los vehículos más antiguos y del tipo Utility, Coupe, Bus y Motorized Caravan. También todas las edades del conductor/a analizado. Por un lado, mediante la observación de los pendientes estimados de estas variables, podemos decir que a medida que incrementaba el tiempo de efecto de la póliza, también lo hacia la probabilidad de ocurrencia de siniestro. Esto se debe al signo positivo del pendiente para este regresor. Lo mismo obtenemos si el vehículo era tipo Coupe, Bus o Motorized Caravan. Es natural pensar que este tipo de vehículos al tener una composición muy distinta a lo que sería el turismo que todos tenemos en mente presenten más posibilidades de siniestro. De lo contrario, si analizamos los pendientes negativos de los regresores estadísticamente significativos apreciamos que los vehículos más antiguos decrementan la probabilidad de darse el suceso. Incluimos a este grupo, los vehículos tipo Utility. Antes hemos detectado que las tres categorías de automóvil que hacían incrementar la probabilidad de la endógena eran especialmente aquellas con aspecto aparentemente distinto a lo que sería el turismo frecuente. Si bien ahora, el tipo Utility es bastante parecido a este turismo frecuente que pensamos y por ello hace disminuir las posibilidades. Finalmente tenemos que a medida que el conductor cumple más años de vida, la probabilidad de sufrir siniestro es cada vez menor. Para acabar destacamos que, nuevamente, el género del conductor/a no afecta significativamente la endógena ni en este caso, tampoco el valor monetario del vehículo.

Pasamos al último capítulo donde hablamos sobre los accidentes graves ocurridos en ámbito español. Mediante el contraste de medias, apreciamos que eran los accidentes durante las horas que comprendía la noche cuando la probabilidad media de accidente grave era mayor, en comparación con el resto del día. Todas las categorías para la variable Hora resultaron ser significativas para determinar la media de probabilidad de tener accidente con muertos y/o heridos graves, en comparación con las observaciones de la 'Tarde'. Centrándonos en los meses del año, decimos que los accidentes producidos en verano eran los que tenían más probabilidad media de acabar siendo graves. Hay que destacar que el accidente del tipo frontal, donde el sentido de la vía doble y la zona era en una travesía, eran las categorías para cada variable ficticia que tenían asociada un valor más elevado de la endógena.

A partir de aquí, modelizamos el número de personas muertas y/o heridas graves en accidentes para poder ver qué variables regresores del modelo eran significativamente

estadísticas para explicar el comportamiento de dicha variable dependiente. Los resultados nos mostraron que todas las franjas horarias eran significativas, pero accidentes en la noche hacían incrementar el número de personas implicadas en accidente grave y contrariamente, los de mañana la disminuían. Tan solo los accidentes en verano explicaban significativamente el comportamiento de la variable dependiente y estos hacían incrementar el número de muertos y/o heridos graves. Las travesías tenían el mismo impacto para la variable dependiente, razonable pensarlo porque estas en la mayoría de los casos pasan por zonas pobladas en un núcleo urbano. Datos que nos revelan que el modelo es fiable es ver cómo vías primitivas como caminos vecinales y vías convencionales resultan significativas para explicar los accidentes graves, incrementando su valor debido a los positivos coeficientes que presentan, y de lo contrario, las autopistas, donde la seguridad vial es más elevada, no resultan significativas. Finalmente, resultan influyentes para nuestra variable dependiente todos los tipos de accidentes, exceptuando los múltiple o en caravana y las salidas de vía por la izquierda con despeñamiento. Todas las categorías para esta variable independiente hacen aumentar el valor de accidente grave, menos los atropellos a animales, que tienen coeficiente con signo negativo.

Explicadas las características que influyen en el número de personas involucradas en un accidente grave, ahora resaltaremos aquellas variables que son significativas para determinar la probabilidad que el accidente cause heridos graves y/o muertos. Vimos que, las categorías referentes a la noche y la mañana tenían una probabilidad de terminar ser accidente grave significativamente distinta en comparación a los de la tarde. Si nos fijamos en las estaciones del año, la categoría 'verano' tenía asociada una probabilidad significativamente superior a la categoría 'Otoño', hecho que no podemos decir de las otras dos estaciones. Las categorías 'Camino vecinal' y 'Vía Convencional' eran las únicas que presentaban tener una probabilidad de accidente grave significativamente superior en comparación a las autovías. La misma conclusión para los accidentes dentro de una intersección y en travesías, cogiendo como referencia los que se dieron fuera de intersección y vías interurbanas, respectivamente. Finalmente, todos los tipos de accidentes, menos los múltiple o en caravana y los atropellos a animales, eran significativos para explicar el hecho que sean graves en comparación a los accidentes por alcance, y a la vez, todos estos incrementaban las posibilidades estudiadas.

Categorías estadísticamente significativas, contraponiéndolas a su categoría base, que disminuyen la probabilidad de accidente son los accidentes ocurridos en la mañana y en intersecciones. Todas las demás categorías significativas hacían incrementar el valor de la probabilidad.

Queremos remarcar el hecho que las observaciones que se produjeron durante los meses de primavera e invierno no presentan una diferente probabilidad con significación estadística respecto los accidentes en otoño, a pesar de que el clima en estas estaciones tiene consecuencias apreciablemente negativas para el tráfico y la conducción.

Finalmente destacar que una potencial mejora del trabajo seria mediante la implementación de técnicas más adecuadas a la tipología de datos disponibles, por ejemplo, a los referentes al número de siniestros. En esta base de datos, tal y como hemos visto, predomina la abundancia de observaciones con valores nulos para las distintas variables, esto hace desestabilizar la muestra utilizada. Sin embargo, distintos modelos econométricos que serían más aptos para el tratamiento de nuestras bases de datos podrían ser: modelos de poisson, count data, binomial negativo inflado de ceros… pero dichas técnicas exceden de los conocimientos vistos a lo largo del grado.

V. BIBLIOGRAFÍA CONSULTADA

Guerrero Casas, F.M.; Melgar Hiraldo, M.C. (2005): Los siniestros en el seguro del automóvil: un análisis econométrico aplicado. Estudios de Economía Aplicada, vol. 23, núm. 1, pp. 355-375. Asociación Internacional de Economía Aplicada.

Celín Ortega; Fabian Alexander; Chérrez Miño, Mónica Cecilia; Gómez García, Antonio Ramon; González Gijón , Luis Alberto; Russo Puga, Marcelo; Suasnavas Bermúdes, Pablo Roberto;(2016): Caracterización de la Mortalidad por Accidentes de Tránsito en Ecuador, 2015. pp. 22-31. Universidad Tecnológica Indoamérica.

Wooldridge, J.M (2009) Introducción a la econometria: un enfoque moderno, Ed. CENGAGE Learning,4º edición.