

MODELOS ECONOMÉTRICOS PARA LA DETECCIÓN DEL FRAUDE EN EL SEGURO DEL AUTOMÓVIL

Tesis Doctoral presentada por Mercedes Ayuso Gutiérrez para la obtención del Título de Doctora. Dirigida por la Dra. Montserrat Guillén Estany.

Programa de Doctorado "Economía y Territorio: Análisis Cuantitativo". Bienio 1993-1995.

Dept. de Econometría, Estadística y Economía Española. Universidad de Barcelona.

Barcelona, mayo de 1998.

Como puede observarse, la concentración de probabilidades en la zona de clasificación del siniestro como no fraudulento se produce no sólo para los expedientes sin fraude detectado sino también para aquellos que presentan fraude. Este resultado, que no hace sino confirmar el obtenido en la Tabla 13 sirve de referencia para pensar que el criterio probabilístico utilizado por defecto en la delimitación de las áreas de clasificación no es el apropiado dado que, tras su aplicación, numerosos siniestros quedan clasificados en la categoría inadecuada. La solución partirá de optimizar el punto de corte utilizado, en el sentido de determinar un vector de probabilidades que amplíe las zonas de clasificación del siniestro como fraudulento. Seleccionando el punto definido por el vector (0.15, 0.15, 0.15), los resultados obtenidos han sido, los que se presentan en la Tabla 14.

Tabla 14. Frecuencias de clasificación (utilizando el criterio $c=0.15$)

	Elección predicha			Total
	Legítimo	Conseguir una indemnización	Incrementar una indemnización	
Elección Observada				
Legítimo	776	101	121	998
Conseguir una indemnización	106	121	75	302
Incrementar una indemnización	89	58	229	376
Total	971	280	425	1676

El modelo predice correctamente el comportamiento observado de 1126 siniestros, siendo, por tanto, el porcentaje de aciertos del 67.2%. Un total de 776 siniestros sin fraude detectado son clasificados de forma correcta (porcentaje de aciertos del 77.8%), y, en relación a los siniestros fraudulentos, la capacidad de predicción puede considerarse elevada. Así, un 40.1% de los siniestros con fraude destinado a conseguir una indemnización y un 60.9% de los siniestros destinados a incrementarla son clasificados correctamente dentro de sus categorías respectivas. El error cometido al clasificar siniestros fraudulentos como no fraudulentos es notablemente inferior al que aparecía en la Tabla 13. De esta forma, el porcentaje de siniestros que conteniendo el primer tipo de fraude son clasificados por el modelo como no fraudulentos es ahora del 35.1%; en relación al segundo tipo de fraude el porcentaje de siniestros incorrectamente clasificados como no fraudulentos es ahora del 23.7%.

La clasificación obtenida para los siniestros fraudulentos puede mejorarse al fijar el punto de corte en el vector de probabilidades (0.1, 0.1, 0.1). En este caso los resultados son los que se muestran en la Tabla 15.

Tabla 15. Frecuencias de clasificación (utilizando el criterio $c=0.10$)

	Elección predicha			Total
	Legítimo	Conseguir una indemnización	Incrementar una indemnización	
Elección Observada				
Legítimo	656	180	162	998
Conseguir una indemnización	69	154	79	302
Incrementar una indemnización	47	83	246	376
Total	772	417	487	1676

Los porcentajes de siniestros fraudulentos correctamente clasificados son del 51% y del 65.4%, respectivamente, para cada tipo de fraude. Los porcentajes de siniestros fraudulentos incorrectamente clasificados como no fraudulentos son ahora notablemente inferiores, pasando a ser del 22.8% y del 12.5%, respectivamente.

La selección de uno u otro criterio probabilístico dependerá, fundamentalmente, del margen de error que la compañía esté dispuesta a aceptar en relación a la clasificación errónea para los siniestros no fraudulentos. Si los costes de investigación no son elevados tal vez se decante por los resultados presentados en la segunda tabla, dado que, en ese caso, los beneficios obtenidos de una mayor detección de fraudes compensarán los costes de investigar un elevado número de siniestros en los que finalmente no existirá comportamiento fraudulento. En cualquiera de los casos, los resultados presentados en las Tablas 14 y 15 evidencian una elevada capacidad predictiva del Modelo 2 de forma que la aplicación del mismo para cuantificar la probabilidad de que el asegurado elija cada una de las alternativas planteadas queda justificada.

Una vez realizada la validación del modelo, estimada la probabilidad de existencia de cada tipo de fraude y determinada la existencia de una elevada bondad del ajuste, disponemos de la información necesaria para estimar la utilidad esperada por el individuo al elegir una alternativa concreta y para estimar el coste esperado del siniestro tras la aplicación por la compañía de una política de detección de fraude mediante el uso del Modelo 2. Ambos aspectos serán tratados en los apartados siguientes.

6.3.2.4 Estimación de la utilidad esperada por el individuo al elegir entre las alternativas

Al igual que veíamos para el Modelo 1, la estimación de la probabilidad de que el individuo elija cualquiera de las tres alternativas analizadas y, la determinación de la probabilidad de que la compañía clasifique correctamente cada uno de los siniestros dentro de la categoría correspondiente, permite ajustar la utilidad que el asegurado espera de su comportamiento.

Dadas W (riqueza inicial del individuo), P (prima pagada en la entidad), s_0 (coste real del accidente), s (indemnización incrementada tras la actuación fraudulenta) y, P' (prima pagada en otra entidad como consecuencia de la expulsión de la compañía inicial, en concepto de penalización tras la detección del fraude), la estimación de la utilidad esperada por el asegurado al cometer fraude para conseguir una indemnización es:

$$E[U(1)] = [0.649U(W-P-P') + 0.351U(W-P+s_0)].$$

Así, la probabilidad de que la compañía detecte la existencia de fraude es 0.649 (la probabilidad de que identifique fraude para conseguir una indemnización es 0.401 ya que $q_1 = 121/302$ y la de que detecte fraude para incrementar la indemnización es 0.248 ya que $q_1' = 75/302$). La probabilidad de que no se detecte fraude es, por tanto, de 0.351. Las estimaciones utilizadas han sido extraídas de los resultados presentados en la Tabla 14.

Siguiendo un proceso análogo al del apartado 6.3.1.4, la utilidad esperada por el individuo que comete fraude para incrementar la indemnización derivada de un siniestro, cuya ocurrencia es real y está cubierto por póliza, es:

$$E[U(2)] = [0.763U(W-P-P') + 0.237U(W-P+s)].$$

En este caso, la probabilidad de que la compañía detecte fraude es 0.763 (la probabilidad de que la entidad identifique fraude para incrementar la indemnización es 0.609 ya que $q_2 = 229/376$ y la de que detecte erróneamente fraude para conseguir una indemnización es 0.154 siendo $q_2' = 58/376$). Por tanto, la probabilidad de no detectar fraude es de 0.237.

La cuantificación de la utilidad esperada por el asegurado considerando, conjuntamente, las tres alternativas de elección será, por tanto:

$$E[U] = t_1[0.649U(W-P-P') + 0.351U(W-P+s_0)] + t_2[0.763U(W-P-P') + 0.237U(W-P+s)] + t_0U(W-P+s_0).$$

Las probabilidades t_0 , t_1 y t_2 corresponderán a las estimadas por el modelo de que el asegurado se comporte honestamente, de que cometa el primer tipo de fraude o el segundo, respectivamente.

6.3.2.5 Cuantificación del coste esperado por siniestro tras la aplicación de un modelo de control fraude

Consideremos las definiciones presentadas en el apartado anterior para t_0 , t_1 , t_2 , q_1 , q_1' , q_2 y q_2' . Sea q_0 la probabilidad de que la compañía clasifique correctamente un siniestro legítimo y q_0' la probabilidad de que lo clasifique como fraude para conseguir una indemnización. La estimación del modelo nos proporciona la información necesaria para evaluar todas estas probabilidades. Sea el coste del siniestro para la compañía el resultado de sumar la indemnización finalmente pagada y las cantidades monetarias invertidas en investigación. De esta forma, el coste esperado del siniestro es el que se muestra en la Tabla 16, según los diferentes casos.

Tabla 16. Coste esperado del siniestro

Tipo de siniestro	Clasificado como		
	Legítimo	Fraude para conseguir una indemnización	Fraude para incrementar la indemnización
Legítimo	s_0	c_1+s_0	c_1+s_0
Fraude para conseguir una indemnización	s_0	c_2	c_2
Fraude para incrementar la indemnización	s	c_2	c_2

En este caso, s_0 y s han sido definidos previamente y, c_1 y c_2 recogen el coste de investigar un siniestro sospechoso para el que no se detecta fraude (y que es, por tanto, legítimo) y el asociado a la identificación de fraude, respectivamente.

De nuevo y, con el objetivo de simplificar la interpretación y la nomenclatura, suponemos que los costes de investigación que deben sufragarse para la identificación como legítimo, de un siniestro en el que existe sospecha de fraude para conseguir una indemnización o, de un siniestro con sospecha de fraude para incrementar una indemnización, son iguales (c_1). Análogamente, la confirmación de la existencia de fraude lleva unos gastos asociados iguales,

en cualquiera de los casos, a c_2 . Teniendo en cuenta las mismas consideraciones que veíamos para el Modelo 1, cabe esperar que $c_1 \leq c_2$.

El coste total esperado del siniestro, $E(T)$, puede escribirse ahora como:

$$E(T) = t_1[q_1 c_2 + q_1' c_2 + (1 - (q_1 + q_1')) s_0] + t_2[q_2 c_2 + q_2' c_2 + (1 - (q_2 + q_2')) s] \\ + t_0[q_0 s_0 + q_0' (s_0 + c_1) + (1 - (q_0 + q_0')) (s_0 + c_1)].$$

Esta expresión es equivalente a:

$$E(T) = t_1[(q_1 + q_1')(c_2 - s_0) + s_0] + t_2[(q_2 + q_2')(c_2 - s) + s] + t_0[q_0 s_0 + (1 - q_0)(s_0 + c_1)].$$

Dado que todos los parámetros que aparecen en la expresión anterior son conocidos o, pueden estimarse a partir del modelo, es posible presentar un ajuste del coste esperado del siniestro.

Terminemos este apartado, y el Capítulo en sí, presentando un pequeño ejemplo en relación a la estimación del coste que la entidad espera del accidente (utilizaremos los resultados que aparecen en la Tabla 14).

La probabilidad de clasificar correctamente un siniestro legítimo es ahora 0.778 siendo $q_0 = 776/998$. El porcentaje de error en relación a la clasificación de los mismos es, por tanto, de 0.222. Como ya sabemos, para los siniestros con fraude dirigido a conseguir una indemnización, la probabilidad de acierto es 0.401, mientras que la probabilidad de que sean clasificados dentro de la tipología alternativa de fraude es 0.248. La probabilidad de clasificar como tal un siniestro con fraude dirigido a incrementar una indemnización es 0.609; siendo la probabilidad de clasificarlo como fraude dirigido a conseguir una indemnización de 0.154.

Sea s_0 el coste real del siniestro y $s = 1.5s_0$ la indemnización pagada por la entidad atendiendo a la declaración presentada por el asegurado tras la ocurrencia real del siniestro (bajo condiciones cubiertas en la póliza). Si asumimos que $c_1 = 0.10s_0$ y, $c_2 = 0.15s_0$, el coste total esperado estimado del siniestro será igual a:

$$E(T) = [t_1(0.45s_0) + t_2(0.47s_0) + t_0(1.02s_0)].$$

Supongamos que la probabilidad estimada de fraude para conseguir una indemnización (t_1) es 0.44 y la de fraude para incrementarla (t_2) es 0.30. El coste esperado del siniestro, tras la aplicación del modelo de detección es, aproximadamente, un 40% más pequeño que el que se derivaría de la no detección del fraude (pago de s_0). Ello supondría para este individuo un ahorro esperado para la entidad de $(1.00 - 0.60)s_0$, es decir $0.40s_0$.

El escenario que se acaba de presentar, de forma análoga al que presentábamos en el apartado 6.3.1.5, es extremo. Si un asegurado tiene probabilidad de no defraudar del 26%, lógicamente, el sistema que utiliza el Modelo 2 debe conducir a la conclusión de que, en términos de costes esperados, implementar dicho modelo como herramienta de clasificación y por lo tanto, de detección es rentable.

En casos más dudosos el beneficio esperado podría no ser tan rotundo. A ello deberíamos añadir que las probabilidades de detección se han estimado agregadamente y son optimistas (puesto que no se ha utilizado una submuestra de validación). Más aún, que una compañía clasifique un siniestro como fraudulento y que lo audite no implica necesariamente que descubra y logre desenmascarar el fraude.

7. APLICACIÓN EMPÍRICA II: MODELIZACIÓN LOGÍSTICA ANIDADA

7.1 Introducción

Cada individuo, a la hora de realizar una determinada elección, hace intervenir una serie de factores que, sin duda, condicionan la alternativa final seleccionada. La existencia de restricciones económicas, familiares, sociológicas,..., determinan su comportamiento, de forma que la toma de decisiones que realiza suele ir acompañada de una abstracción sobre las posibilidades existentes.

Pensar en aquellas situaciones en las que se elige, implica reflexionar, en la mayoría de los casos, mediante un proceso de decisión por etapas.

A la hora de comprar un vehículo (Koujianou, 1995) el individuo estructura mentalmente sus posibilidades. Primero decide entre si quiere comprar o no. Si decide comprar, puede pensar en un vehículo nuevo o en un vehículo de segunda mano. Quizá si opta por la segunda alternativa sus posibilidades se vean reducidas, pero si opta por la primera un amplio conjunto aparece a su alcance. A la decisión sobre la clase de vehículo que quiere (turismo, motocicleta,...) le seguirá la decisión sobre las características que busca en el mismo. Las preferencias sobre el lugar de fabricación, sobre el tamaño o sobre la cilindrada son algunos de los factores que permitirán continuar con el proceso de elección, de forma sucesiva, hasta llegar a la alternativa final seleccionada. La elección del modo de viajar es otro de los ejemplos utilizado en la literatura (Hensher, 1986; Greene, 1995) para analizar procesos de decisión por etapas.

Ahora bien, ¿qué implicaciones tiene la toma de decisiones de forma sucesiva en el diseño de un modelo que intente explicar los factores que la determinan?. ¿Cuál es la diferencia con lo que sería la modelización en base a alternativas finales, sin tener en cuenta etapas

intermedias?. En definitiva, ¿por qué podemos plantear una modelización alternativa a la presentada en el Capítulo 6, para cuantificar la probabilidad de existencia de fraude?.

Anidar significa tener en cuenta que las alternativas de elección pueden no ser independientes. En el Capítulo 6, se ha supuesto un enfoque que implica considerar una única etapa de decisión. El asegurado decide de forma directa entre no defraudar, defraudar en favor de sí mismo o defraudar para beneficiar a otra persona y, en la modelización de este comportamiento, únicamente se tienen en cuenta características relacionadas con diferentes aspectos que, en ninguno de los casos, son propios de una alternativa específica de elección. La información contenida en la muestra disponible no refleja, para ninguna de las observaciones, características propias de un determinado tipo de fraude. No disponemos de atributos de elección y, por lo tanto, únicamente podemos determinar qué factores o circunstancias generales del siniestro pueden tener mayor impacto en la explicación de comportamientos, desde la óptica de fraude.

El modelo presentado en el capítulo anterior supone la existencia de una elección múltiple y los resultados obtenidos evidencian que determinadas variables pueden tener una influencia diferente en la probabilidad de elegir cada una de las alternativas posibles. La existencia de niveles de significación diferentes o la presencia de coeficientes con signos contrarios para una misma variable en categorías alternativas permite obtener importantes conclusiones relacionadas, por ejemplo, con la información que la entidad debe utilizar para detectar ciertos comportamientos. Igualmente, puede orientar sobre la forma de encaminar la investigación sobre el fraude.

Sin embargo, ¿cómo podemos mejorar los resultados obtenidos teniendo en cuenta la existencia de niveles en el proceso de decisión?. La respuesta a esta cuestión constituye el objetivo de este último Capítulo. La estructuración para el mismo es la siguiente: en el apartado 7.2 presentamos el modelo logístico anidado como una forma alternativa a la modelización del fraude; en el apartado 7.3. rediseñamos la base de datos para hacerla susceptible de aplicación de este tipo de modelos y presentamos las variables introducidas en la modelización final; en el apartado 7.4 presentamos la estimación de los modelos.

7.2 Justificación de la aplicación de un modelo logístico anidado: solución al incumplimiento de la hipótesis de Independencia de Alternativas Irrelevantes

Cuando el asegurado piensa en la posibilidad de cometer o no fraude, es posible, que estructure mentalmente un proceso secuencial de elección: primero decide entre no defraudar o defraudar y si opta por la segunda alternativa, decide de qué forma hacerlo.

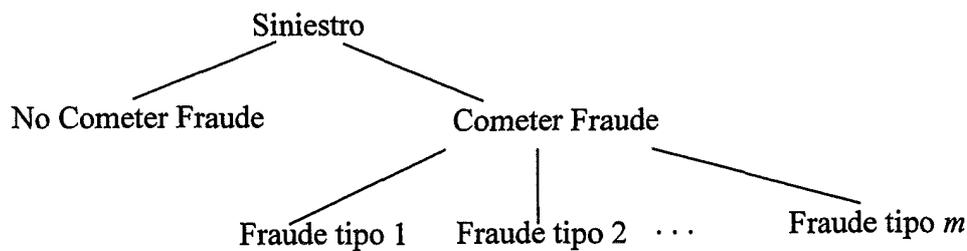


Figura 20

En este caso, existen m tipos diferentes de fraude (formas de defraudar).

En el nivel superior, los factores que intervienen para diferenciar entre comportamientos son comunes, en lo que a fraude se refiere. Determinadas características ayudarán a incrementar la sospecha de fraude, genéricamente hablando. El historial de siniestros del asegurado, la existencia de coincidencia de apellidos entre las partes, la intervención policial en el siniestro o la presencia de relatos estadísticamente verificados como sospechosos, son algunas de las variables que pueden ayudar a modelizar el comportamiento de fraude *versus* no fraude, pero ¿hasta qué punto pueden ayudar a diferenciar entre tipos de fraude?

Lo ideal sería disponer de variables que tomaran valores diferentes según el comportamiento fraudulento existente. En la misma línea, por ejemplo, que el “tiempo de espera” o el “coste monetario de viaje” son variables utilizadas en la modelización de la elección final del modo de viajar -avión, tren, autobús o coche- (ver Greene, 1995) por proporcionar información diferenciada para cada alternativa, la búsqueda de atributos específicos para cada tipo de fraude ha de permitir hacer lo propio en nuestra aplicación.

El funcionamiento del seguro del automóvil, sobre todo, desde el punto de vista de la aplicación de convenios, dicta unas pautas bastante claras respecto a las posibilidades de cobrar una indemnización. Así, si el asegurado posee cobertura de daños propios resulta complicado, *a priori*, elaborar conclusiones sobre fraude ya que, tanto puede actuar para beneficiarse a sí mismo como para beneficiar a un tercero. Ahora bien, si el asegurado posee

cobertura a terceros las interpretaciones pueden ser más exhaustivas. Ante esta situación, es posible pensar que si el asegurado admite la culpa del siniestro, la existencia de fraude pueda estar relacionada con un beneficio al contrario¹ mientras que si lo que pretende es beneficiarse a sí mismo un posible acuerdo llevará a que sea el contrario quien admita la culpa del accidente. El cruce de datos realizado para la variable *SUBFRAU* (tipo de fraude), la variable *DCULPA* (culpa del asegurado=1) y la variable *DRAMO* (cobertura a terceros=1) permite verificar estas conclusiones tal y como se manifiesta en las Tablas 17 y 18, mostradas a continuación:

Tabla 17. Frecuencias absolutas de siniestros fraudulentos según tipo de cobertura del asegurado y culpa

Existencia de <i>fraude a favor del asegurado</i> (Total muestra=299 casos)	
culpa del asegurado y cobertura a terceros	culpa del contrario y cobertura a terceros
71	216

Tabla 18. Frecuencias absolutas de siniestros fraudulentos según tipo de cobertura del asegurado y culpa

Existencia de <i>fraude a favor del tercero</i> (Total muestra=314 casos)	
culpa del asegurado y cobertura a terceros	culpa del contrario y cobertura a terceros
221	69

La culpa del siniestro y el tipo de cobertura son, por lo tanto, variables importantes a la hora de considerar la alternativa final elegida por el asegurado.

La fundamentación teórica de los modelos logísticos anidados, tal y como veíamos en el Capítulo 4 (apartado 4.3), relaja el cumplimiento de las hipótesis de independencia y

¹ Lógicamente esto puede estar sujeto a matizaciones y como veremos más adelante, en algunos casos es posible hablar de cobertura a terceros, culpa del asegurado y fraude en beneficio propio. Concretamente en la base de datos encontramos 71 casos en los que concurren estas circunstancias. Las razón viene fundamentada básicamente en aquellas situaciones en las que el asegurado admite su culpa para eludir casos excluidos en la póliza (en 63 casos se detecta este comportamiento). Hechos como los derivados de falsas declaraciones para beneficiar a familiares servirían como ejemplo y permiten justificar la existencia de un coeficiente positivo para la variable *dculpa* (culpa del asegurado) en la modelización presentada en el Capítulo anterior para la categoría *fraude a favor del asegurado* (a pesar, de que si recordamos, dicha variable no goza de significación estadística).

homocedasticidad para los términos de perturbación. Desde este punto de vista, admite el incumplimiento de la hipótesis de independencia de alternativas a nivel global.

Cuando en la modelización, junto al resto de variables (relacionadas con el asegurado, con el siniestro o con la póliza) tenemos en cuenta atributos específicos de cada elección es posible que la independencia de alternativas no se cumpla. La existencia de correlación puede venir provocada porque en la etapa inferior del árbol de decisión aparecen posibilidades de elección definidas en función de atributos propios pero también de variables comunes. Si en la definición del modelo somos capaces de agrupar categorías es posible que logremos demostrar la independencia de la mismas dentro del grupo al que pertenecen. A nivel global la hipótesis se incumplirá: el asegurado decide entre categorías que pueden formar parte de un mismo grupo de elección y, desde este punto de vista, parece lógico pensar en una posible correlación. Ahora bien, a nivel inferior, las categorías, definidas en función de sus propios atributos, deberán ser independientes, en el sentido de que aquellos factores que las hacen comunes quedan recogidos en el nivel superior.

Lógicamente es posible que modelizando globalmente, en el sentido de aplicar un modelo multinomial o condicional, se cumpla la independencia sin que sea necesario considerar una estructura jerárquica. Si al ejecutar el test de Hausman y McFadden (teniendo en cuenta únicamente las alternativas finales) no se puede rechazar la hipótesis nula de independencia, la aplicación de un modelo lógit multinomial o condicional queda justificada. En el Capítulo anterior aceptábamos ese cumplimiento bajo la aplicación de un multinomial², pero no teníamos en cuenta atributos de cada una de las elecciones. Será ahora, considerando características propias de cada alternativa, cuando justificaremos la utilidad de modelizar el comportamiento mediante regresión logística anidada.

7.3 Diseño de la base de datos y selección de las variables explicativas introducidas en la modelización

La aplicación de modelos logísticos anidados (al igual que los lógit condicionales) requiere una preparación apropiada de la base de datos empleada. A diferencia de lo que ocurre con los lógit multinomiales, cuando tenemos en cuenta características específicas de la elección, cada una de las observaciones de la muestra ha de estar adecuadamente preparada para recoger el comportamiento de variables con valores diferentes para cada una de las alternativas.

² Una posible solución al incumplimiento de la hipótesis de Independencia de Alternativas Irrelevantes en un modelo lógit multinomial es la aplicación de un probit multivariante (Greene, 1997).

Originalmente, tal y como hemos comentado en páginas anteriores, no disponemos dentro de la base de datos de variables que cambien de valor al considerar diferentes alternativas. Su uso, como matriz de N filas y K columnas, siendo N el número de individuos (siniestros) que hay en la muestra y K el número de variables explicativas consideradas, es apropiado en la aplicación de regresión logística multinomial.

Sin embargo, cuando planteamos un lógit anidado, el número de filas de la matriz de datos queda multiplicado por $m+1$, siendo $m+1$ el número de alternativas finales consideradas. El diseño de la matriz de datos para estimar un modelo lógit condicional nos permitirá, por su similitud, extraer las principales consideraciones sobre la estructura que ha de presentar la base de datos de cara a estimar un lógit anidado mediante el software especializado. Para verlo, planteemos un pequeño ejemplo.

Supongamos que disponemos de dos atributos (T , Z) para cada una de las elecciones finales consideradas en nuestro estudio (utilicemos, a modo de ejemplo, la clasificación asociada a no fraude, fraude a favor del asegurado y fraude a favor de un tercero). Supongamos que nuestra base de datos consta únicamente de información relativa a tres siniestros. El individuo realiza una elección entre tres alternativas, de las cuáles sólo elige una: supongamos que el primer asegurado elige la tercera alternativa (defrauda a favor del contrario), el segundo elige la primera (no defrauda) y el tercero la segunda (comete fraude en beneficio propio). La matriz de datos estará formada por 9 filas, con 3 bloques de tres filas cada uno. Dentro de cada bloque habrá información referida a los atributos considerados y a la variable dependiente. Esta última, en cada fila, tomará el valor 1 si la alternativa es elegida y 0 si no lo es. Así, para este caso hipotético la matriz de datos sería la que se especifica en la Tabla 19.

Tabla 19. Matriz de datos transformada

	Y	T	Z
$i=1 \Rightarrow$	0	$t_{1,1}$	$z_{1,1}$
	0	$t_{2,1}$	$z_{2,1}$
	1	$t_{3,1}$	$z_{3,1}$
$i=2 \Rightarrow$	1	$t_{1,2}$	$z_{1,2}$
	0	$t_{2,2}$	$z_{2,2}$
	0	$t_{3,2}$	$z_{3,2}$
$i=3 \Rightarrow$	0	$t_{1,3}$	$z_{1,3}$
	1	$t_{2,3}$	$z_{2,3}$
	0	$t_{3,3}$	$z_{3,3}$

En este caso Y es la variable dependiente (elección realizada) y T y Z , atributos de las elecciones finales ($t_{j,i}$ es el valor de T para el individuo i respecto a la alternativa j . Análogamente se define $z_{j,i}$). De esta forma, y generalizando, la matriz inicial formada por N filas pasa a estar formada por $N \cdot (m+1)$ (N bloques de $(m+1)$ filas) y K variables explicativas. En el modelo anidado, el diseño que ha de presentar la base de datos es el mismo que para el condicional pero con un matiz importante.

El modelo logístico anidado permite la definición de funciones de utilidad para cada uno de los niveles del árbol de decisión. La utilidad de cada una de las elecciones estará definida, tal y como vimos en el Capítulo 4, por una combinación de atributos y coeficientes de la regresión y por un término de error aleatorio, de forma que en la estructuración de la matriz de datos se ha de tener en cuenta el nivel donde se incorporan las variables de estudio.

En el árbol que aparece en la Figura 20 (particularizado, para continuar con el ejemplo que venimos utilizando, a tres posibilidades de elección en el nivel inferior) la utilidad de cada una de las alternativas finales vendrá definida en base a determinados atributos de las mismas. Así, la matriz de datos presentará una estructura similar a la que aparece en la Tabla 19. Sin embargo, el modelo consta también de funciones de utilidad asociadas a las alternativas intermedias (nofraude-fraude) y en su definición intervendrán atributos cuya variación ha de quedar adecuadamente recogida en la base de datos: tomarán valores constantes para las alternativas finales agrupadas dentro de una misma alternativa intermedia. Así, si P y Q son dos atributos específicos de las alternativas no fraude-fraude, la matriz de datos tendrá la forma que se presenta en la Tabla 20.

Tabla 20. Estructura de la Matriz de datos en un modelo anidado.

			T	Z	P	Q
raíz 1 (siniestro)	rama 1 1 (no fraude)	alt. final 1 1,1 (no fraude)	$t_{(1 1),i}$	$z_{(1 1),i}$	$P_{(1 1),i}$	$q_{(1 1),i}$
	rama 2 1 (fraude)	alt. final 1 2,1 (benef. propio)	$t_{(1 2),i}$	$z_{(1 2),i}$	$P_{(2 1),i}$	$q_{(2 1),i}$
		alt. final 2 2,1 (benef. tercero)	$t_{(2 2),i}$	$z_{(2 2),i}$	$P_{(2 1),i}$	$q_{(2 1),i}$

En la Tabla anterior, $t_{(1|1),i}$, $t_{(1|2),i}$, $t_{(2|2),i}$ corresponderían con $t_{1,i}$, $t_{2,i}$ y $t_{3,i}$, respectivamente, para $i=1,2,3$, en la Tabla 19, y $z_{(1|1),i}$, $z_{(1|2),i}$, $z_{(2|2),i}$, con $z_{1,i}$, $z_{2,i}$ y $z_{3,i}$, respectivamente.

Como puede observarse, los atributos para las elecciones intermedias toman el mismo valor en las alternativas finales pertenecientes a las mismas.

Nuestra matriz de datos se transforma, en base al proceso definido, en una matriz de grandes dimensiones. Concretamente, y para el Modelo 1A, esquematizado mediante el árbol de decisión de la Figura 21, el orden de la matriz es de 4833 filas por 28 columnas, siendo el número de filas el resultado de multiplicar por 3 (número de alternativas finales) las observaciones iniciales (1661 casos) y las columnas, el número de variables consideradas³ incluido el vector de constantes.

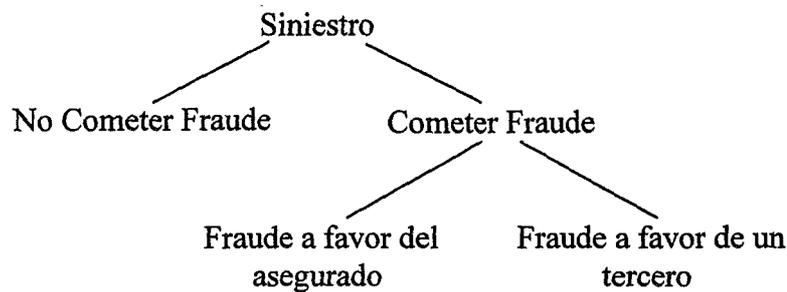


Figura 21

Para el Modelo 2A, representado gráficamente en la Figura 22, el orden de la matriz es aún mayor, siendo el número de filas igual a 5028 (1676 siniestros por 3 alternativas) y el número de columnas el mismo que en el modelo anterior.

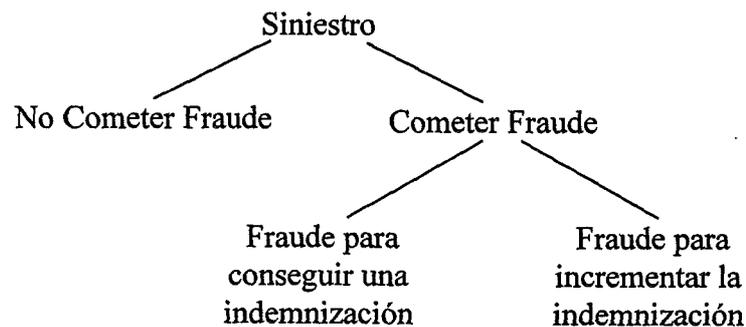


Figura 22

Sin embargo y, a pesar de disponer de información para un elevado número de variables, en ninguno de los casos poseemos, tal y como hemos comentado en páginas anteriores, de información específica asociada a cada uno de los comportamientos fraudulentos estudiados. ¿Cómo podemos generar atributos de cada una de las elecciones a partir de nuestra base de datos?. Señalando de antemano que lo deseable sería disponer de los mismos a partir de

³ Dichas variables aparecen definidas en el Capítulo 5.

información directamente recogida por la entidad, presentamos a continuación el proceso seguido en la definición de las variables que incluiremos en la modelización.

7.3.1 Especificación del Modelo 1A: definición de variables

Comencemos por los términos constantes. En un modelo lógit anidado, de la misma forma que en un condicional, la definición del término independiente supone la creación de una serie de constantes específicas de elección para cada una de las alternativas. En este sentido, es necesario incorporar en la base de datos un vector de variables dicotómicas para cada una de las posibilidades, que tomarán valor uno si la elección considerada es realizada y cero, en caso contrario⁴. De esta forma, y teniendo en cuenta las alternativas consideradas se ha procedido a la creación de tres nuevas variables:

SELEC: variable constante específica de elección de la primera alternativa considerada (no fraude). Toma el valor 1 en el registro correspondiente a la decisión de no defraudar y 0 en caso contrario.

AELEC: variable constante específica de elección de la segunda alternativa considerada (fraude a favor del asegurado en el modelo 1A y, fraude para conseguir una indemnización en el 2A). Toma el valor 1 en el registro correspondiente a la decisión de elegir la alternativa considerada y 0 en caso contrario.

TELEC: variable constante específica de elección de la tercera alternativa (fraude a favor del tercero en el modelo 1A y, fraude para incrementar la indemnización en el 2A). Toma el valor 1 en el registro correspondiente a la decisión de elegir la alternativa considerada y 0 en caso contrario⁵.

⁴ Cabe destacar que bajo la utilización del programa LIMDEP, v.7.0, y para un lógit condicional, no es necesario incorporar variables *dummies* específicas de elección en la base de datos. La incorporación en la especificación del modelo de la variable *one* permitirá obtener la estimación de m constantes específicas, siendo $m+1$ el total de alternativas consideradas. Sin embargo, en el lógit anidado es necesario incorporar la constante específica correspondiente dentro de la función de utilidad definida para cada una de las alternativas (en una de ellas tomará valor cero), a pesar de que el programa también proporciona una forma para evitar tener que crearlas dentro de la base de datos (ver Greene, 1997, págs. 516-517).

⁵ El diseño en la matriz de datos para las constantes específicas de elección, considerando los dos primeros individuos, será,

	<i>AELEC</i>	<i>SELEC</i>	<i>TELEC</i>
	1	0	0
	0	1	0
	0	0	1
	1	0	0
	0	1	0
	0	0	1

¿Qué ocurre con el resto de variables explicativas?. Bien, cuando en una base de datos aparecen, junto a atributos de cada una de las elecciones, variables relacionadas directamente con el individuo, es posible incluir éstas últimas dentro del modelo anidado (o, en su caso del condicional) interaccionándolas con las constantes específicas de elección. Así, cuando una variable es medida para el individuo y no para la elección, como variable sólo puede ser incorporada en el modelo usando el equivalente en términos de interacción con una *dummy*. De lo contrario, la variable es la misma para todas las elecciones y el modelo no puede ser estimado (Greene, 1997). En nuestro caso, no disponemos de ningún atributo y las variables son en la mayoría de los casos dicotómicas, lo, que sin duda, complica el proceso de estimación, fundamentalmente al aumentar la probabilidad de obtener matrices Hessianas singulares (debido a elevada presencia de ceros). Sin embargo, la definición de una serie de atributos para explicar el comportamiento en el nivel inferior del árbol de decisión y la interacción de variables originales con constantes específicas de elección para explicar el comportamiento en el nivel intermedio ha permitido, como veremos más adelante, la obtención de resultados interesantes tras la estimación de los modelos.

Veamos como hemos definido los atributos específicos de la elección. En el Modelo 1A (Figura 21) y teniendo en cuenta el comportamiento explicado en el apartado 7.2 para las variables que recogen la culpa del siniestro y el tipo de cobertura (*DCULPA* y *DRAMO*), los atributos considerados serán:

COBERTER: interacción de la variable *DRAMO* con la constante específica de elección para la tercera alternativa (*TELEC*). La variable vale uno cuando existiendo cobertura a terceros (*DRAMO=1*) se considera la elección de fraude a favor del tercero, siendo cero en caso contrario.

CULPAASE: resultado de interaccionar la variable *DCULPA* con la constante específica de elección para la tercera alternativa (*TELEC*). La variable toma valor uno cuando existiendo culpa del asegurado (*DCULPA=1*) en el siniestro se considera la elección de fraude a favor del tercero y cero, en caso contrario.

ANTC: interacción de la variable *ANTIGUO*⁶ con la existencia de cobertura a terceros (*DRAMO=1*), de culpa del contrario (*DCULPA=0*) y con la constante específica de elección para la segunda alternativa (*AELEC*). El objetivo es recoger la importancia que la antigüedad del vehículo asegurado puede tener en

⁶ Antigüedad en años del vehículo en el momento del siniestro.

la realización de fraude, pareciendo lógico interaccionarla con la constante de la segunda alternativa atendiendo a su propia definición. El hecho de multiplicarla por una variable dicotómica que identifica a la cobertura a terceros $DRAMO=1$ y otra que identifica que la culpa es de un tercero $DCULPA=0$, nos permite convertirla en un atributo específico de la elección fraude a favor del asegurado. Existiendo cobertura a terceros (daños al vehículo no cubiertos), la única forma de que esta tenga *a priori* poder explicativo en el fraude a favor del asegurado es interaccionándola con existencia de culpa de contrario. Así, la variable $ANTC$ toma el valor igual al número de años del vehículo cuando, existiendo cobertura a terceros y declarándose el contrario culpable del siniestro, se considera su influencia en la elección de fraude a favor del asegurado; vale cero en caso contrario.

$ANTC2$: cuadrado de la variable anterior, $ANTC$. El objetivo es recoger de forma adecuada la existencia de una influencia cuadrática para la variable considerada.

$OCURANT$: atributo resultado de interaccionar la variable $DEFOCEMI$ ⁷ con la constante específica de la elección fraude a favor del asegurado. En este caso, el criterio utilizado parte de suponer que cuando el siniestro ocurre entre las fechas de efecto y de emisión (por este orden) de la póliza, parece más lógico pensar en la existencia de fraude a favor del asegurado (por ejemplo, puede pensarse en situaciones en las que el agente retrocede el efecto de la póliza para beneficiar a un determinado asegurado). Al intervenir la parte aseguradora (no directamente, sino a través de agentes) resulta más difícil pensar en un posible beneficio a un tercero. El atributo tomará valor uno cuando el siniestro haya ocurrido entre la fecha de efecto y la fecha de emisión ($DEFOCEMI=1$) y consideremos el fraude en beneficio propio; valdrá cero cuando ello no ocurra (siniestro posterior a la emisión de la póliza o elección de otra alternativa).

Al haber definido estas variables de la forma anterior hemos considerado implícitamente y con consideraciones apriorísticas que influyen en una elección y no en la otra (en un tipo de fraude y no en el otro). Ello es restrictivo pero es la única forma de solucionar el problema de que el valor observado no varía entre las alternativas.

⁷ Variable dicotómica que toma valor 1 si la fecha de ocurrencia del siniestro es anterior o igual a la fecha de emisión de la póliza y posterior o igual a la fecha de efecto; toma valor 0 en caso contrario.

Continuando con las variables que introduciremos en la especificación del Modelo 1A, quedan por determinar aquellas que consideraremos en la definición de las funciones de utilidad para las alternativas intermedias. En este caso, y como resultado de interaccionar con la constante específica de elección para la rama de fraude ($1-SELEC$)⁸, se han definido las siguientes variables:

HISTC: variable que indica el número de siniestros asociados a la póliza anteriores al estudiado, en los registros correspondientes a la elección de fraude. Surge como resultado de multiplicar la variable $HISTSTR$ ⁹ por la constante de elección ($1-SELEC$), de forma que para cada individuo, la variable valdrá cero para la primera elección (no fraude) mientras que tomará el valor igual al número de siniestros en las dos alternativas de fraude consideradas (ver Tabla 20).

APELAPCC: variable que indica la existencia de alguna coincidencia de apellidos entre la parte asegurada (asegurado/propietario y/o conductor) y la parte contraria, en las alternativas de elección de fraude. Siguiendo el mismo criterio que en la variable anterior, surge como resultado de interaccionar la variable $DESTAF$ ¹⁰ con ($1-SELEC$), de forma que valdrá cero para la elección de no fraude y uno si existe coincidencia en los apellidos para las alternativas relativas a la ocurrencia de fraude.

SAUA: variable que recoge la intervención de policía en los siniestros para las alternativas correspondientes al fraude. Resulta de multiplicar la variable $SAUTDAD$ ¹¹ por ($1-SELEC$), y su disposición en la matriz de datos será análoga a la variable **APELAPCC**: valdrá cero para la primera elección (no fraude) mientras que tomará el mismo valor (uno o cero, en función de si ha intervenido policía o no) en las dos alternativas de fraude consideradas.

⁸ Podríamos haber realizado las interacciones con la constante específica de no fraude ($SELEC$). Sin embargo, y dado que los resultados no varían (únicamente cambia el signo de los parámetros que aparecen en el nivel intermedio) hemos preferido utilizar como constante ($1-SELEC$) de cara a facilitar la interpretación de los coeficientes en términos de probabilidad de elegir fraude frente a la probabilidad de actuar honestamente.

⁹ Número de siniestros asociados a la póliza anteriores al estudiado.

¹⁰ Variable dicotómica que toma valor 1 si existe coincidencia de apellidos entre la parte asegurada en la entidad (asegurado/propietario/conductor) y la parte contraria; vale 0 en caso contrario.

¹¹ Variable dicotómica que toma valor 1 si hay intervención de policía en el siniestro; toma valor 0 en caso contrario.

RELSTRO: interacción entre la variable *DRELATO*¹² y (*1-SELEC*). Creada bajo el objetivo de recoger la presencia de los relatos sospechosos en las alternativas del nivel intermedio, tomará valor cero para la elección de no fraude y valdrá uno o cero (según si aparecen o no los relatos en la declaración del asegurado) en las alternativas de fraude (el mismo valor para las dos posibilidades consideradas).

El nuevo diseño de la base de datos teniendo en cuenta las constantes específicas de cada elección, los atributos definidos para las alternativas finales consideradas y las variables explicativas introducidas en el nivel intermedio del árbol de decisión conduce a una matriz de 4833 filas, teniendo en cuenta también, en relación a las constantes específicas de elección, la definida como (*1-SELEC*). Esta matriz es la que utilizaremos en la estimación del primer modelo (Figura 21).

7.3.2 Especificación del Modelo 2A: definición de variables

Siguiendo el proceso presentado hasta ahora, ¿qué variables consideraremos en la especificación del modelo 2A (Figura 22). La creación de atributos específicos que definan comportamientos fraudulentos como los que aparecen en este segundo árbol de decisión ha resultado más complicada que en el caso anterior. A diferencia de lo que hemos visto para el modelo 1A, las variables *DCULPA* y *DRAMO* no tienen por que ayudarnos ahora a crear escenarios que faciliten la identificación de características propias de cada tipo de fraude. En la realización de fraude para conseguir una indemnización o para incrementarla pueden intervenir múltiples factores y las normas que regulan el contrato del seguro automovilístico¹³, aunque siempre presentes, puede que no tengan tanta importancia como a la hora de diferenciar entre fraude a favor del asegurado o fraude a favor del contrario. En base a ello, la definición de atributos para las alternativas finales se ha realizado introduciendo una serie de supuestos, que sin alejarse del mecanismo asegurador, tal y como veremos a continuación, no dejan de gozar de un cierto matiz de subjetividad.

Las constantes específicas de elección que incorporaremos en el modelo ya han sido definidas anteriormente (*SELEC*, *AELEC*, *TELEC*)¹⁴. De la misma forma que en el modelo 1A, la incorporación en la matriz de datos de la constante (*1-SELEC*) permitirá realizar la

¹² Variable dicotómica que toma valor 1 si en la declaración del asegurado aparecen algunos de los términos de descripción de lo ocurrido sospechosos; vale 0 en caso contrario.

¹³ Fundamentadas, básicamente, en la determinación de la culpa del siniestro y en las coberturas contratadas por el asegurado.

¹⁴ Hemos mantenido los mismos nombres para las constantes específicas de elección teniendo en cuenta su definición para la primera (no fraude), la segunda (fraude para conseguir una indemnización) y la tercera categoría (fraude para incrementar una indemnización), respectivamente.

interpretación de los coeficientes del nivel intermedio del árbol de decisión desde la óptica de la honestidad como categoría de referencia.

La creación de atributos para las alternativas ahora planteadas se ha realizado previa incorporación de los dos supuestos siguientes:

- la ocurrencia de un siniestro con aceptación de culpa por parte del contrario sirve, en algunos casos, para que el asegurado intente incrementar la indemnización que realmente le corresponde (reparación de daños anteriores del vehículo, alargamiento de días de baja,...), y
- la experiencia en siniestros del asegurado como miembro de una determinada entidad ayuda a que éste conozca el funcionamiento de la misma en relación a la investigación habitual realizada (peritaje, control de bajas médicas,...), de forma que aumenta su habilidad para tratar de conseguir una mayor indemnización.

Así, podemos definir:

CULPACON: directamente relacionado con el primero de los supuestos establecidos, este atributo se ha creado mediante la interacción de la variable que recoge la presencia de culpa de contrario en el siniestro ($DCULPA=0$) y la constante específica de elección asociada a la tercera alternativa (*TELEC*). De esta forma, valdrá 1 cuando existiendo culpa de contrario, se considere el fraude dirigido a incrementar la indemnización derivada de un siniestro; valdrá 0 en caso contrario.

HISTC1: creado en base al segundo de los supuestos considerados, este atributo recoge la interacción entre el número de siniestros asociados a la póliza anteriores al estudiado (variable *HISTSTR*) y la constante de elección *TELEC*. Corresponde al número de siniestros cuando el asegurado considere la posibilidad de defraudar para aumentar la indemnización que le corresponde tras el siniestro y será cero en caso contrario.

OCURANTI: atributo obtenido al interaccionar la variable *DEFOCEMI* con la constante específica de elección de fraude para conseguir una indemnización (*AELEC*). Retomando la explicación presentada para la variable *OCURANT* en el Modelo 1A parece lógico pensar que, cuando el siniestro ocurre entre la fecha de efecto

y la fecha de emisión de la póliza, el fraude cometido es para conseguir una indemnización. De esta forma, la variable definida valdrá uno cuando el siniestro se produzca entre las fechas mencionadas, en el orden presentado ($DEFOCEMI=1$) y el asegurado considera la segunda alternativa; valdrá cero cuando el siniestro sea posterior a la fecha de emisión de la póliza o, lógicamente, cuando la alternativa final sea otra.

Una vez definidas las variables a introducir en la especificación de las funciones de utilidad para cada una de las alternativas que aparecen en el nivel inferior del árbol de decisión, únicamente nos quedan por definir aquellas que intervendrán en el nivel intermedio. Dado que las variables introducidas para este nivel en el Modelo 1A, *APELAPCC*, *SAUA* y *RELSTRO*, poseen carácter genérico de cara a explicar cualquier tipo de comportamiento fraudulento, han sido seleccionadas también para especificar las funciones de utilidad asociadas a no fraude y fraude, respectivamente, en el Modelo 2A. Su definición es, por tanto, la presentada en páginas anteriores, recordando que al ser el resultado de interaccionar determinadas variables con la constante, *1-SELEC*, los coeficientes estimados que las acompañan deberán ser interpretados en términos de una mayor o menor probabilidad de aparición de fraude.

Como resultado del proceso seguido, la matriz de datos que utilizaremos en la estimación de los Modelo 2A (Figura 22), es una matriz de 5028 filas.

Una vez analizado el diseño que ha de presentar la base de datos para hacerla susceptible de modelización logística anidada y, definidas las variables que introduciremos en la especificación de los modelos, pasamos a presentar los resultados obtenidos en la estimación.

7.4 Estimación de los modelos anidados

La estimación de los modelos logísticos anidados suele realizarse bien, mediante el método de Máxima Verosimilitud Completa¹⁵ (ver por ejemplo, Greene, 1997), bien, utilizando el método de estimación Secuencial o Máxima Verosimilitud Limitada¹⁶ (Amemiya, 1985; Greene, 1995, 1997).

Mientras que las dificultades computacionales asociadas a la utilización de Máxima Verosimilitud Completa en modelos con más de tres niveles de decisión sugieren la

¹⁵ Full Information Maximum Likelihood (FIML).

¹⁶ Two Step Maximum Likelihood o Limited Information Maximum Likelihood (LIML).

utilización de estimación Secuencial (Koujianou, 1995), el hecho de que los estimadores obtenidos a partir del primer método de estimación sean más eficientes, en términos asintóticos, le convierten en el criterio de estimación utilizado, por defecto, en modelos de dos o tres niveles (McFadden, 1983; Greene, 1995).

¿En qué consisten ambos métodos de estimación?

Como vimos en el Capítulo 4, la probabilidad de que un individuo elija una determinada alternativa teniendo en cuenta una estructura jerárquica o árbol de decisión a dos niveles puede calcularse a partir de la expresión:

$$P_{i(cj)} = P_{i(j|c)}P_{i(c)}, \quad (7.1)$$

en donde, $P_{i(cj)}$ es la probabilidad de que el individuo i elija la alternativa cj y se obtiene a partir del producto entre la probabilidad de que elija la alternativa intermedia c , $P_{i(c)}$, y la probabilidad condicionada de que elija j una vez ha elegido c , $P_{i(j|c)}$.

La estimación del modelo anidado por Máxima Verosimilitud Completa permite obtener valores para todos los coeficientes de la regresión simultáneamente, maximizando la función de verosimilitud incondicional (en términos logarítmicos) definida anteriormente, o sea, haciendo máxima la expresión,

$$\ln L = \sum_{i,c,j} Y_{i(cj)} \ln P_{i(cj)} = \sum_{i,c,j} Y_{i(cj)} (\ln P_{i(j|c)} + \ln P_{i(c)}). \quad (7.2)$$

Como resultado del proceso obtendremos un vector de estimadores para cada nivel del árbol de decisión, es decir, para los parámetros β (como veíamos en el Capítulo 4, aparecen multiplicando a los atributos específicos de cada una de las elecciones finales) y para los parámetros α que aparecen multiplicando a las variables explicativas incluidas en la especificación de las funciones de utilidad de las alternativas intermedias. Asimismo, obtendremos los estimadores de los coeficientes de los valores inclusivos ($1-\sigma$), teniendo en cuenta que para la alternativa de no fraude dicho coeficiente estará fijado en la unidad dado que se trata de una "rama" con una única elección.

La estimación Secuencial o en dos etapas consiste en obtener los coeficientes estimados mediante el uso de sucesivos lógit condicionales.

En primer lugar, la maximización de la función de verosimilitud condicional (en términos logarítmicos) definida por,

$$\ln L_c = \sum_{i,j} Y_{i(cj)} \ln P_{i(j|c)} = \sum_{i,j} Y_{i(cj)} \ln \left[\frac{e^{\beta' X_{i(cj)}}}{\sum_{J=1}^{m_c} e^{\beta' X_{i(cJ)}}} \right] = \sum_{i,j} Y_{i(cj)} \ln \left[\frac{e^{\beta' X_{i(cj)}}}{e^{I_{i(c)}}} \right],$$

permitirá obtener la estimación del vector de coeficientes β , considerando el modelo como un simple lógit condicional siendo $X_{i(cj)}$ las observaciones de las explicativas para el individuo i -ésimo relativas a la alternativa cj . A partir de los resultados obtenidos y de las observaciones para los atributos se calcularán los valores inclusivos¹⁷ para las “ramas” (en nuestro caso, para la rama de fraude).

En una segunda etapa se obtendrán los estimadores para los coeficientes α y para $(1-\sigma)$ tratando, de nuevo, la elección en el nivel intermedio (entre las “ramas”) como si de un lógit condicional se tratara, ahora con variables $Z_{i(c)}$ e $I_{i(c)}$ (variables incluidas en la especificación de las funciones de utilidad para no fraude y fraude y, valor inclusivo, respectivamente). De esta forma el modelo condicional estimado en la segunda etapa especifica un probabilidad de elección de la alternativa intermedia c para el individuo i -ésimo:

$$P_{i(c)} = \frac{e^{\alpha' Z_{i(c)} + (1-\sigma) I_{i(c)}}}{\sum_{b=1}^C e^{\alpha' Z_{i(b)} + (1-\sigma) I_{i(b)}}} = \frac{e^{\alpha^* Z_{i(c)}^*}}{\sum_{b=1}^C e^{\alpha^* Z_{i(b)}^*}} \quad (7.3)$$

¹⁷ Recordemos que el valor inclusivo para cada individuo se calcula según la expresión siguiente y su valor se estima tras estimar β ,

$$I_{i(c)} = \ln \left[\sum_{J=1}^{m_c} e^{\beta' X_{i(cJ)}} \right]$$

En esta última expresión α^* es el vector de coeficientes de la regresión, incluido el asociado al valor inclusivo (incluido como una variable más dentro del vector de explicativas, $Z_{i(c)^*}$). El hecho de que el valor inclusivo, calculado en la primera etapa teniendo en cuenta valores estimados para los coeficientes, sea incorporado como una variable del modelo hace que el proceso fracase en la obtención de estimadores consistentes para la matriz de covarianzas en la segunda etapa, siendo necesaria su corrección (McFadden, 1981; Amemiya, 1985; Greene, 1995). La obtención de estimadores de los parámetros consistentes (aunque no eficientes) y las ventajas computacionales hacen atractivo el uso de este método de estimación.

No obstante, para nuestro caso, la posibilidad de realizar la estimación por Máxima Verosimilitud Completa (trabajamos únicamente con dos niveles) y las ventajas que este método proporciona en términos de eficiencia nos ha llevado a utilizarlo en el proceso de estimación.

7.4.1 Estimación logística anidada de la probabilidad de existencia de fraude a favor del asegurado o a favor de un tercero

Pasamos a continuación a mostrar los resultados obtenidos al estimar el modelo asociado al árbol de decisión que aparece en la Figura 21, considerando, por tanto, que el individuo realiza la toma de decisiones por etapas: el asegurado decide en primer lugar entre comportarse honesta o fraudulentamente, y si opta por la segunda alternativa, decide el tipo de fraude a cometer.

La cuantificación de la probabilidad de que el individuo decida no defraudar, defraudar a favor de sí mismo o a favor de un tercero (variable dependiente del modelo), supondrá, a diferencia de lo que ocurría con el lógit multinomial, realizar una estimación de las probabilidades de elección, por niveles. Así, para cada uno de los expedientes analizados, el producto entre las probabilidades condicionadas estimadas de elección en el nivel inferior del árbol de decisión y las correspondientes probabilidades ajustadas para el nivel intermedio permitirá determinar la probabilidad de que el individuo elija cada una de las tres alternativas consideradas. La elección predicha por el modelo será aquella que presente una mayor probabilidad estimada, aunque este criterio podemos modificarlo.

7.4.1.1 Análisis de la significación individual y global del modelo

Dado que el cálculo de cada una de las probabilidades de elección vendrá determinado por los coeficientes estimados para el modelo logístico anidado, presentamos a continuación los resultados obtenidos, analizando tanto la significación individual de los mismos, como la del modelo globalmente considerado.

La estimación del modelo, como hemos comentado en el apartado anterior, se ha realizado por el método de Máxima Verosimilitud Completa. La muestra está formada, tal y como ya sabemos, por 1611 siniestros (ver apartado 6.2), lo que supondrá estimar el modelo con un total de 4833 observaciones, al tener en cuenta tres posibles alternativas de elección para cada uno de los individuos.

La existencia de estratificación endógena en la base de datos ha sido considerada inicialmente en el proceso de estimación. Las observaciones de la variable dependiente han sido ponderadas siguiendo el mismo criterio que veíamos para el lógit multinomial (apartado 6.2). De esta forma, para los casos de no fraude se ha utilizado un factor de ponderación de 1.26 y para los casos de fraude a favor del asegurado y fraude a favor del tercero, respectivamente, uno de 0.58. Sin embargo, la obtención de una matriz Hessiana singular para la log-verosimilitud ponderada del modelo, ha impedido corregir la matriz de varianzas y covarianzas para los estimadores de la forma que veíamos en el Capítulo anterior, cuando hablábamos de estimación corregida por “muestreo basado en la elección” (apartado 6.2). A pesar de que la solución se ha encontrado aplicando el estimador de la matriz de covarianzas normalmente utilizado cuando el algoritmo de iteración es el D.F.P.¹⁸, los productos externos de las primeras derivadas de la log-verosimilitud (Greene, 1995), hemos creído conveniente presentar también los resultados obtenidos al estimar el modelo sin introducir ponderaciones para la variable dependiente. Lógicamente, y para este último caso, el hecho de trabajar con la función de verosimilitud sin ponderar facilitará el proceso de estimación. La significación individual y global del modelo prácticamente no sufrirá variaciones en relación al modelo ponderado.

Las variables explicativas introducidas en la especificación del modelo han sido comentadas en el apartado anterior. Un total de 12 variables han sido utilizadas para definir la elección final realizada, entre ellas se han considerado constantes específicas de elección, atributos asociados a las alternativas y variables explicativas propias del individuo o del siniestro. La característica principal de todas las variables utilizadas continua siendo, al igual que comentábamos para la especificación del modelo multinomial, su inmediatez: la compañía dispone de información referente a las mismas en un breve periodo tras la comunicación del siniestro ya que, en algunos casos, simplemente ha de analizar el contenido del parte de declaración del accidente (culpa del siniestro, intervención policial, coincidencia de apellidos, existencia de descripciones sospechosas) y, en otros, estudiar la póliza en cuestión (tipo de cobertura, fecha de efecto y de emisión, antigüedad del vehículo, historial de siniestros).

¹⁸ Davidon, Fletcher, Powell. La estimación no lineal de los parámetros del modelo se ha realizado mediante la aplicación de este algoritmo de iteración. Una explicación exhaustiva del mismo, usado normalmente cuando la función de verosimilitud adopta formas complicadas, puede encontrarse en Greene, 1993, 1995.

La especificación del modelo, atendiendo a la definición de funciones de utilidad¹⁹ para cada una de las alternativas²⁰, en cada uno de los niveles de decisión, ha sido la siguiente:

$$U(A1, A2, A3) = \beta_1 AELEC + \beta_2 COBERTER + \beta_3 CULPAASE + \beta_4 ANTC \\ + \beta_5 ANTC2 + \beta_6 OCURANT + e_j$$

$$U(R1, R2) = \alpha_1 (1-SELEC) + \alpha_2 HISTC + \alpha_3 APELAPCC + \alpha_4 SAUA \\ + \alpha_5 RELSTRO + e_j$$

siendo,

A1= alternativa 1= no fraude;

A2= alternativa 2= fraude *a favor del asegurado*;

A3= alternativa 3= fraude *a favor del tercero*;

R1= rama 1= no fraude;

R2= rama 2= fraude,

y,

$e_j, e_{j'}$, con $j=1,2,3$ (alternativas finales) y $j'=1,2$ (alternativas intermedias), los términos de aleatoriedad.

Los coeficientes estimados quedan recogidos en la siguientes tablas, presentándose los resultados para el modelo ponderado y sin ponderar, respectivamente.

¹⁹ En el Capítulo 3 comentábamos la derivación del modelo lógit anidado a partir de la teoría de la maximización de utilidad aleatoria (McFadden, 1978).

²⁰ En realidad, la función de utilidad se define separadamente para cada una de las alternativas. Sin embargo y, dado que para explicar cada una de las elecciones usamos las mismas variables (con la salvedad, lógicamente, de las constantes específicas de elección), podemos sintetizar la especificación del modelo utilizando una expresión genérica para la función de utilidad asociada a cada uno de los niveles (Greene, 1995).

Tabla 21. Estimación del modelo lógit anidado con ponderaciones
[MODELO 1A]

variable	coeficiente	t-test	P-valor
<i>Nivel inferior:</i>			
CONSTANTE	1.573	4.606	0.000
COBERTER	0.689	2.705	0.007
CULPAASE	2.006	8.008	0.000
ANTC	0.070	2.039	0.041
ANTC2	-0.004	-1.882	0.060
OCURANT	0.867	1.697	0.090
<i>Nivel superior:</i>			
CONSTANTE	-5.990	-8.540	0.000
HISTC	0.155	4.385	0.000
APELAPCC	1.414	5.073	0.000
SAUA	-1.624	-5.620	0.000
RELSTRO	0.509	3.686	0.000
<i>Valores Inclusivos:</i>			
NO FRAUDE	1.000	----	----
FRAUDE	1.835	5.614	0.000
número de observaciones ^a	4833	Chi-cuadrado	1264.489
ln función verosimilitud	-909.315	grados de libertad	12
ln verosimilitud restringida	-1541.559	nivel signif.	0.000

^a Cada individuo es replicado una vez por cada elección.

Tabla 22. Estimación del modelo lógit anidado sin ponderar
[MODELO 1A]

Variable	Coefficiente	t-test	P-valor
<i>Nivel inferior:</i>			
CONSTANTE	1.495	5.441	0.000
COBERTER	0.654	3.072	0.002
CULPAASE	2.018	10.485	0.000
ANTC	0.070	2.471	0.013
ANTC2	-0.004	-2.213	0.027
OCURANT	0.885	2.094	0.036
<i>Nivel superior:</i>			
CONSTANTE	-5.098	-9.031	0.000
HISTC	0.161	5.039	0.000
APELAPCC	1.393	5.209	0.000
SAUA	-1.570	-6.855	0.000
RELSTRO	0.502	4.182	0.000
<i>Valores Inclusivos:</i>			
NO FRAUDE	1.000	----	----
FRAUDE	1.822	7.116	0.000
número de observaciones ^a	4833	Chi-cuadrado	630.727
ln función verosimilitud	-1226.196	grados de libertad	12
ln verosimilitud restringida	-1541.559	nivel signif.	0.000

^a Cada individuo es replicado una vez por cada elección.

El modelo, tanto ponderado como sin ponderar, es globalmente significativo. Para el primer caso, el estadístico Chi-cuadrado es igual a 1264.5 y para el segundo, 630.7. Su comparación con el valor de las tablas para una χ^2 con 12 grados de libertad²¹ nos lleva a aceptar la significación global del modelo al 1%, en ambos casos.

Los resultados obtenidos para los coeficientes de la regresión individualmente considerados (ver Tabla 21), permiten extraer conclusiones interesantes en relación a la importancia que determinadas variables tienen en la explicación del fraude y en sus diferentes formas de manifestarse. La aparición de un comportamiento esperado en signos y el hecho de que todos los parámetros sean estadísticamente significativos son evidencias de una especificación correcta del modelo, teniendo en cuenta la existencia de niveles de decisión. La definición de atributos específicos de una elección ayuda a determinar qué circunstancias, relacionadas con la póliza o con el siniestro, gozan de mayor importancia a la hora de intentar detectar un tipo concreto de fraude. El hecho de que éstos se definan a partir de la interacción con la constante específica de elección de una determinada alternativa, implica interpretar los resultados en términos de la misma.

Teniendo en cuenta que los coeficientes estimados muestran un comportamiento prácticamente análogo para la modelización ponderada y sin ponderar, pasamos a interpretar los resultados obtenidos (de forma simultánea para ambos modelos), comenzando por el nivel inferior del árbol de decisión. Todas las comparaciones en términos de probabilidad relacionan un tipo de fraude con el otro al no existir categoría de referencia por la propia especificación del modelo.

La existencia de un valor positivo para el coeficiente que acompaña a la constante específica de elección de fraude a favor del asegurado (significativo al 1%) permite señalar, que cuando el resto de variables consideradas en el modelo son nulas, aumenta la probabilidad de que el individuo elija la segunda alternativa en relación a la tercera²².

La contratación en póliza de cobertura a terceros está positivamente relacionada con la existencia de fraude a favor de un tercero. La existencia de un coeficiente positivo y estadísticamente significativo (al 1%) para la variable *COBERTER* así lo confirma, de manera que cabe esperar que cuando el asegurado ha contratado el seguro para cubrir básicamente la responsabilidad civil (obligatoria y/o voluntaria) derivada de un siniestro, sin cubrir los daños

²¹ Diferencia entre los grados de libertad del modelo general (con todas las variables explicativas) y del modelo restringido (incorpora la hipótesis nula de que todos los coeficientes son cero).

²² Lógicamente, para el caso de no fraude, la probabilidad condicionada siempre será uno.

ocasionados a su vehículo, aumenta la probabilidad de que el fraude cometido sea para favorecer al contrario.

Un comportamiento análogo se observa para la variable *CULPAASE*. La existencia de un coeficiente positivo y estadísticamente significativo (al 1%) confirma que cuando el asegurado acepta la culpa del siniestro aumenta la probabilidad de que cometa fraude para beneficiar al contrario. Lógicamente, este resultado es más fácil de interpretar si tenemos en cuenta que en la mayoría de los casos la cobertura contratada por el asegurado es a terceros, y que, por tanto, parece más difícil pensar en un posible beneficio para sí mismo²³.

La antigüedad del vehículo está positivamente asociada a la propensión al fraude a favor del asegurado (existiendo cobertura a terceros y siendo el contrario quien admite la culpa del siniestro). El coeficiente para la variable *ANTC* es estadísticamente significativo (al 5%) y su signo positivo permite afirmar que la probabilidad de que el asegurado cometa fraude a su favor, por ejemplo, reparando daños anteriores al siniestro (más frecuentes cuanto más antiguo es el vehículo), aumenta con la edad del vehículo. No obstante, el hecho de que el coeficiente que acompaña a la variable *ANTC2* sea estadísticamente significativo (al 10% en el modelo ponderado y al 5% en el no ponderado) evidencia que la variable que recoge la antigüedad del vehículo presenta una tendencia no lineal y su signo negativo permite interpretar que a partir de una determinada antigüedad (punto de inflexión de la parábola) la probabilidad de existencia de fraude en beneficio propio disminuye (cuando el vehículo es muy antiguo su valor venal es muy bajo y, por lo tanto, es posible que el asegurado no obtenga ninguna ventaja al cometer fraude y decide beneficiar al contrario).

Por último, cuando el siniestro ocurre entre la fecha de efecto y la fecha de emisión de la póliza aumenta la probabilidad de que el asegurado haya cometido fraude en beneficio propio respecto a un tercero. El signo positivo para el coeficiente que multiplica a la variable *OCURANT* permite confirmar lo anterior, siendo estadísticamente significativo tanto en el modelo con ponderaciones para la variable dependiente (al 10%) como en el modelo sin ponderar (al 5%).

¿Qué ocurre con las variables explicativas de la existencia o no de fraude, en el nivel intermedio del árbol de decisión?

²³ La existencia de casos en los que el asegurado, a pesar de tener contratada únicamente cobertura a terceros admite la culpa del siniestro para actuar en beneficio propio ya ha quedado argumentada anteriormente.

Los resultados obtenidos para los coeficientes que acompañan a las variables introducidas en la especificación de las funciones de utilidad para no fraude y fraude, genéricamente considerado, han sido también los esperados, tanto en signos como en significación estadística (todos ellos son estadísticamente significativos al 1%). Teniendo en cuenta que todas las variables han sido definidas mediante la interacción con la constante específica de elección de fraude, la interpretación de los coeficientes estimados debe hacerse en términos de aumento (signo positivo) o reducción (signo negativo) de la probabilidad de aparición de comportamientos fraudulentos, frente a los no fraudulentos, ya que la ausencia de fraude se considera la categoría de referencia.

La constante específica de elección de fraude presenta un coeficiente estimado con signo negativo.

El historial de siniestros (anteriores al estudiado) que el asegurado posee en la compañía está directamente relacionado con la aparición de fraude tal y como se desprende de la existencia de un coeficiente estimado positivo y significativo para la variable *HISTC*.

Un comportamiento análogo se observa para las variables que recogen la coincidencia de algún apellido entre la parte asegurada y la parte contraria, y la presencia en la declaración del asegurado, de los relatos sospechosos (variables *APELAPCC* y *RELSTRO*, respectivamente).

Por último, la intervención de policía en el siniestro aparece inversamente relacionada con la aparición de fraude en el siniestro. El coeficiente negativo para la variable *SAUA* permite señalar una reducción en la probabilidad de existencia de comportamiento fraudulento cuando existe atestado del accidente.

Una vez interpretados los coeficientes estimados para todas las variables incluidas en la especificación del modelo anidado, únicamente nos quedan por analizar los coeficientes ajustados para los valores inclusivos. En realidad, la interpretación se limita al coeficiente estimado para la rama de fraude, pues, la existencia de una única alternativa en la rama de no fraude ha supuesto fijar el coeficiente del valor inclusivo correspondiente en la unidad. Como se desprende de la tabla de resultados (Tabla 21), la estimación obtenida para fraude (estadísticamente significativa al 1%) permite obtener una primera conclusión importante en relación al modelo especificado. El hecho de que el coeficiente sea superior a la unidad impide señalar que la estructura anidada sea consistente con la maximización de la utilidad aleatoria. Desde este punto de vista, no resulta posible interpretar la toma secuencial de

decisiones como el criterio utilizado por el individuo para seleccionar, entre diferentes alternativas, aquella que le proporciona un mejor resultado, en términos de utilidad.

Sin embargo, tal y como ya sabemos, es posible dar una interpretación alternativa al árbol de decisión cuando, como ocurre en este caso, el coeficiente del valor inclusivo se encuentra fuera del intervalo $[0,1]$. El uso de una estructura anidada puede responder más a una forma de solucionar la posible correlación entre los factores no observados (términos de error) para las alternativas que al deseo de tratar analíticamente el proceso de decisión seguido por el asegurado. El esquema presentado en la Figura 21 puede utilizarse para describir el proceso de decisión asumiendo, por ejemplo, la existencia de costes asociados a la elección de cada una de las alternativas. Así, la preparación de un fraude puede no sólo suponer un coste de oportunidad en tiempo (diseño del plan utilizado, ensayo de la declaración que harán el asegurado y el contrario para evitar errores que levanten sospechas,...) sino también un coste moral. El individuo que defrauda es consciente de que si su comportamiento es detectado quedará en cuestión su honestidad, siendo la ética de la persona o su temor a las penalizaciones lo que en numerosos casos frenará la actuación fraudulenta. Desde este punto de vista parece lógico pensar que cuando el asegurado esquematiza mentalmente un árbol de decisión no lo hace para maximizar la utilidad esperada sino más bien para ir eliminando alternativas de su conjunto de elección, y, por tanto, no debe sorprendernos la aparición de un coeficiente estimado para el valor inclusivo superior a la unidad.

Ahora bien, ¿es este coeficiente poblacionalmente distinto de uno?

7.4.1.2 Contrastes de especificación para el modelo logístico anidado 1A

En el Capítulo 4 veíamos como la fundamentación teórica básica del modelo logístico anidado se centraba en la aparición de un coeficiente para el valor inclusivo diferente a la unidad. De lo contrario, la Independencia de Alternativas Irrelevantes quedaba justificada y no era necesario estimar considerando etapas en el proceso de decisión (aplicación de un logit condicional). De esta forma, la respuesta a la pregunta de si el coeficiente del valor inclusivo es a nivel poblacional diferente de uno permitirá obtener conclusiones en relación a la especificación correcta del modelo.

La especificación de un modelo logit anidado puede contrastarse respecto a un logit condicional utilizando el test del ratio de verosimilitud o el estadístico de Wald²⁴. Bajo la

²⁴ También puede utilizarse el test de los multiplicadores de Lagrange.

hipótesis nula de que la especificación de un modelo lógit condicional es correcta (coeficiente del valor inclusivo igual a la unidad), ambos estadísticos se distribuyen según una Chi-cuadrado con un grado de libertad.

El valor para el test del ratio de verosimilitud (LR), calculado a partir de la expresión,

$$LR = 2 (\ln L_a - \ln L_c)$$

siendo $\ln L_a$, el logaritmo de la verosimilitud del modelo logístico anidado (-909.3) y $\ln L_c$, el logaritmo de la verosimilitud del correspondiente lógit no anidado (-915.6), es igual a 12.6. Su comparación con el valor en tablas para una χ^2 con un grado de libertad nos lleva a aceptar la hipótesis alternativa de especificación correcta de un logístico anidado (coeficiente del valor inclusivo poblacionalmente distinto de uno), al 1% de significación. Para el modelo sin ponderar, llegamos a la misma conclusión. En este caso, el test toma un valor igual a 18.2, siendo el $\ln L_a$ igual a -1226.2 y el $\ln L_c$ igual a -1235.3 .

El estadístico de Wald nos conduce, de nuevo, a aceptar la existencia de un coeficiente para el valor inclusivo diferente de uno. Bajo la hipótesis nula, el estadístico calculado mediante la expresión,

$$W = \frac{(\hat{\lambda} - 1)^2}{(e.s(\hat{\lambda}))^2},$$

siendo $\hat{\lambda}$ el coeficiente estimado del valor inclusivo, es igual a 6.5 y su comparación con el valor en tablas para una χ^2 con un grado de libertad nos lleva a aceptar la hipótesis alternativa a un 5% de significación (prácticamente, al 1%). La conclusión es la misma para el modelo sin ponderar. En este caso el estadístico es igual a 10.3 y por tanto permite aceptar la hipótesis alternativa al 1%.

Una explicación exhaustiva de los tests de especificación para el modelo lógit anidado puede encontrarse en Hausman y McFadden (1984).

7.4.1.3 Cálculo de las probabilidades predichas por el modelo

Teniendo en cuenta los resultados presentados en la Tabla 21 para los coeficientes estimados de la regresión, las probabilidades predichas de elección de cada una de las alternativas serán (omitimos el subíndice i para simplificar la expresión), para el nivel inferior ($P_{j|c}$),

$$P(b^{\circ} \text{ aseg.} | \text{fraude}) = \frac{e^{1.573+0.070 \text{ ANTC}-0.004 \text{ ANTC}2+0.867 \text{ OCURANT}}}{e^{1.573+0.689 \text{ COBERTER}+2.006 \text{ CULPAASE}+0.07 \text{ ANTC}-0.004 \text{ ANTC}2+0.867 \text{ OCURANT}}}$$

$$P(b^{\circ} \text{ ter.} | \text{fraude}) = \frac{e^{0.689 \text{ COBERTER}+2.006 \text{ CULPAASE}}}{e^{1.573+0.689 \text{ COBERTER}+2.006 \text{ CULPAASE}+0.07 \text{ ANTC}-0.004 \text{ ANTC}2+0.867 \text{ OCURANT}}}$$

$$P(\text{nofraude} | \text{nofraude}) = 1$$

y para el nivel superior (P_c),

$$P(\text{fraude}) = \frac{e^{-5.990+0.155 \text{ HISTC}+1.414 \text{ APELAPCC}-1.624 \text{ SAUA}+0.509 \text{ RELSTRO}+1.835 I_c}}{e^{I_t}}$$

$$P(\text{nofraude}) = 1 - P(\text{fraude}) = \frac{1}{e^{I_t}}$$

siendo,

$$I_c = \ln \left[e^{1.573+0.689 \text{ COBERTER}+2.006 \text{ CULPAASE}+0.070 \text{ ANTC}-0.004 \text{ ANTC}2+0.867 \text{ OCURANT}} \right]$$

$$I_t = \ln \left[1 + e^{-5.990+0.155 \text{ HISTC}+1.414 \text{ APELAPCC}-1.624 \text{ SAUA}+0.509 \text{ RELSTRO}+1.835 I_c} \right].$$

La probabilidad incondicional estimada (dados los valores para las variables explicativas) de la elección hecha por el individuo es,

$$P_{(ej)} = P_{(j|c)} P_c$$

Para el modelo sin ponderar el cómputo de la probabilidades es idéntico, lógicamente, utilizando los coeficientes estimados para el mismo.

7.4.1.4 Análisis de la calidad del ajuste

Una vez analizada la significación individual y global del modelo, contrastada la especificación correcta anidada y presentado el cálculo de las probabilidades predichas, únicamente nos queda por estudiar la calidad del ajuste del modelo, en el sentido de determinar en qué medida clasifica correctamente los siniestros según el tipo de fraude existente.

El modelo clasifica al siniestro en aquella categoría final cuya probabilidad estimada es mayor. Desde este punto de vista, la comparación entre la clasificación real y la predicha para cada observación es la presentada en la Tabla 23.

Tabla 23. Frecuencias de clasificación. (utilizando el criterio $c=0.15$)

	Elección predicha			Total
	Legítimo	Beneficio propio	Beneficio tercero	
Elección Observada				
Legítimo	796	59	143	998
Beneficio propio	172	53	74	299
Beneficio tercero	62	16	236	314
Total	1030	128	453	1611

A partir de la Tabla 23, podemos observar como el porcentaje de clasificación correcto es de un 67.3% (un total de 1085 casos son clasificados correctamente por el modelo). Sin embargo, y pesar de que para los siniestros sin fraude detectado y con fraude a favor del tercero el porcentaje de aciertos es elevado (prácticamente de un 79.7% y de un 75.1%, respectivamente), para los siniestros con fraude a favor del asegurado no podemos decir lo mismo. Como puede observarse, para estos últimos, el porcentaje de aciertos es sólo de 17.7%. Además y, en relación a los mismos, se observa un hecho preocupante: el elevado número de casos clasificados como no fraudulentos (un 57.5%).

Por lo tanto, en base a los resultados obtenidos, la compañía dejaría de detectar un elevado número de casos fraudulentos. Aunque es posible que este hecho venga provocado por la utilización de una muestra de siniestros de ambos tipos de fraude no representativa (sólo disponemos de 299 casos para el primer tipo de fraude y de 314 para el segundo frente a 998 casos sin fraude detectado), la clasificación para los casos fraudulentos puede mejorarse al seleccionar un criterio probabilístico que cambie las regiones de clasificación, tal y como puede observarse en la Tabla 24.

Tabla 24. Frecuencias de clasificación (utilizando el criterio $c=0.10$)

	Elección predicha			Total
	Legítimo	Beneficio propio	Beneficio tercero	
Elección Observada				
Legítimo	495	347	156	998
Beneficio propio	68	156	75	299
Beneficio Tercero	24	48	242	314
Total	587	551	473	1611

En este caso, los resultados para los siniestros sin fraude empeora, pero quizá la compañía esté dispuesta a asumir los costes de investigar como fraudulentos siniestros que no lo son, ante las ventajas obtenidas en relación a detección de comportamientos deshonestos. El ratio de clasificación correcta es ahora tan sólo del 55.4% pero, para los siniestros fraudulentos de uno y otro tipo, el porcentaje de aciertos es del 52% y del 77%, respectivamente.

Si se dispusiera de información más concreta sobre los costes de investigación podría incluso fijarse un criterio de clasificación mejor. Aquí se supone que la optimalidad de un sistema de clasificación viene dada por su capacidad de detectar los fraudes.

Las conclusiones extraídas, por tanto, en relación a la calidad del ajuste nos llevan a señalar que la capacidad predictiva del modelo es bastante buena. Los resultados evidencian que el modelo lógit anidado proporciona una técnica de predicción aceptable.

Su capacidad como herramienta explicativa ha quedado también justificada a lo largo de este capítulo. Así, es de destacar el hecho de que gracias a su aplicación hemos sido capaces de proponer un modelo para el proceso de decisión que lleva al asegurado a cometer o no fraude. Además hemos determinado qué variables tienen influencia en cada elección, siendo destacable la intervención de las mismas en las diferentes etapas de decisión. De esta forma, la compañía puede saber *a priori* sobre qué variables tiene que prestar más atención, no sólo para detectar fraude, genéricamente hablando, sino también para detectar aquellos comportamientos fraudulentos que con mayor frecuencia aparecen en la cartera de asegurados. Posiblemente la obtención de muestras más representativas para los tipos de fraude analizados pudiera ayudar a que la modelización logística anidada ganase en capacidad predictiva.

La comparación con los resultados que obteníamos para este mismo árbol de decisión tras la aplicación de un lógit multinomial nos impide establecer conclusiones definitivas sobre cuál de las dos modelizaciones resulta más adecuada. Las razones se pueden sintetizar en dos.

Mientras que el lógit multinomial proporciona una mejor herramienta de predicción (los porcentajes de clasificación correcta son más elevados para los mismos criterios probabilísticos) y resulta más fácil de implementar, el lógit anidado permite señalar la dirección que ha de seguir la entidad a la hora de investigar determinados tipos de fraude. Para este último, la implementación es más complicada pero es posible demostrar el poder explicativo de determinadas variables que no resultan estadísticamente significativas en el modelo multinomial (véase la variable que recoge la antigüedad del vehículo del asegurado en el momento del siniestro o la que indica la ocurrencia del siniestro entre la fecha de efecto y la de emisión de la póliza).

7.4.2 Estimación logística anidada de la probabilidad de existencia de fraude para conseguir una indemnización o para incrementar la indemnización derivada de un siniestro (MODELO 2A)

Pasamos, a continuación a analizar los resultados obtenidos al modelizar el segundo árbol de decisión representado gráficamente en la Figura 22.

7.4.2.1 Análisis de la significación individual y global del modelo

La estimación de un modelo como el presentado en la Figura 22, en el que de nuevo tenemos en cuenta la existencia de dos niveles de decisión ha supuesto desarrollar un proceso análogo al presentado para el Modelo 1A y que hemos comentado en el subapartado anterior.

De esta forma, y siguiendo el mismo criterio, hemos cuantificado la probabilidad de que el individuo elija la alternativa de no defraudar, de defraudar para conseguir una indemnización o de defraudar para incrementar la indemnización que le corresponde tras la ocurrencia legal del siniestro (variable dependiente del modelo) mediante la aplicación del método de Máxima Verosimilitud Completa.

La muestra utilizada consta de un total de 1676 siniestros (ver apartado 6.2). De esta forma, la estimación se realizará, al tener en cuenta tres alternativas de elección para cada uno de los asegurados, con un total de 5028 observaciones.

Al igual que en el modelo 1A hemos partido de considerar en la estimación la existencia de estratificación endógena en la base de datos. Así, hemos introducido las ponderaciones correspondientes para cada uno de los siniestros observados dependiendo de los valores de la variable dependiente. Tal y como veíamos en el capítulo anterior (apartado 6.2), las

observaciones de no fraude han sido ponderadas por un factor de 1.3, mientras que las asociadas al primer y segundo tipo de fraude considerado, por los factores 0.6 y 0.5, respectivamente. No obstante, y como explicábamos para el modelo 1A, de nuevo la imposibilidad de calcular directamente la matriz de varianzas y covarianzas de los estimadores teniendo en cuenta la existencia de muestreo basado en la elección²⁵ nos ha llevado a presentar, adicionalmente, los resultados obtenidos al estimar el modelo sin ponderar.

Las variables explicativas introducidas en la especificación del modelo ya han sido comentadas en el subapartado 7.3.2. Para todas ellas, se dispone de información en un breve periodo de tiempo tras la ocurrencia del siniestro, con lo que la rapidez en la detección de fraudes se puede ver considerablemente incrementada.

La especificación de las funciones de utilidad para cada una de las alternativas, considerando la existencia de dos niveles de decisión, ha sido la siguiente:

$$U(A1, A2, A3) = \beta_1 AELEC + \beta_2 CULPACON + \beta_3 HISTC1 + \beta_4 OCURANT1 + e_j$$

$$U(R1, R2) = \alpha_1 (1-SELEC) + \alpha_2 SAUA + \alpha_3 APELAPCC + \alpha_4 RELSTRO + e_j$$

siendo,

A1= alternativa 1= no fraude;

A2= alternativa 2= fraude *para conseguir una indemnización*;

A3= alternativa 3= fraude *para incrementar la indemnización*;

R1=rama1=no fraude;

R2= rama 2= fraude,

y,

$e_j, e_{j'}$, con $j=1,2,3$ (alternativas finales) y $j'=1,2$ (alternativas intermedias), los términos de error aleatorio del modelo.

Los coeficientes estimados para el modelo ponderado y sin ponderar han sido los presentados en las Tablas 25 y 26, respectivamente.

²⁵ La corrección se ha realizado, como veíamos para el modelo 1A, utilizando los productos externos de las primeras derivadas de la log-verosimilitud del modelo (estimador habitual cuando el algoritmo de iteración es, como en nuestro caso, el D.F.P.).

Tabla 25. Estimación del modelo lógit anidado con ponderaciones
[MODELO 2A]

variable	coeficiente	t-test	P-valor
<i>Nivel inferior:</i>			
CONSTANTE	-0.654	-4.239	0.000
CULPACON	-1.392	-6.411	0.000
HISTC1	0.069	2.726	0.006
OCURANT1	1.499	4.531	0.000
<i>Nivel superior:</i>			
CONSTANTE	-1.458	-5.015	0.000
APELAPCC	1.373	5.020	0.000
SAUA	-1.655	-6.056	0.000
RELSTRO	0.405	3.037	0.002
<i>Valores Inclusivos:</i>			
NO FRAUDE	1.000	----	----
FRAUDE	2.182	5.457	0.000
número de observaciones ^a	5028	Chi-cuadrado	1298.318
ln función verosimilitud	-982.509	grados de libertad	9
ln verosimilitud restringida	-1631.668	nivel signif.	0.000

^a Cada individuo es replicado una vez por cada elección.

Tabla. 26. Estimación del modelo lógit anidado sin ponderar
[MODELO 2A]

variable	coeficiente	t-test	P-valor
<i>Nivel inferior:</i>			
CONSTANTE	-0.869	-7.354	0.000
CULPACON	-1.398	-8.650	0.000
HISTC1	0.072	3.355	0.001
OCURANTI	1.583	5.376	0.000
<i>Nivel superior:</i>			
CONSTANTE	-0.363	-1.862	0.062
APELAPCC	1.323	5.047	0.000
SAUA	-1.557	-7.340	0.000
RELSTRO	0.408	3.554	0.000
<i>Valores Inclusivos:</i>			
NO FRAUDE	1.000	----	----
FRAUDE	2.088	7.223	0.000
Número de observaciones ^a	5028	Chi-cuadrado	531.005
ln función verosimilitud	-1366.166	grados de libertad	9
ln verosimilitud restringida	-1631.668	nivel signif.	0.000

^a Cada individuo es replicado una vez por cada elección.

Antes de pasar a analizar el comportamiento observado para cada una de las variables introducidas en la especificación cabe señalar que los dos modelos, ponderado y sin ponderar, son globalmente significativos. Así lo demuestran los valores obtenidos para el estadístico Chi-cuadrado, iguales a 1298.3 en el modelo ponderado y a 531.0 en el modelo sin ponderar. Su comparación en tablas con una χ^2 con 9 grados de libertad nos lleva a aceptar la hipótesis alternativa de significación global al 1%, en ambos casos.

En relación a los coeficientes estimados, cabe señalar que todos ellos gozan de significación estadística (salvo para el coeficiente que acompaña a la constante de elección (1-SELEC) que es estadísticamente significativo al 5%, el resto los son al 1%) y además, salvo en un caso, presentan los signos esperados. Para el nivel inferior del árbol, los atributos a los que acompañan han sido definidos en función de la constante de elección de una determinada alternativa y, por tanto, deberán ser interpretados en referencia a ella y en comparación con la otra alternativa de fraude. Asimismo, para el nivel superior, las variables utilizadas han sido definidas mediante la interacción con la constante (1-SELEC) lo que supondrá interpretar los coeficientes estimados en términos de una mayor (signo positivo) o menor (signo negativo) probabilidad de aparición de fraude.

Observando como los coeficientes estimados muestran un comportamiento paralelo en el modelo que posee ponderaciones para la variable dependiente y en el modelo sin ponderar, pasamos a interpretar los resultados obtenidos comenzando por el nivel inferior del árbol de decisión.

Los atributos utilizados en la especificación de las funciones de utilidad para las alternativas finales están directamente relacionados con el siniestro (culpa del accidente y fecha de ocurrencia) y con circunstancias asociadas al asegurado en el momento de ocurrencia del mismo (historial de siniestros en la compañía).

Para la constante específica de elección, la obtención de un coeficiente estimado con signo negativo permite señalar que, si el resto de variables consideradas son nulas, se producirá una disminución en la probabilidad de que el individuo elija cometer fraude para conseguir una indemnización en relación a que lo haga para incrementarla.

El coeficiente para la variable *CULPACON* indica que cuando el contrario acepta la culpa del siniestro disminuye la probabilidad de que el fraude cometido sea para incrementar una indemnización frente a conseguirla. De esta forma, el signo obtenido muestra un efecto contrario al que habíamos supuesto cuando definíamos el atributo, suponiendo entonces que la

aceptación de culpa por el contrario afectaría directamente a la aparición del segundo tipo de fraude considerado. Como consecuencia del resultado obtenido cabe interpretar que en aquellos casos en los que existe fraude para conseguir una indemnización es posible que exista un acuerdo entre la parte asegurada y la parte contraria, siendo esta última la que se declare responsable del siniestro.

La ocurrencia del siniestro entre la fecha de efecto y la fecha de emisión de la póliza está directamente relacionada con la probabilidad de existencia de fraude para conseguir una indemnización. El signo positivo para la variable *OCURANTI* así lo indica, de forma que, como cabía esperar, el retroceso en la fecha de efecto del contrato de seguro se realizará normalmente para cubrir un siniestro que no lo estaba en el momento de su ocurrencia.

El número de siniestros que el asegurado ha declarado en la compañía muestra una relación positiva con la aparición de fraude destinado a incrementar la indemnización derivada del accidente. La obtención de un coeficiente estimado con signo positivo para la variable *HISTCI* confirma el supuesto que introducíamos en su definición, de forma que cuando el asegurado posee un mayor conocimiento de la operativa seguida por la compañía en la tramitación de los siniestros aumenta la probabilidad de que, al sufrir un nuevo siniestro, intente defraudarla.

¿Qué resultados obtenemos para las variables introducidas en la especificación de las funciones de utilidad para las alternativas intermedias?

En este caso, las variables consideradas están asociadas, bien a la relación que existe entre el asegurado y el contrario (coincidencia de apellidos), bien a determinadas circunstancias relacionadas con el propio siniestro (existencia, en la declaración del asegurado, de los relatos sospechosos e intervención de autoridad policial).

La obtención de un coeficiente negativo para la constante de elección específica de fraude permite interpretar una disminución en la probabilidad de que el asegurado elija dicha alternativa en relación a no fraude, cuando el resto de variables son nulas.

La interpretación para el resto de coeficientes, dado que aparecen asociados a variables explicativas de fraude, genéricamente considerado, será similar a la que presentábamos para el Modelo 1A. Así, la existencia de coeficientes positivos para las variables *APELAPCC* y *RELSTRO* permite interpretar que cuando existe coincidencia de apellidos o aparecen en la declaración términos que indican una descripción sospechosa aumenta la probabilidad de

existencia de fraude. De la misma forma pero, en sentido contrario, la aparición de un coeficiente negativo para la variable *SAUA* indica una disminución en la probabilidad de aparición de fraude cuando la policía ha intervenido en el siniestro. De nuevo, los signos observados confirman el comportamiento que para las variables *DESTAF*, *DRELATO* y *SAUTDAD*, ya avanzábamos en el Capítulo 5.

¿Qué ocurre con los coeficientes estimados para los valores inclusivos?

El haber fijado en uno el coeficiente del valor inclusivo asociado a la rama de no fraude (existe una única alternativa final dependiente de la misma), supone realizar la interpretación únicamente para el coeficiente del valor asociado a la rama de fraude. Como se observa en la tabla de resultados, dicho coeficiente, positivo y estadísticamente significativo, es superior a la unidad. Las consecuencias son, por tanto, las mismas que veíamos para el modelo 1A, y por tanto no podemos interpretar la estructura anidada desde el punto de vista de la maximización de la utilidad aleatoria. La existencia de una toma secuencial de decisiones vendrá provocada más por un deseo del asegurado de eliminar alternativas que de maximizar la utilidad que espera de las mismas.

En muchos casos la alternativa final seleccionada estará condicionada por las circunstancias en que ha ocurrido el accidente. Por ejemplo, al detectarse que el siniestro ha ocurrido entre la fecha de efecto y la fecha de emisión de la póliza la secuencia más lógica de toma de decisiones es la siguiente: el asegurado que no tiene cobertura se plantea la posibilidad de conseguirla (decide defraudar²⁶) y, posteriormente, elimina la alternativa que no le interesa, en este caso, cometer fraude para incrementar una indemnización. Desde este punto de vista, el individuo no pretende maximizar la utilidad esperada eligiendo indistintamente entre alternativas; simplemente, selecciona aquella que se adecua a sus necesidades.

Pasamos a continuación a analizar si el coeficiente estimado para el valor inclusivo es poblacionalmente distinto de uno.

7.4.2.2 Contrastes de especificación para el modelo anidado 2A

El contraste de la hipótesis nula, $H_0: \lambda = 1$, relacionada con una especificación de modelo no anidada, ha sido realizada de nuevo, mediante la aplicación del test del ratio de verosimilitud (LR) y mediante el estadístico de Wald.

²⁶ Intervendrá personal de la compañía.

En el modelo ponderado, el LR es igual a 26.9, siendo el logaritmo neperiano de la verosimilitud del modelo anidado igual a -982.5 y el correspondiente a la verosimilitud del modelo logístico sin anidar, -996.0. El contraste en tablas con una χ^2 con 1 grado de libertad supone aceptar la hipótesis alternativa al 1% de significación y por tanto, el coeficiente para el valor inclusivo es poblacionalmente distinto de uno, es decir, el modelo anidado está correctamente especificado.

Para el modelo sin ponderar, se obtiene la misma conclusión. En este caso, el logaritmo neperiano de la verosimilitud del modelo lógit anidado es -1366.2 mientras que el del lógit sin anidar es -1381.6. El LR es, por tanto, igual a 30.9 y su contraste en tablas con una χ^2 con 1 grado de libertad nos lleva, de nuevo, a aceptar la hipótesis alternativa al 1%.

El cálculo del estadístico de Wald confirma el resultado anterior. Obteniéndose un valor para el mismo igual a 8.7, en el modelo ponderado y a 14.2, en el modelo sin ponderar, el contraste con el valor en tablas para una χ^2 con 1 grado de libertad supone, en ambos casos, aceptar al 1% la hipótesis alternativa de especificación correcta anidada.

7.4.2.3 Cálculo de las probabilidades predichas por el modelo

Considerando los resultados que aparecen en la Tabla 25 para los coeficientes estimados de la regresión, las probabilidades predichas de elección de cada una de las alternativas dada las características del individuo para el nivel inferior ($P_{j|c}$), son (de nuevo omitimos el subíndice i para simplificar la notación),

$$P(\text{conse. indemn.} | \text{fraude}) = \frac{e^{-0.654+1.499\text{OCURANT1}}}{e^{-0.654-1.392\text{CULPACON}+0.069\text{HISTC1}+1.499\text{OCURANT1}}$$

$$P(\text{increm. indemn.} | \text{fraude}) = \frac{e^{-1.392\text{CULPACON}+0.069\text{HISTC1}}}{e^{-0.654-1.392\text{CULPACON}+0.069\text{HISTC1}+1.499\text{OCURANT1}}$$

$$P(\text{nofraude} | \text{nofraude}) = 1$$

y para el nivel superior (P_c),

$$P(\text{fraude}) = \frac{e^{-1.458+1.373\text{APELAPCC}-1.655\text{SAUA}+0.405\text{RELSTRO}+2.182I_c}}{e^{I_t}}$$

$$P(\text{nofraude}) = 1 - P(\text{fraude}) = \frac{1}{e^{I_t}}$$

siendo,

$$I_c = \ln \left[e^{-0.654 - 1.392CULPACON + 0.069HISTC1 + 1.499OCURANT1} \right]$$

$$I_t = \ln \left[1 + e^{-1.458 + 1.373APELAPCC - 1.655SAUA + 0.405RELSTRO + 2.182 * I_c} \right].$$

La probabilidad incondicional estimada (dados los valores para las variables explicativas) de la elección hecha por el individuo como sabemos será,

$$P_{(cj)} = P_{(j|c)} P_c.$$

Para el modelo sin ponderar el cómputo de la probabilidades sería idéntico, considerando lógicamente los coeficientes estimados para el mismo.

7.4.2.4 Análisis de la calidad del ajuste

La comparación entre la clasificación real de siniestros y la predicha por el modelo (sólo para el modelo con ponderaciones) es la siguiente,

Tabla 27. Frecuencias de clasificación (utilizando el criterio $c=0.15$)

	Elección predicha			Total
	Legítimo	Conseguir una indemnización	Incrementar una indemnización	
Elección Observada				
Legítimo	829	22	147	998
Conseguir una indemnización	174	42	86	302
Incrementar una indemnización	117	14	245	376
Total	1120	78	478	1676

El porcentaje de casos correctamente clasificados en la categoría correspondiente es del 66.6% (1116 casos). Para los siniestros no fraudulentos y con fraude dirigido a incrementar la indemnización, el porcentaje de aciertos es elevado, del 83% y del 65%, respectivamente. Para los siniestros con fraude dirigido a conseguir una indemnización dicho porcentaje es

notablemente inferior, del 14%. El modelo posee, por lo tanto, una pobre capacidad predictiva en relación al primer tipo de fraude. No obstante, de nuevo, resulta preocupante el elevado número de siniestros que, conteniendo fraude, son clasificados por el modelo como no fraudulentos (un 57.6% de los siniestros con fraude dirigido a conseguir una indemnización y, un 31% del siniestros con fraude destinado a incrementarla).

Las conclusiones son, por tanto, análogas a las que veíamos al final del subapartado 7.4.1.4, en relación a la calidad del ajuste para el Modelo 1A. La selección de un punto de corte fijado en un criterio de asignación que amplía la zona de clasificación de los distintos tipos de fraude ayuda a mejorar los resultados, en cuanto a fraude se refiere, tal y como puede observarse en la Tabla 28.

Tabla 28. Frecuencias de clasificación (utilizando el criterio $c=0.10$)

	Elección predicha			Total
	Legítimo	Conseguir una indemnización	Incrementar una indemnización	
Elección Observada				
Legítimo	455	393	150	998
Conseguir una indemnización	59	155	88	302
Incrementar una indemnización	50	80	246	376
Total	564	628	484	1676

El porcentaje de casos correctamente clasificados es ahora del 51%, observándose, no obstante, una notable mejoría para el ratio de clasificación de los siniestros con fraude dirigido a conseguir una indemnización (el porcentaje de aciertos para el segundo tipo de fraude considerado, prácticamente no varía, en relación al presentado en la tabla 27). Los resultados ponen de manifiesto una reducción en el porcentaje de siniestros con fraude, clasificados por el modelo como no fraudulentos (pasa a ser del 19.5% para los casos de fraude dirigido a conseguir una indemnización y del 13% para aquéllos con fraude destinado a incrementarla). Sin embargo, la capacidad predictiva del modelo en relación a los siniestros no fraudulentos, empeora de forma considerable (sólo un 45.6% de los casos son correctamente clasificados). De nuevo, será la compañía la que determinará la conveniencia de asumir unos mayores costes por la investigación de siniestros no fraudulentos, en aras a detectar un mayor número de siniestros fraudulentos de uno y otro tipo.

Los ratios de clasificación obtenidos permiten señalar que modelo anidado 2A posee una calidad predictiva muy moderada, aunque quizá la disponibilidad de una mayor número de casos para cada uno de los tipos de fraude considerados permitiría mejorar notablemente su capacidad de predicción. No obstante, el hecho de que gracias a su aplicación seamos capaces de diferenciar la influencia que determinadas variables tienen en la elección concreta de uno u otro tipo de fraude, además de considerar el proceso de decisión seguido por el asegurado, hace que la modelización presentada sea una herramienta muy importante a tener en cuenta en el estudio del fraude y la investigación de siniestros, debido a la explicación que aporta respecto a la influencia de las variables.

Los resultados obtenidos al modelizar este segundo árbol de decisión y, al igual que comentábamos para el primero, no nos permiten decidir cuál de las dos modelizaciones, la multinomial o la anidada, resulta más adecuada. Las razones son las mismas que aludíamos al final del apartado 7.4.1.4. Así mientras que del modelo multinomial se deriva una mejor calidad del ajuste (los porcentajes de clasificación correcta son más elevados), del anidado se obtiene una interesante herramienta de investigación para la compañía, que sabrá hacia donde debe enfocar sus esfuerzos de cara a detectar determinados tipos de fraude.

Finalmente, el análisis de costes refleja, utilizando el mismo escenario que presentábamos en el Capítulo 6, un ahorro esperado de $0.7s_0$ y de $0.3s_0$ para el Modelo 1A y 2A, respectivamente. Como puede observarse, estos valores son inferiores a los que obteníamos tras la aplicación del Modelo 1 y del 2, hecho que viene provocado por la inferior calidad predictiva de los modelos anidados (la comparación se ha realizado considerando en todos los casos el mismo criterio probabilístico fijado en $c=0.15$).

8. CONCLUSIONES

La existencia de fraude en el seguro de automóviles provoca un incremento en el número y coste medio de los siniestros que ha despertado el interés de las entidades. La solución tradicional utilizada por las compañías ante la presencia de costes superiores a los esperados ha sido el incremento de las primas aplicadas. Sin embargo, este mecanismo no ha podido evitar los resultados técnicos negativos que han caracterizado al ramo de automóviles en los últimos años.

El objetivo perseguido en la presente Tesis Doctoral es modelizar el comportamiento deshonesto en las declaraciones de siniestros del seguro del automóvil. Para ello, se ha utilizado información de una muestra de reclamaciones del mercado asegurador español y se ha aplicado la metodología econométrica para especificar, estimar y validar los modelos.

La mayor preparación del personal asegurador, la cooperación entre compañías y la colaboración con las fuerzas de seguridad, entre otras posibilidades, son esenciales. Sin embargo, no son totalmente efectivas si no van acompañadas de mecanismos de control suficientemente validados desde un punto de vista estadístico, hecho que ha dado lugar a una involucración de la comunidad científica en el tratamiento del fraude.

Los resultados obtenidos constituyen un modelo explicativo a la vez que un mecanismo de detección. En el tema estudiado a lo largo de este trabajo confluyen dos intereses: 1) realizar un análisis microeconómico del fraude y 2) elaborar un procedimiento que permita agilizar la tramitación y la investigación de las reclamaciones.

Este segundo aspecto permite desarrollar una herramienta de decisión que puede ayudar a las entidades aseguradoras a controlar el fraude. Más aún, la Tesis profundiza en la identificación de los factores que influyen en los diversos comportamientos deshonestos. Por lo tanto, los modelos no sólo ayudan a decidir si se debe investigar o no un siniestro sospechoso, sino a concretar de qué forma debe investigarse. Los modelos estudiados permiten concluir que

algunas características se asocian a un tipo de fraude y no a otro y, de este modo, aportan evidencias de cómo encaminar el seguimiento de la reclamación.

A pesar de que en otros países (fundamentalmente, EE.UU, Canadá y Francia) es posible encontrar grupos de trabajo o instituciones directa o indirectamente asociadas al estudio del fraude, hasta el momento no podíamos decir lo mismo a nivel nacional. La investigación presentada es, por tanto, totalmente novedosa en el marco del seguro automovilístico español, aunque las diferencias metodológicas en relación al tratamiento econométrico realizado por otros autores permiten ampliar dicho calificativo a nivel internacional.

Las principales conclusiones obtenidas a lo largo del trabajo se detallan a continuación:

1.- La importancia concedida en la última década a la detección de comportamientos fraudulentos en el mercado asegurador en general queda patente a nivel nacional e internacional. La prevención del fraude en la selección de riesgos y en la liquidación de siniestros es uno de los objetivos reflejados en la nueva Ley de Ordenación y Supervisión de los Seguros Privados en España, en vigor desde 1995. Además, como resultado de la integración comunitaria, todos los países miembros están homogeneizando sus organismos sectoriales de lucha contra el fraude. El seguro del automóvil es el más afectado y, dado que suele llevar asociados resultados técnicos negativos, es el que ha despertado un mayor interés por el tratamiento del problema, aunque se acepta la existencia de fraude en la práctica totalidad de los ramos aseguradores. A pesar de todo ello, no existen en Europa trabajos científicos que traten el fenómeno del fraude en el seguro del automóvil desde un punto de vista cuantitativo. La única excepción son las investigaciones de Pierre Picard, que mediante un enfoque de teoría económica demuestran la necesidad de crear organismos de control de fraude. Otras aportaciones no van más allá de la mera sugerencia de porcentajes de fraude y de una constante, pero pasiva, preocupación por el problema.

2.- Las estadísticas existentes sobre el fraude en el seguro automovilístico español muestran divergencias notables entre el porcentaje de casos en los que se sospecha una posible actuación deshonestas y el porcentaje de casos en los que realmente llega a confirmarse dicho comportamiento. Mientras que para los primeros, las cifras revelan valores superiores en algunos casos al 20% de los siniestros declarados, para los segundos, la cifra es notablemente inferior y tiende a situarse entre el 1 y el 10% (C.E.S., 1992). Estos resultados son completamente concordantes con los observados por investigadores de EE.UU. y Canadá. Así, cualquier sistema dirigido a identificar las actuaciones deshonestas sospechosas que realmente

responden a comportamientos fraudulentos será de gran valía para las compañías y la sociedad en general.

3.- La existencia de tipos de comportamientos muy variados dificulta la obtención de una definición genérica de fraude. Resulta patente que, en cualquiera de los casos, la actuación deshonesta por parte del asegurado supone la alteración del principio de buena fe, que sustenta el contrato de seguro. Así pues, la consideración de las diferentes posibilidades que el individuo tiene a su alcance a la hora de actuar fraudulentamente puede ayudar de modo notable a la compañía para generar un mecanismo de control.

La tipología existente de fraude en el marco del seguro automovilístico español se concentra fundamentalmente en ocho categorías. Sin embargo, en el presente trabajo se ha adoptado una categorización más genérica en base a la agrupación de algunas de las clases anteriores. El objetivo es ceñirse a las características específicas del funcionamiento del sistema de seguros de automóviles en España. La clasificación se ha realizado en un doble sentido. Por un lado, se ha considerado el fraude realizado por el asegurado en beneficio propio, *versus* el perpetrado en beneficio de un tercero; por otro, el fraude destinado a conseguir una indemnización que no corresponde por contrato, *versus* el dirigido a incrementar la indemnización derivada de la ocurrencia de un siniestro sujeto a cobertura.

La determinación de variables significativas en la explicación de las probabilidades de aparición de cada una de las categorías puede ayudar a que la compañía optimice su política de control de fraude. La sospecha sobre el tipo de fraude existente inducirá a analizar prioritariamente el comportamiento de aquellas variables para las que se ha demostrado poder explicativo sobre el mismo.

4.- La dificultad para obtener muestras con información completa sobre las variables que pueden ser utilizadas en la modelización del fraude es una de las principales razones que justifican la escasez de tratamientos metodológicos exhaustivos. Weisberg y Derrig (1993) utilizan una muestra no aleatoria formada por 127 siniestros (aproximadamente en un 60% de los mismos se sospecha la existencia de fraude). Belhadji y Dionne (1997) usan una muestra aleatoria formada por 2.068 siniestros (un 93.77% de los casos no son fraudulentos; en un 5.46% existe sospecha de fraude y tan sólo en un 0.87% se detecta existencia de fraude).

En nuestro caso, la muestra utilizada contiene información para un total de 1995 expedientes de siniestros, de los cuáles el 50% (997 casos) contienen fraude detectado. El 50% restante corresponde a siniestros clasificados por la compañía como no fraudulentos. Para estos

últimos resulta más correcto hablar en términos de “ausencia de fraude detectado”, dado que la entidad los ha clasificado como legítimos, pero es posible que, aunque haya realizado alguna investigación, no haya logrado verificarlo (una de las líneas de mejora en relación al trabajo presentado que propondremos al final, se centra precisamente en este aspecto: la consideración de errores de medida en la variable dependiente). La existencia en el diseño de la muestra de estratificación endógena (con el objetivo de obtener una muestra representativa para los siniestros fraudulentos, la selección de expedientes no se ha realizado de forma estrictamente aleatoria) ha estado presente en la modelización mediante la incorporación de ponderaciones en los procesos de estimación.

Los datos, referidos a un amplio conjunto de variables, han sido depurados con el objetivo de recoger información precisa para un conjunto de indicadores, presentados en la literatura como señales de fraude. Las variables literales (por ejemplo, la que recoge la descripción del relato del siniestro) han sido objeto de tratamiento textual. De esta forma se ha conseguido averiguar, entre otras cosas, cuáles son las circunstancias más propicias a la ocurrencia de siniestros fraudulentos. Mediante un identificador de coincidencia de apellidos entre asegurado y contrario se ha pretendido detectar relaciones familiares entre las partes que intervienen en el accidente.

El estudio de las fechas relacionadas con diferentes aspectos del siniestro y de la póliza ha permitido realizar tratamientos exhaustivos asociados, entre otras cosas, a la determinación del plazo que transcurre entre la ocurrencia del siniestro y la comunicación del mismo a la entidad, o a la detección de casos en los que la ocurrencia del siniestro se produce entre la fecha de efecto y la fecha de emisión de la póliza.

Para numerosos atributos, como los relativos a la intervención policial en el siniestro o a la participación de testigos, se han utilizado variables *dummies* que han permitido realizar una recodificación en términos de la presencia o ausencia de una determinada situación. Otras variables, como la edad del individuo y su antigüedad de carnet en el momento del accidente, o el historial de siniestros que posee en la entidad, han sido directamente introducidas en las ecuaciones de regresión.

5.- Aunque las razones que pueden llevar al asegurado a cometer fraude pueden ser variadas (deseo de recuperar la prima pagada, intención de beneficiar a un amigo, conseguir la cobertura para situaciones excluidas de la póliza,...), su actuación está fundamentada principalmente en dos hechos concretos: la mayor o menor probabilidad con la que cree que

será detectado y las circunstancias en las que se produce el siniestro (que incluyen el caso de que el accidente sea provocado por el asegurado).

La utilidad final esperada por el individuo, definida por Picard (1996) como el resultado de adicionar la utilidad esperada al actuar de forma honesta y la esperada al actuar fraudulentamente (ambas utilidades ponderadas por la probabilidad de elegir una u otra alternativa, respectivamente), debería recoger la existencia de diferentes clases de fraude. Para cada una de ellas el asegurado tendrá en cuenta dos situaciones posibles. Si su comportamiento es detectado por la compañía sufrirá una penalización (en el caso español normalmente será la expulsión de la compañía o el pago únicamente de la indemnización que corresponde por contrato). Si no lo es, él y/o el contrario percibirán una cantidad monetaria en concepto de reparación de daños, no justificada.

El proceso necesario para estimar la utilidad esperada por el asegurado a la hora de decidir si comportarse honestamente o cometer un determinado tipo de fraude supondrá la cuantificación de dos conceptos básicos. Por un lado, se determinará la probabilidad de que el asegurado elija cada una de las alternativas; por otro, se cuantificará la probabilidad de que, tras la aplicación del modelo, se detecte la existencia de actuación ilícita. El alcance de ambos objetivos ha de estar presente en la selección de la técnica econométrica utilizada en la modelización.

6.- El trabajo desarrollado está dirigido a la aplicación de métodos cuantitativos en el diseño de un modelo de control de fraude, teniendo en cuenta la tipología existente para el mismo. El objetivo principal se ha centrado en considerar, dentro de la modelización, el proceso de decisión seguido por el asegurado a la hora de elegir una determinada alternativa. Este proceso puede ser unietápico o polietápico. En el segundo caso, el individuo decide, en primer lugar, cometer fraude o bien no hacerlo y, si opta por la primera alternativa, selecciona la forma más conveniente de realizarlo. La representación gráfica de árboles de decisión ayuda a comprender el proceso seguido, teniendo en cuenta, por tanto, una elección secuencial o por etapas. La aplicación de técnicas de regresión logística (modelos lógit multinomiales y anidados) ha permitido cuantificar la probabilidad de elegir cada una de las alternativas, determinando, asimismo, las variables estadísticamente significativas en la explicación de cada comportamiento.

Los resultados obtenidos permiten interpretar el árbol de decisión atendiendo a un doble criterio: el asegurado realiza una elección por etapas bajo el deseo de maximizar la utilidad esperada de su comportamiento (conclusión obtenida tras la aplicación de los modelos lógit

multinomiales), o bien lo hace para estructurar mentalmente las diferentes opciones que posee con el objetivo de minimizar los costes -en tiempo, en preparación del siniestro,...- asociados a la elección (conclusión derivada tras aplicar modelos lógit anidados).

No podemos decidir si es mejor utilizar un modelo multinomial o un modelo anidado. Ambos proporcionan conclusiones complementarias. El primero constituye una herramienta de clasificación fácil de implementar pero, a diferencia del anidado, no orienta suficientemente sobre la forma de conducir la investigación. Sirve para identificar las variables relevantes o aquellas características en que deberían fijarse los tramitadores para luego enviar el expediente a una unidad especial de investigación. Creemos que algunas variables importantes no aparecen como significativas en este modelo porque no son capaces de discriminar entre las categorías de la variable dependiente (incluyen las reclamaciones honestas y los distintos tipos de fraude). En cambio, el modelo anidado, al tener en cuenta probabilidades condicionadas, puede tener una utilidad muy diferente. En este caso, el modelo discrimina entre las situaciones sospechosas y las que no lo son. Una vez que identifica el comportamiento deshonesto es capaz de encaminar la investigación hacia el tipo de fraude más probable. Esto se consigue determinando qué variables son explicativas para cada una de las alternativas de fraude. Los resultados permiten concluir que determinadas variables que no son significativas en el modelo multinomial, sí lo son en este caso. Una vez se sospecha de la existencia de fraude, ciertas circunstancias resultan explicativas del tipo de proceder más probable.

La implementación de un modelo anidado, aunque más compleja, permite estudiar el comportamiento de las probabilidades condicionadas, aspecto que no había sido tratado hasta este momento.

Los modelos planteados permiten, asimismo, determinar la capacidad de predicción de los comportamientos fraudulentos. Los resultados obtenidos para el porcentaje de siniestros correctamente clasificados en sus categorías respectivas son interpretados como la capacidad que la compañía tiene de detectar, en caso de que exista, un determinado tipo de fraude. La optimización del criterio probabilístico utilizado en la delimitación de las zonas de clasificación de los siniestros dentro de una determinada categoría ha permitido mejorar notablemente la calidad predictiva de los modelos utilizados. La demostración del proceso a seguir para optimizar el punto de corte en el marco de la modelización logística multinomial y anidada constituye, sin duda, una de las principales aportaciones, en lo que a la vertiente teórica se refiere.

En la presentación final se ha optado por dar mayor relevancia a la capacidad explicativa de los modelos estimados. Es por ello que no se ha dividido la muestra con la finalidad de poder validar los resultados con datos no empleados en la estimación. A ello se ha sumado la intención de mantener el tamaño muestral que conduce a la estimación de los parámetros, lo mayor posible.

En relación a las variables explicativas cabe destacar un hecho importante: para todas ellas la compañía dispone de información de forma inmediata, de manera que la investigación sobre la existencia o no de fraude puede realizarse sin una demora excesiva.

A diferencia de los estudios realizados por otros autores, los modelos presentados no incluyen indicadores subjetivos o aquéllos para los que la información se recaba durante la tramitación del siniestro pero no en el momento inicial de su apertura.

7.- Las diferencias existentes con las aportaciones realizadas por otros autores se hacen patentes en el tratamiento metodológico aplicado y en los indicadores de fraude utilizados, además de que aquí consideramos un conjunto de alternativas de elección más amplio que el asociado a la dicotomía fraude/no fraude.

7.1.- El enfoque planteado es distinto al utilizado por Picard (1996), que realiza un análisis microeconómico teórico del comportamiento deshonesto de los asegurados y sugiere la creación de un organismo (agencia) común a todas las aseguradoras, que se encargue total o parcialmente de los gastos derivados de la investigación de siniestros, para fomentar que las entidades auditen el fraude sin alejarse del criterio económico de minimización de costes.

Nuestro trabajo está más en la línea de los estudios realizados por Weisberg y Derrig (1993, 1997) y por Belhadji y Dionne (1997), ya que presentan análisis cuantitativos para identificar los indicadores de fraude.

7.2.- Weisberg y Derrig (1993, 1997) tienen en cuenta la existencia de categorías genéricas de fraude. Se consideran cuatro posibilidades a la hora de clasificar el siniestro: legítimo, con sospecha de fraude planeado -siniestro construido o simulado-, con sospecha de fraude oportunista -ocurrencia del accidente real pero sin daños aparentes- y con sospecha de gastos médicos inflados.

Sin embargo, mientras que en esta Tesis la variable dependiente es cualitativa y viene definida por el tipo de comportamiento observado para los individuos de la muestra, Weisberg y Derrig

especifican como variable dependiente la “sospecha” de existencia de una determinada actuación puntuada de 0 a 10. En su modelización hacen intervenir una serie de indicadores relacionados con diferentes aspectos del siniestro, de las partes que intervienen y de los daños derivados (otorgan especial importancia a los costes médicos). Mediante la aplicación de técnicas de regresión múltiple, determinan cuáles resultan estadísticamente significativos en la explicación de la dependiente. Las diferencias entre nuestro trabajo y el de dichos autores quedan patentes no sólo en la metodología aplicada, sino también en el objetivo final perseguido. Para estos últimos, el resultado final del estudio es la determinación de las características básicas que permiten fundamentar la existencia de sospecha de un tipo de fraude (los patrones de comportamiento definidos son susceptibles de implementación automática). En esta Tesis, el análisis de la significación estadística de las variables introducidas en el modelo goza de una especial importancia, pero forma parte de un proceso más amplio, dirigido a cuantificar la probabilidad de existencia de cada tipo de fraude y a aumentar la probabilidad de que la compañía logre detectarlos lo antes posible.

7.3.- En relación al trabajo realizado por Belhadji y Dionne (1997), las principales similitudes se encuentran en los objetivos perseguidos tras la modelización. La cuantificación de la probabilidad de existencia o sospecha de fraude (previa selección de los indicadores a introducir en la especificación del modelo y análisis estadístico de los mismos) y la determinación del criterio probabilístico óptimo a utilizar por la compañía para clasificar correctamente los siniestros en su categoría respectiva (fraude/no fraude) son los principales resultados extraídos del análisis realizado por ambos autores. Presentan, también, una aproximación al tratamiento del fraude teniendo en cuenta los costes de investigación (análisis coste-beneficio asociado a la realización de un estudio exhaustivo del siniestro). Sin embargo las diferencias se hacen evidentes al considerar el tratamiento metodológico aplicado. Así, y respecto a nuestra aportación, Belhadji y Dionne no tienen en cuenta la existencia de diferentes alternativas al alcance del individuo que quiere actuar fraudulentamente. En su estudio, todos los resultados aparecen asociados a la dicotomía fraude/no fraude, sin que puedan interpretarse teniendo en cuenta un proceso de decisión por etapas. La técnica econométrica utilizada es, de esta forma, más simple: frente a la aplicación de modelos lógit multinomiales y anidados, ellos proponen el uso de un modelo próbit. Entendemos que esta limitación les viene provocada por el tipo de información estadística de que disponen. Al tratarse de una muestra aleatoria de siniestros, el porcentaje de casos fraudulentos es excesivamente bajo.

La definición de patrones de comportamiento por parte de estos autores queda limitada al caso general de presencia-ausencia de fraude y no es posible determinar la existencia de atributos

directamente relacionados con formas alternativas de fraude. Ello es especialmente peligroso para aquellas variables cuya influencia en la probabilidad de cometer un cierto tipo de fraude difiere de la de otro tipo, sobre todo si el parámetro asociado cambia de signo.

Tanto Weisberg y Derrig como Belhadji y Dionne utilizan indicadores de fraude caracterizados, en algunos casos, por una elevada subjetividad (por ejemplo, la valoración que los tramitadores hacen de la conducta del asegurado) y, en otros, por una tardanza en su disposición (número de días en que el asegurado ha recibido tratamiento médico,...). Ésta es también una de las grandes diferencias con nuestro trabajo, dado, que como hemos manifestado anteriormente, las variables utilizadas en la modelización poseen una elevada objetividad y para todas ellas la compañía dispone de información de forma inmediata. No obstante, la comparación con los indicadores utilizados a nivel internacional nos ha alertado sobre la existencia de ciertas variables, cuya información puede ayudar notablemente en el estudio del fraude, que no son recogidas actualmente en los partes de siniestros de las entidades españolas (por ejemplo, la situación socioeconómica del demandante o indicadores indirectos de la misma).

Respecto al análisis de la significación individual de las variables no se han encontrado diferencias alarmantes respecto a las conclusiones de otros autores, o que no puedan ser explicadas por los diferentes sistemas aseguradores.

8.- Pasamos a continuación a destacar las principales conclusiones de los resultados obtenidos en la modelización, tanto multinomial como anidada.

Las variables explicativas que tienen un efecto significativo en todos los modelos son la coincidencia de apellidos entre las partes implicadas, la intervención policial, los antecedentes de siniestralidad que el asegurado posee en la compañía y la descripción textual de lo ocurrido. Podríamos calificar a estas variables como las más relevantes o las que se ha de considerar siempre en la investigación del fraude. Esta conclusión coincide con la de otros investigadores que insisten en la conveniencia de que las autoridades emitan informes de siniestros con más frecuencia. El resto de variables corresponden a circunstancias propias del suceso y, por lo tanto, poco se puede hacer institucional o corporativamente para cambiarlas.

Cuanto mayor es el historial de siniestros que el asegurado posee en la compañía, mayor es la probabilidad de que cometa cualquier tipo de fraude (el conocimiento de la forma en que la entidad tramita los siniestros -rapidez, peritajes poco exhaustivos,...- y la no detección de posibles fraudes anteriores, pueden ser algunas de las razones que ayudan a interpretar el

resultado obtenido). La coincidencia de apellidos entre la parte asegurada y la parte contraria es, sin duda, uno claro indicador de cooperación que induce a sospechar de la existencia de fraude. El diseño de un programa destinado a realizar el análisis de los apellidos entre las partes que intervienen en el accidente ha permitido detectar un elevado porcentaje de casos en los que existe coincidencia, en relación a la muestra de siniestros fraudulentos. La presencia en la declaración del asegurado de descripciones alusivas al estacionamiento, adelantamiento, cruce y marcha atrás aparece relacionada con un aumento en la propensión de que exista cualquiera de los tipos de fraude comentados.

A modo de síntesis presentamos otras variables que muestran cierto poder explicativo en los modelos multinomiales y, por lo tanto, las podemos considerar interesantes para diferenciar tipos de comportamientos.

- La ocurrencia del siniestro a altas horas de la madrugada y la demora en la comunicación del siniestro son variables directamente relacionadas con la aparición de fraude. En relación a la existencia de testigos del accidente se observa una influencia contrapuesta en los tipos de fraude analizados: produce una disminución en la probabilidad de aparición de fraude para conseguir una indemnización y de fraude cometido en beneficio propio, e incrementa la probabilidad de aparición de fraude cometido en beneficio de un tercero (existencia de testigos falsos). Ello constituye una evidencia de que las variables pueden tener influencia distinta en la probabilidad de aparición de cada tipo de fraude. Este resultado es muy interesante puesto que otros autores concluyen que la influencia de esta variable en la presencia de fraude no es significativa, pese a que éste no sea el resultado esperado. En nuestro caso se confirma su diferente influencia en los distintos tipos de comportamiento, que se anula si se consideran éstos agregadamente.
- Cuando el vehículo asegurado es un turismo de uso particular disminuye la probabilidad de aparición de fraude (la variable es estadísticamente significativa en la explicación de fraude en beneficio propio y de fraude para conseguir una indemnización). La propensión a defraudar es mayor, por tanto, en aquellos casos en los que la utilización del turismo es diferente o cuando los vehículos pertenecen a otras categorías como motocicletas o camiones.
- A medida que aumenta la edad del individuo disminuye la probabilidad de que cometa fraude, siendo más frecuente la aparición de comportamientos fraudulentos en los individuos jóvenes (con edades, básicamente, entre 25 y 34 años, en el momento del siniestro).

- La antigüedad de carnet del conductor asegurado, considerada en la modelización multinomial de fraude a favor del asegurado y a favor del tercero, no goza de significación estadística. La misma conclusión se obtiene para la existencia de franquicia en la póliza y para la contratación de cobertura de accesorios, introducidas al modelizar la probabilidad de aparición de fraude para conseguir una indemnización o para incrementarla.

Respecto a la modelización del tipo de fraude en las ramas inferiores de un esquema anidado, como hemos señalado anteriormente, las variables significativas indican cómo dirigir la investigación del siniestro, es decir, como identificar cada tipo de fraude.

- Cuanto más antiguo sea el vehículo del asegurado mayor es la probabilidad de que actúe fraudulentamente en beneficio propio y no en beneficio de otro (por ejemplo, reparando averías cada vez más frecuentes) aunque se ha detectado la existencia de un punto de inflexión (aproximadamente, en el séptimo año de antigüedad) a partir del cual dicha probabilidad disminuye, como consecuencia, probablemente, del bajo valor venal del automóvil.
- Cuando se detecta que el siniestro ha ocurrido entre la fecha de efecto y la fecha de emisión de la póliza (por este orden) aumenta la probabilidad de que el asegurado haya cometido fraude para conseguir una indemnización y no para incrementarla.
- Cuando el asegurado tiene contratada únicamente cobertura a “terceros” y cuando admite la culpa del siniestro, aumenta la probabilidad de que cometa fraude en beneficio de un tercero frente a que lo haga en el suyo propio, como es de esperar.

Todos los modelos presentados en los Capítulos anteriores son globalmente significativos y los estadísticos de ajuste son satisfactorios.

9.- Las líneas de mejora propuestas para un futuro, en relación al proceso de recogida de datos, se centran básicamente en dos aspectos: 1) el análisis de la existencia de comportamientos fraudulentos considerando categorías de fraude alternativas a las presentadas en la Tesis y, 2) el diseño de una base de datos más amplia que nos permita trabajar con un mayor número de casos para cada una de las alternativas consideradas.

9.1.- En relación a la primera línea, creemos que sería positivo, no sólo a nivel investigador, sino también a nivel de entidades, el diseño de “patrones de tipos de fraude”,

adicionales a los presentados en la Tesis, que ayudase a los tramitadores a clasificar los siniestros con fraude detectado en la categoría conveniente. En la actualidad ya se trabaja con categorías de fraude preestablecidas, pero la existencia de un importante número de casos en los que la compañía no identifica el tipo de fraude nos ha alertado de la dificultad asociada al proceso de clasificación. El tramitador puede tener problemas a la hora de identificar de forma explícita la clase de fraude existente en el siniestro. En este sentido podría pensarse en alternativas adicionales de clasificación a las presentadas en este trabajo, entre las que queremos destacar la que nos llevaría a diferenciar (siguiendo la pauta establecida por el Insurance Fraud Bureau de Massachusetts) entre aquellos siniestros en los que existe fraude “*a priori*” (el asegurado fabrica deliberadamente el siniestro) y aquellos en los que existe fraude “*a posteriori*” (la ocurrencia del siniestro es legal, pero el asegurado y/o el contrario la utilizan para obtener un beneficio no justificado).

9.2.- El reducido número de observaciones para cada una de las submuestras utilizadas (teniendo en cuenta cada una de las categorías de fraude planteadas) ha dificultado la realización de un estudio de indicadores más exhaustivo dirigido, fundamentalmente, a la inclusión de interacciones entre variables. La disponibilidad de un mayor número de casos para cada tipo ayudaría a refinar los patrones de comportamiento definidos, teniendo en cuenta la posibilidad de recoger en el modelo, mediante la creación adecuada de variables, algunas de las normas que regulan el funcionamiento del seguro del automóvil (considerando, por ejemplo, de forma conjunta la culpa del siniestro y el tipo de cobertura,...). Las interacciones entre variables utilizadas en la estimación han estado dirigidas a la definición de atributos de elección específicos para cada una de las alternativas consideradas.

10.- Desde el punto de vista metodológico, las líneas de mejora irían encaminadas al tratamiento de dos aspectos adicionales: 1) la introducción de errores en la medida de la variable dependiente y 2) el estudio longitudinal de la vida de las reclamaciones.

10.1.- Los modelos de variable dependiente cualitativa con errores de medida han sido recientemente introducidos en la literatura estadística. Su aplicación permitiría mejorar resultados analizando si la omisión de este error tiene efectos significativos en los resultados que hemos presentado. El contexto donde se debería aplicar la corrección permite afirmar que el error de medida se produce únicamente en una de las categorías: la de las reclamaciones consideradas legítimas (que bien podrían ser fraudes no detectados).

10.2.- La consideración de un estudio longitudinal requeriría trabajar con muestras de expedientes de siniestros desde su apertura hasta su liquidación. Con ello, se lograría detectar

qué indicadores inciden en cada etapa de la tramitación y, por tanto, qué momento es óptimo para realizar los controles de fraude.

11.- Las líneas de mejora propuestas, desde el punto de vista de la eficiencia asociada a la implantación de un mecanismo de control de fraude por parte de una entidad, son también fundamentalmente dos: 1) la realización de un análisis exhaustivo de la influencia que el fraude tiene en el coste del siniestro para la entidad y 2) la determinación de la rentabilidad asociada a la instalación de un sistema automático de control de fraude teniendo en cuenta la relación coste-beneficio asociada al mismo.

11.1.- El hecho de que en la actualidad no dispongamos de información pormenorizada referida al coste del siniestro ha impedido estudiar cómo influye la detección del fraude en la cantidad finalmente pagada por una compañía. La consideración del flujo de costes asociados (coste inicial estimado del siniestro según valoraciones técnicas -constitución de la reserva;- coste del siniestro tras la valoración pericial -modificación de la reserva; coste final del siniestro teniendo en cuenta la posible existencia de recobros; coste final tras la detección del fraude) y la determinación de los costes de investigación puede ayudar a determinar las cantidades realmente ahorradas por una compañía al instalar un mecanismo de detección. Los resultados presentados al final del Capítulo 6 y 7 sólo pueden considerarse una primera aproximación en términos de costes esperados y han sido determinados utilizando hipótesis muy restrictivas.

11.2.- Por último, y si bien ha quedado suficientemente demostrado a lo largo de la Tesis que los modelos desarrollados pueden considerarse importantes herramientas para detectar y controlar el fraude, queda pendiente determinar la rentabilidad que para una entidad puede suponer su incorporación en el sistema de tramitación de los siniestros. La comparación entre el coste de instalación del mecanismo de detección y los beneficios derivados del mismo (ahorro en las cantidades finalmente pagadas por cada siniestro) determinará la rentabilidad que tiene para una compañía el controlar el fraude.

BIBLIOGRAFÍA

- A.I.B. (1997) 1997 Annual Report. *Automobile Insurers Bureau of Massachusetts*. Boston.
- Addis, K.K. (1989) "The Fraud Squad". *Security Management*, September, vol. 33, nº 9, 53-57.
- Allingham, M.G. y A. Sandmo (1972) "Income Tax Evasión: A Theoretical Analysis". *Journal of Public Economics*, vol. 1, nº 3/4, 323-338.
- Amemiya, T. (1985) *Advanced Econometrics*. Ed. Basil Blackwell Ltd., Oxford, U.K.
- Anderson, K. (1990) "The Phoney-Claims Drain". *Insurance Review*, vol. 51, nº 8, 20-23.
- Artís, M., M. Ayuso y M. Guillén (1997) "Modelling Different Types of Automobile Insurance Fraud Behaviour in the Spanish Market". Ponencia presentada en el primer Congreso Internacional en *Insurance: Mathematics and Economics*, celebrado en Amsterdam del 25 al 27 de agosto de 1997. En revisión para su posible publicación en la revista *Insurance: Mathematics and Economics*, Elsevier Science Publishers, North-Holland.
- Artís, M., M. Ayuso y M. Guillén (1998) "Aplicación de un Modelo Lógit a la Cuantificación de la Probabilidad de Fraude en el Seguro del Automóvil". *Working Paper*. Departamento de Econometría, Estadística y Economía Española. Universidad de Barcelona.
- Ayuso, M. (1995) *El Fraude en el Seguro del Automóvil*. Trabajo de Investigación realizado dentro del Programa de Doctorado *Economía y Territorio: Análisis Cuantitativo*. Departamento de Econometría, Estadística y Economía Española.

- Ayuso, M. y M. Guillén (1995) "Modelos de Elección del Fraude en el Seguro de Automóviles". *Actas de la IX Reunión de la Asociación Científica Europea de Economía Aplicada ASEPELT-España* celebrada en Santiago de Compostela, 22 y 23 de junio de 1995, vol. II (*Economía Sectorial*), 383-394.
- Ayuso, M. y M. Guillén (1996) "Modelos de Detección de Fraude en el Seguro del Automóvil". Aceptado para su publicación en la Revista *Cuadernos Actuariales*.
- Belhadji, E.B. y G. Dionne (1997) "Development of an Expert System for the Automatic Detection of Automobile Insurance Fraud". *Working Paper 97-06*. École des Hautes Études Commerciales. Université de Montréal.
- Bernouilli, D. (1954) "Exposition of a New Theory of the Measurement of Risk". *Econometrica*, vol. 22, nº 1, 23-36.
- Betés de Toro, A. (1996) "La Estafa en el Seguro". *Actualidad Aseguradora*, nº 2, 22 de enero, 9.
- Bollinger, C. R. y M. H. David (1997) "Modeling Discrete Choice With Response Error: Food Stamp Participation". *Journal of the American Statistical Association*, vol. 92, nº 439, 827-835.
- Bond, E.W. y K.J. Crocker (1997) "Hardball and the Soft Touch: The Economics of Optimal Insurance Contracts with Costly State Verification and Endogenous Monitoring Costs". *Journal of Public Economics*, vol. 63, nº 2, 239-264.
- Borch, K. (1990) *Economics of Insurance*. Advanced Textbooks in Economics, vol. 29, Ed. North-Holland, Amsterdam.
- Brockett, P.L., X. Xia y R. A. Derrig (1995) "Using Kohonen's Self-Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud". Ponencia presentada en *American Risk and Insurance Association Annual Meeting*, Seattle.
- Brotman, B.A. y P. Fox (1988) "The Impact of Economic Conditions on the Incidence of Arson: Comment". *Journal of Risk and Insurance*, vol. 55, nº 4, 751-755.

- Brundin, I. y F. Salanié (1997) "Fraud in the Insurance Industry: An Organizational Approach", preliminary, April 1997. *mimeo*.
- Bujold, L., G. Dionne y R. Gagné (1997) "Assurance Valeur à Neuf et Vols d'Automobiles: Une Étude Statistique". *Cahier de recherche* 97-01, février 1997. École des Hautes Études Commerciales. Université de Montréal.
- Caron, L. y G. Dionne (1996) "Insurance Fraud Estimation: More Evidence from the Quebec Automobile Insurance Industry". *Working Paper* 96-02, december 1996. École des Hautes Études Commerciales. Université de Montréal.
- C.E.S. (1992) *El Fraude en el Seguro de Automóviles*. Centro de Estudios del Seguro. Madrid.
- Cercós, J. y R. Mayol (1990) Seguro de Automóviles. *Tratado General de Seguros, Teoría y Práctica de los Seguros Privados*, tomo III, vol. 3. Ed. Consejo General de los Colegios de Agentes y Corredores de Seguros de España, 145-301.
- Chao, J. (1994) "Nuevo Frente contra el Fraude". *Mercado Previsor*, nº 195, 30 de octubre, 12-13.
- Clarke, M. (1990) "The Control of Insurance Fraud. A comparative view". *The British Journal of Criminology*, vol. 30, nº 1, 1-23.
- Cloninger, D.O. (1990) "Arson and Abandonment: A Restatement". *The Journal of Risk and Insurance*, vol. 57, nº 3, 540-545.
- Cobo, P. (1993) *Manual de Investigación de Siniestros y Lucha contra el Fraude en el Seguro de Automóviles*. Ed. Mapfre, Madrid.
- Comité Europeo de Seguros (1996) *Le Guide de l'anti-fraude à l'assurance en Europe*. CEA Info, Hors - Série nº 4, Mai 1996.

- Correlation Research, Inc. (1989) *Marb Baseline Study of Bodily Injury Liability Claims*.
- Correlation Research, Inc. (1990) *AIB Study of 1989 Bodily Injury Liability Claims*.
- Correlation Research, Inc. (1991) *AIB Study of 1989 PIP Claims*.
- Correlation Research, Inc. (1993) *AIB Study of Suspected Fraud. Supplementary Coding Form for 1989 BI and PIP Claims*.
- Crocker, K.J. y J. Morgan (1998) "Is Honesty the Best Policy? Curtailing Insurance Fraud through to Optimal Incentive Contracts". *Journal of Public Economics*, vol. 106, n° 2, 355-375.
- Cummins, J.D. y R.A. Derrig (1993) "Fuzzy Trends in Property-Liability Insurance Claim Costs". *Journal of Risk and Insurance*, vol. 60, n° 3, 429-465.
- Cummins, J.D. y R.A. Derrig (1997) "Fuzzy Financial Pricing of Property-Liability Insurance". *North American Actuarial Journal*, vol. 1, n° 4, 21-44.
- Cummins, J.D. y S. Tennyson (1992) "Controlling Automobile Insurance Costs". *Journal of Economic Perspectives*, vol. 6, n° 2, 95-115.
- Cummins, J.D. y S. Tennyson (1996) "Moral Hazard in Insurance Claiming: Evidence from Automobile Insurance". *Journal of Risk and Uncertainty*, vol. 12, n° 1, 29-50.
- Cummins, J.D. y M.A. Weiss (1991) "The Effects of No-Fault on Automobile Insurance Loss Costs". *Geneva Papers on Risk and Insurance*, vol. 16, n° 1, 20-38.
- Cummins, J.D. y M.A. Weiss (1993) "Measuring Cost Efficiency in the Property-Liability Insurance Industry". *Journal of Banking and Finance*, vol. 17, n° 3, 463-482.
- Derrig, R.A. y L.K. Krauss (1994) "First Steps to Fight Workers' Compensation Fraud". *Journal of Insurance Regulation*, vol. 12, n° 3, 390-415.

- Derrig, R.A. y K.M. Ostaszewski (1994a) Fuzzy Sets. *Cas Ratemaking Seminar*. Atlanta, Georgia, March 10-11, 1994.
- Derrig, R.A. y K.M. Ostaszewski (1994b), "Fuzzy Techniques of Pattern Recognition in Risk and Claim Classification", *4th International Conference on Insurance Solvency and Finance*. Philadelphia, PA. April 25-27, 1994.
- Derrig, R.A. y H.I. Weisberg (1995) "Identification and Investigation of Suspicious Claims. A Report on the AIB Study of 1993 Personal Injury Protection Claims". *AIB Filing on Fraudulent Claims Payment*, DOI Docket R95-12, Boston.
- Derrig, R.A. y H.I. Weisberg (1996) "A Report on the AIB Study of 1993 Personal Injury Protection and Bodily Injury Liability Claims. Coping with the Influx of Suspicious Strain and Sprain Claims". *AIB Filing on Cost Containment and Fraudulent Claims Payment*, DOI Docket R96-36, Boston.
- Derrig, R.A. y H.I. Weisberg (1997) "AIB PIP Claim Screening Experiment Interim Report. Understanding and Improving the Claim Investigation Process". *AIB Filing on Fraudulent Claims Payment*, DOI Docket R97-37, Boston.
- Derrig, R.A., H.I. Weisberg, y X. Chen (1994) "Behavioral Factors and Lotteries Under No-Fault with Monetary Threshold: A Study of Massachusetts Automobile Claims". *Journal of Risk and Insurance*, vol. 61, n° 2, 245-275.
- Dionne, G. y R. Gagné (1997) The Non-Optimality of Deductible Contrats Against Fraudulent Claims: An Empirical Evidence in Automobile Insurance. *Working Paper 97-05*, june 1997. École des Hautes Études Commerciales. Université de Montréal.
- Elguero, J. M. (1996) "Código Penal: ¿Cómo Afecta al Fraude en el Seguro?. *Actualidad Aseguradora*, n° 2, 22 de enero, 36-37.
- Fagart, M.C. y P. Picard (1997) "Optimal Insurance Under Random Auditing". Preliminary and incomplete, September 1997. *mimeo*.

- Fernández de Castro, J. y Tugores, J. (1988) *Fundamentos de Microeconomía*. Ed. McGraw-Hill, Madrid.
- Fernández de Castro, J. y Tugores, J. (1992) *Fundamentos de Microeconomía*. Segunda Edición. Ed. McGraw-Hill, Madrid.
- García, J.L. (1989) "Fraud: The Hidden Element in Health Care Costs". *Compensation and Benefits Management*, vol. 6, nº 1, 49-51.
- Garrigues, J. (1974) *Curso de Derecho Mercantil*, vol. 2, 6ª edición revisada con la colaboración de F. Sánchez Calero, Ed. Aguirre, Madrid.
- González, Y. (1993) *La Demanda de Seguro Sanitario: Una Elección Bajo Incertidumbre*. Tesis Doctoral. Departamento de Econometría, Estadística y Economía Española de la Universidad de Barcelona.
- Greene, W.H. (1993) *Econometric Analysis*. Second Edition. Ed. Macmillan Press. New York.
- Greene, W.H. (1995) *LIMDEP, Version 7.0, User's Manual*. Ed. Econometric Software, Inc. New York.
- Greene, W.H. (1997) *Econometric Analysis*. Third Edition. Ed. Prentice-Hall International, Inc. New York.
- Greenouff, M. (1990) "El Fraude en el Seguro del Automóvil. Una comparación Internacional". Información Comité Europeo de Seguros, C.E.A. IV Jornadas Comunitarias del Seguro del Automóvil. *Previsión y Seguros*, nº 8, noviembre-diciembre, 109-120.
- Hausman, J. (1975) "An Instrumental Variable Approach to Full Information Estimators for Linear and Certain Nonlinear Econometric Models", *Econometrica*, vol. 43, nº 4, 727-738.
- Hausman, J. y D. McFadden (1984) "Specification Tests for the Multinomial Logit Model", *Econometrica*, vol. 52, nº 5, 1219-1240.

- Hensher, D. (1986) "Simultaneous Estimation of Hierarchical Logit Mode Choice Models" *Working Paper*, 24, MacQuarie University, School of Economic and Financial Studies.
- Hershberger, R.A. y R.K. Miller (1978) "The Impact of Economic Conditions on the Incidence of Arson: A Reply", *Journal of Risk and Insurance*, vol. 45, nº 2, 275-290.
- Hershberger, R.A. y R.K. Miller (1988) "The Impact of Economic Conditions on the Incidence of Arson: A Reply", *Journal of Risk and Insurance*, vol. 55, nº 4, 755-757.
- Hevia, E. y J.J. Lafuente (1992) *Como Luchar contra el Fraude en la Empresa*. Ed. Instituto de Auditores Internos de España, Madrid.
- Hoyt, R.E. (1990) "The Effect of Insurance Fraud on the Economic System". *Journal of Insurance Regulation*, vol. 8, nº 3, 304-315.
- I.C.E.A. (1995) "Manual de Lucha contra el Fraude en el Seguro de Automóviles". *Informe nº 607*, abril 1995.
- I.C.E.A. (1996) "Estadística de Fraudes en Seguros, 1995". *Informe nº 642*, mayo 1996.
- I.C.E.A. (1997) "Fraude: Una Lucha Permanente". *Actualidad Aseguradora*, nº12, 7 de abril, 60-62.
- Institute of Internal Auditors, Inc. (1985) "Deterrence, Detection, Investigation and Reporting of Fraud", *Statement on Internal Auditing Standards (SIAS)*, nº 3, 1-20.
- I.F.B. (1994) *1994 Annual Report*. Ed. Insurance Fraud Bureau of Massachusetts, Boston.
- I.F.B. (1994) *Insurance Fraud Research Register*. Ed. Insurance Fraud Bureau of Massachusetts, Boston.
- I.F.B. (1995) *1995 Annual Report*. Ed. Insurance Fraud Bureau of Massachusetts, Boston.

- I.F.B. (1995) *Insurance Fraud Research Register*. Ed. Insurance Fraud Bureau of Massachusetts, Boston.
- I.F.B. (1996) *1996 Annual Report*. Ed. Insurance Fraud Bureau of Massachusetts, Boston.
- I.F.B. (1996) *Insurance Fraud Research Register*. Ed. Insurance Fraud Bureau of Massachusetts, Boston.
- I.F.B. (1997) *1997 Annual Report*. Ed. Insurance Fraud Bureau of Massachusetts, Boston.
- I.F.B. (1998) "Focus Fraud". *Insurance Fraud Bureau of Massachusetts Publications*, vol. 5, nº 1.
- I.N.E.S.E. (1996a) "Organización del Sector en la Lucha contra el Fraude". Departamento de Investigación. *Actualidad Aseguradora*, nº 2, 22 de enero, 2-4.
- I.N.E.S.E. (1996b) "El Fraude en el Ramo de Autos en Europa". *Actualidad Aseguradora*, nº 2, 22 de Enero, 5-8.
- I.N.E.S.E. (1996c) "Fraude al Seguro: Alarma Roja". *Actualidad Aseguradora*, nº 24, 1 de julio, 2-5.
- I.N.E.S.E. (1996d) "¡Ojo al Fraude!". *Actualidad Aseguradora*, nº 24, 1 de julio, 6-8.
- Johnston, J. (1992) *Métodos de Econometría*. Ed. Vicens-Vives, Barcelona.
- Kleinbaum, D. G. (1992) *Logistic Regression: a self-learning text*. Ed. Springer, New York.
- Kreps, D.M. (1990) *A Course in Microeconomic Theory*. Ed, Harvester Wheatsheaf, New York.
- Koujianou, P. (1995) "Product Differentiation and Oligopoly in International Markets: The Case of the U.S. Automobile Industry". *Econometrica*, vol. 63, nº 4, 891-951.
- Lara, I. (1996) "La solución se llama FIVA". *Actualidad Aseguradora*, nº 2, 22 de enero, 34-35.

- Lemaire, J. (1985) *Automobile Insurance*. Ed. Kluwer-Nijhoff Publishing, U.S.A.
- Lemaire, J. (1990) Fuzzy Insurance. *Astin Bulletin*, vol. 20, nº 1, 33-55.
- Lemaire, J. (1995) *Bonus-Malus Systems in Automobile Insurance*. Ed. Kluwer Academic Publishers, U.S.A.
- Ley 122/1962, de 24 de diciembre, de Uso y Circulación de Vehículos de Motor (Boletín-Gaceta nº 310 de 27 de diciembre de 1962).
- Ley 50/1980, de 8 de octubre, de Contrato de Seguro (B.O.E. nº 250 de 11 de octubre de 1980).
- Ley 30/1995, de 8 de noviembre, de Ordenación y Supervisión de los Seguros Privados (B.O.E. nº 268 de 9 de noviembre de 1995).
- Ley 10/1995, de 23 de noviembre, Código penal (B.O.E. nº 281 de 24 de noviembre de 1995).
- Luce, R.D. (1959) *Individual Choice Behaviour*. Ed. Wiley, New York.
- Macho, I y D. Pérez (1994) *Introducción a la Economía de la Información*, 1ª edición, Ed. Ariel, Madrid.
- Maddala, G.S. (1983) *Limited-Dependent and Qualitative Variables in Econometrics*. Econometric Society Monographs. Ed. Cambridge University Press, Cambridge.
- Manski, C. y S. Lerman (1977) "The Estimation of Choice Probabilities from Choice-Based Samples". *Econometrica*, vol. 45, 1977-1988.
- Manski, C y D. McFadden (eds.) (1981) *Structural Analysis of Discrete Data with Econometric Applications*. Ed. MIT Press, Cambridge, Massachusetts.

- Marter, S.S. y H.I. Weisberg (1992) "Medical Costs and Automobile Insurance: A Report on Bodily Injury Liability Claims in Massachusetts". *Journal of Insurance Regulation*, vol. 9, nº 3, 381-422.
- Marter, S.S. y H.I. Weisberg (1992) "Medical Expenses and The Massachusetts Automobile Tort Reform Law: First Review of 1989 Bodily Injury Liability Claims". *Journal of Insurance Regulation*, vol. 10, nº 4, 462-514.
- Martin Boyer, M. (1997) Insurance Taxation and Insurance Fraud. *Working Paper 97-10*, September 1997. École des Hautes Études Commerciales. Université de Montréal.
- Martín Pliego, F.J. (1994) *Introducción a la Estadística Económica y Empresarial*. Ed. AC, Madrid.
- McFadden, D. (1978) "Modelling the Choice of Residential Location". En *Spatial Interaction Theory and Planning Models (Studies in Regional Science and Urban Economics, vol. 3)*. Ed. A. Karlquist et al., North-Holland, Amsterdam, 75-96.
- McFadden, D. (1981) "Econometric Models of Probabilistic Choice". En Manski, C. y D. McFadden (eds.) *Structural Analysis of Discrete Data with Econometric Applications*, Ed. MIT Press, Cambridge, Massachusetts. 198-272.
- McFadden, D. (1983) "Econometric Analysis Qualitative Response Models". En Griliches, Z. y M. Intrilligator (eds.) *Handbook of Econometrics*, vol. 2, cap. 24, Ed. North-Holland, Amsterdam, 1395-1457.
- Morillas Jarillo, M.J. (1992). *El Seguro del Automóvil: El Aseguramiento Obligatorio de la Responsabilidad Civil Automovilística*. Ed. José M^a Bosch, Barcelona.
- Mossin, J. (1992) "Aspects of Rational Insurance Purchasing". En Dionne, G. y S. Hamigton (eds.) *Foundations of Insurance Economics (Readings in Economics and Finance)*. Ed. Kluwer Academic Publishers, Boston, 119-133.
- Münchener Rück (1988) *El Fraude en el Seguro*. Ed. Münchener Rückversicherungs-Gesellschaft.. München.

- Münchener Rück (1994) *El Fraude en el Seguro de Personas*. Ed. Münchener Rückversicherungs-Gesellschaft, München.
- Norusis, M.J. (1990) *SPSS Advanced Statistics User's Guide*. SPSS Inc, U.S.A.
- Novales, A. (1993) *Econometría*. Segunda Edición. Ed. McGraw-Hill, Madrid.
- Paul, M. (1995) "El Fraude en el Seguro de Autos enciende la Alarma en Europa". *La Tribuna de Automoción*, 15 de noviembre, 8-9.
- Picard, P. (1996) "Auditing Claims in the Insurance Market with Fraud: The Credibility Issue". *Journal of Public Economics*, vol. 63, nº 1, 27-56.
- Picazo, A. (1995) "La Eficiencia en los Seguros". *Revista de Economía Aplicada, sección de Notas*, vol. 3, nº 8, 197-215.
- Pratt, J. W. (1964) "Risk Aversion in the Small and in the Large". *Econometrica*, vol. 32, nº 1-2, 122-136.
- Pudney, S. (1989) *Modelling Individual Choice. The Econometrics of Corners, Kinks and Holes*. Ed. Basil Blackwell Ltd, New York.
- Real Decreto Legislativo 1301/1986 de 28 de junio 1986, (Ministerio de Economía y Hacienda). Automóviles-Comunidad Económica Europea. Adaptación de la Ley de Uso y Circulación de Vehículos de Motor al Derecho de las Comunidades Europeas (B.O.E. nº 155 de 30 de junio de 1986).
- Savage, L. (1954) *The Foundations of Statistics*. Ed. John Wiley and sons., New York.
- Solana de Ariza, G. (1988) *La Declaración Amistosa de Accidente de Automóvil y el Convenio de Indemnización Directa Español (C.I.D.E.)*. Ed. Einsa, Madrid.

- Taricco, A. (1988) "Medical Consultant Key to Dealing Effectively with Insurance Fraud and Abuse", *Physician Executive*, September/October, 22-25.
- Texto refundido de la Ley de Uso y Circulación de Vehículos de Motor, aprobado por Decreto 632/1968 de 21 de marzo (B.O.E. nº 85 de 8 de abril de 1968).
- Tennyson, S. (1997) "Economic Institutions and Individual Ethics: A Study of Consumer Attitudes toward Insurance Fraud". *Journal of Economic Behavior and Organization*, vol. 32, 247-265.
- U.N.E.S.P.A. (1990) *Estadística de Seguros Privados. Análisis Económico Actuarial del Sector Asegurador Español, Ejercicio 1989*, Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1991a) *Estadística de Seguros Privados. Información Estadística del Seguro Privado, 1981-90*, vol. 2. Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1991b) *Informe Actuarial sobre Determinación de las Primas de Riesgo. Recargo de Seguridad y Recargo para Bonificaciones por No Siniestralidad (Sistema Bonus) en el Seguro de Responsabilidad Civil de Automóviles. Agrupación Nacional de Seguros de Automóviles*, Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1992) *Estadística de Seguros Privados. Información Estadística del Seguro Privado, 1982-91.*, Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1993a) *Informe Actuarial sobre Determinación de las Primas de Riesgo. Recargo de Seguridad y Recargo para Bonificaciones por No Siniestralidad (Sistema Bonus) en el Seguro de Responsabilidad Civil de Automóviles. Comisión Técnica de Seguros de Automóviles*, Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1993b) *Estadística de Daños Propios, Incendio, Robo y Rotura de Lunas del Seguro del Automóvil. Datos 1991. Comisión Técnica de Seguros de Automóviles*, Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1993c) *Estadística de Seguros. Cifras de Avance e Informe Económico, 1992*, Ed. Aseguradora, Madrid.

- U.N.E.S.P.A. (1993d) *Estadística de Seguros Privados. Información Estadística del Seguro Privado, 1983-92.*, Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1993e) *Estadística del Seguro de Responsabilidad Civil del Automóvil. Datos 1991. Comisión Técnica de Seguros de Automóviles.* Ed, Aseguradora, Madrid.
- U.N.E.S.P.A. (1993f) *Informe Actuarial sobre Determinación de las Primas de Riesgo. Recargo de Seguridad y Recargo para Bonificaciones por No Siniestralidad (Sistema Bonus) en las Modalidades de Daños Propios, Robo y Lunas del Seguro de Automóviles 1993.* Comisión Técnica de Seguros de Automóviles, Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1994a) *Estadística de Daños Propios, Rotura de Lunas e Incendio del Seguro del Automóvil. Datos 1992. Comisión Técnica de Seguros de Automóviles,* Ed. Aseguradora., Madrid.
- U.N.E.S.P.A. (1994b) *Estadística de Seguros. Cifras de Avance e Informe Económico, 1993,* Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1994c) *Estadística de Seguros Privados, 1984-93,* Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1994d) *Estadística del Seguro de Responsabilidad Civil del Automóvil. Datos 1992. Comisión Técnica de Seguros de Automóviles,* Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1994e) *Informe Actuarial sobre Determinación de las Primas de Riesgo. Recargo de Seguridad y Recargo para Bonificaciones por No Siniestralidad (Sistema Bonus) en el Seguro de Responsabilidad Civil de Automóviles. Comisión Técnica de Seguros de Automóviles,* Ed. Aseguradora, Madrid.
- U.N.E.S.P.A. (1994f) *Manual de Criterios. Comisión de Vigilancia y Arbitraje. C.I.D.E. y A.S.C.I.D.E. Comisión Técnica de Seguros de Automóviles,* Ed. Aseguradora, Madrid.

- Varian, H. (1992) *Análisis Microeconómico*. Tercera Edición. Ed. Antoni Bosch editor, Barcelona.
- Vaughan, E.J. y C.M. Elliot (1978) *Fundamentals of Risk and Insurance*. Second Edition. Ed. Wiley/Hamilton, New York.
- Von Neumann, J. y O. Morgenstern (1944) *Theory of Games and Economic Behaviour*. Ed. Princeton University Press.
- Weiss, M.A. (1991) "Efficiency in the Property-Liability Insurance Industry". *Journal of Risk and Insurance*, vol. 58, nº 3, 452-479.
- Weisberg, H.I. y R.A. Derrig (1991) "Fraud and Automobile Insurance: A Report on the Baseline Study of Bodily Injury Liability Claims in Massachusetts". *Journal of Insurance Regulation*, vol. 9, nº 4, 497-541.
- Weisberg, H.I. y R.A. Derrig (1992a) "Massachusetts Automobile Bodily Injury Tort Reform". *Journal of Insurance Regulation*, vol. 10, nº 3, 384-440.
- Weisberg, H.I. y R.A. Derrig (1992b) "Compensation for Automobile Injuries Under No-Fault with a Monetary Threshold: A Study of Massachusetts Claims". *AIB Filing for 1993 Private Passenger Automobile Rates*, DOI Docket G92-21, 28-94.
- Weisberg, H.I. y R.A. Derrig (1992c) "The System Misfired". *Best's Review*, December, vol. 93, nº 8, 37-40 y 87.
- Weisberg, H.I. y R.A. Derrig (1992d) "Pricing Auto No-Fault and Bodily Injury Liability Coverages Using Micro-Data and Statistical Models". *Casualty Actuarial Society*, 1993 Ratemaking Call Paper Program.
- Weisberg, H.I. y R.A. Derrig (1993) "Quantitative Methods for Detecting Fraudulent Automobile Bodily Injury Claims". *AIB Cost Containment/Fraud Filing*, 49-82.
- Wilmot, R.T.D. (1991) "World Insurance Forum: London". *Best's Review*, (Property/Casualty), vol. 92, nº 7, 97-98.

ANEXO 1

A continuación se muestran los principales indicadores de fraude seleccionados por el Insurance Fraud Bureau de Massachusetts y por la École des Hautes Études Commerciales de la Universidad de Montreal. La mayoría de ellos contienen información que bien directa o indirectamente es considerada por los aseguradores españoles. No obstante, es de destacar la importancia que ambos organismos conceden a las valoraciones que el tramitador hace de ciertos aspectos de la declaración del siniestro (nerviosismo del asegurado, rápida aceptación de culpa...). Destaca, asimismo, el interés sobre el tratamiento médico derivado del accidente y sobre la situación laboral y familiar del demandante.

1.1 Relación de indicadores potenciales de fraude utilizados por el I.F.B. (Insurance Fraud Bureau of Massachusetts).

A. Características del accidente

- No hay informe policial sobre el accidente (ausencia de atestado).
- No hay testigos del accidente.
- Colisión en la parte trasera del vehículo.
- Accidente referente a un único vehículo.
- Colisión en un cruce.
- El demandante estaba en un vehículo aparcado.
- Los dos conductores estaban emparentados o eran amigos.
- Accidente a altas horas de la noche (por la noche).
- Explicación no admisible del accidente.
- Demandante en un vehículo antiguo y de bajo valor.
- Coche alquilado implicado en el accidente.
- No remolque desde el lugar del accidente a pesar de que el coche estaba seriamente dañado.
- La investigación del lugar del suceso levantó preguntas.
- El daño material era inconsistente con el accidente.
- Choque con impacto sin importancia.
- El vehículo del demandante paró en seco.
- El vehículo del demandante hizo una maniobra inesperada.
- Discrepancias entre las versiones del demandante y asegurado.
- El asegurado se sintió ofendido, negando su culpa.

B. Características del demandante (Daños Corporales)

- Contrató un abogado muy rápidamente.
- Tenía un historial de siniestros anteriores.
- Dio su dirección como hotel o apartado de correos.
- Residía fuera del estado.
- Gastó mucho en abogados.
- Fue difícil contactar con él; no cooperó.
- Era uno de los tres o más demandantes del vehículo.
- Residía en una ciudad de elevado riesgo (con nivel de siniestros elevado).
- Evitó el uso del teléfono o del correo.
- Estaba sin trabajo.
- Parecía ser "prudente para los siniestros".

C. Características del conductor asegurado (Daños Corporales)

- Tenía un historial de siniestros anteriores.
- Dio su dirección como hotel o apartado de correos.
- Aceptó fácilmente su culpa por el accidente.
- Conocía a los ocupantes del otro vehículo.
- No quiso proporcionar una declaración jurada.
- Fue difícil contactar con él; no cooperó.
- El accidente ocurrió poco después de la fecha de entrada en vigor de la póliza.
- Parecía ser "prudente en los siniestros".

D. Características del daño

- El daño consistía sólo en torceduras.
- No hay evidencia objetiva de daño.
- El informe de la policía mostraba que no se había producido daño.
- El demandante rehusa someterse al "IME".
- No se dio un tratamiento de urgencia para el daño.
- El tratamiento (no urgente) fue demorado (aplazado).
- El primer tratamiento (no urgente) fue por un "DC".
- La inspección arrojó dudas sobre el daño.
- Los daños eran inconsistentes con el informe policial.
- El "IME" sugiere que el daño no tenía relación con el accidente.
- Se trata de un daño inusual para ese accidente de coche.
- Hay evidencia de una causa alternativa del daño.

E. Características del tratamiento

- gran número de visitas al masajista.
- El "DC" proporcionó 3 o más recetas médicas en la mayoría de visitas.
- Gran número de visitas al terapeuta físico.
- Registro MRI o CT pero no cargos hospitalarios por haber estado interno.
- Empleo de un elevado volumen de "usos médicos".

- Intervalos significativos a lo largo del tratamiento.
- El tratamiento fue exageradamente prolongado (más de 6 meses).
- El "IME" cuestionó la extensión del tratamiento.
- La auditoría médica levantó cuestiones sobre los cargos.

F. Características de la pérdida de salario

- El demandante trabajaba para él o era miembro de una familia.
- El salario como empleado difería de la pérdida de salario reclamada.
- El demandante empezó recientemente a trabajar.
- Empresario (patrón) desconocido o difícil de encontrar.
- La declaración sobre pérdida de salario parece no oficial (espontánea).
- Larga incapacidad para trabajar por un daño menor.

1.2 Relación de indicadores potenciales de fraude utilizados por la Cátedra de Gestión de Riesgos de la École des Hautes Études Commerciales de la Universidad de Montreal.

A. Variables relacionadas con el accidente y daño

- Existencia de informe de la policía.
- Coste excesivo de un daño menor.
- Existencia de un daño anterior no relacionado con la pérdida o inconsistente con los daños declarados sobre el siniestro.
- El coche se declara robado y es localizado poco después con daños importantes.

B. Variables relacionadas con el vehículo

- Modelo nuevo y caro.
- El asegurado no puede asumir los costes de mantenimiento del vehículo.
- El vehículo es antiguo y de bajo valor venal.
- Vehículo robado en un centro comercial.

C. Variables relacionadas con el agente o asegurador

- Justo antes del accidente el asegurado comprobó su cobertura con el agente.
- El agente nunca ha visto el vehículo asegurado.

D. Variables relacionadas con las características financieras

- El asegurado posee dificultades financieras personales o de su negocio.
- La ocupación del asegurado no justifica que tenga un vehículo de ese valor o sus accesorios.

E. Variables relacionadas con la existencia de acuerdos tras el accidente

- El asegurado prefiere una compensación monetaria en lugar de la reparación del vehículo.
- El asegurado acepta una pequeña indemnización en lugar de seguir todos los trámites ligados a la reclamación.

F. Variables relacionadas con el demandante/asegurado

- El asegurado insiste en un acuerdo rápido.
- El asegurado está extraordinariamente familiarizado con el proceso de tramitación y la jerga empleada en seguros y reparaciones.
- El asegurado está dispuesto a ir a las oficinas de la compañía para recibir su pago.
- El asegurado acepta rápidamente su culpa en el accidente.
- El accidente tiene lugar poco después de que el vehículo fuera matriculado y asegurado o en los meses justo anteriores al final de la cobertura.
- El asegurado, acogiéndose al Acto 68, rehusa dar su consentimiento para investigaciones del siniestro.
- Elevado número de facturas reclamadas en relación a daños corporales (taxis, alquiler de vehículos, masajistas,...).
- Pérdida u olvido de facturas o pruebas de pago.
- Historial de fraude del asegurado: existencia de fraudes anteriores.

G. Variables relacionadas con la propiedad del vehículo

- No se conocen los antecedentes del vehículo.

H. Otros indicadores

- El siniestro es declarado por una tercera persona.
- No disponibilidad de documentación referente a la estimación y reparación de los daños.
- Todos los vehículos dañados han sido enviados al mismo taller.
- El perito no puede examinar el vehículo.
- El vehículo fue reparado antes de ser examinado por el perito.
- Los testigos aportan versiones contradictorias en relación a las circunstancias del siniestro.
- El asegurado niega las versiones de los testigos.
- Accidente relacionado con un único vehículo.
- Accidente relacionado con una tercera parte no identificada.
- Vehículo con un historial elevado de averías o fuera de garantía.
- Vehículo alquilado con elevado kilometraje.
- Llave de contacto no manipulada tras la recuperación del vehículo.
- No hay signos de haberlo forzado.
- No paga el crédito concedido para la compra del vehículo (o se retrasa en el pago).
- Vehículo comprado al contado.
- El asegurado está en paro; trabaja en una empresa con problemas de solvencia o vive en una región económicamente pobre.

- Vehículo encontrado por el asegurado.
- Varios tipos de cobertura; varias pólizas; pérdidas sobre-aseguradas.
- Prima pagada personalmente y al contado.
- Problemas con la localización del asegurado (dirección falsa, teléfono, etc...).
- El demandante es muy agresivo y amenaza con llamar a un abogado.
- El demandante rehusa responder preguntas sobre el accidente.
- Durante la investigación, el asegurado está nervioso y parece confuso.
- Muchos siniestros anteriores.
- Coche comprado con matrícula de otra provincia.
- La documentación del vehículo contiene todavía los datos del anterior propietario.

ANEXO 2

A.2.1 Estimación del Modelo Lógit Multinomial

La estimación de un modelo logit multinomial suele realizarse usando el método de Máxima Verosimilitud.

La definición del logaritmo neperiano de la función de verosimilitud atiende a la siguiente expresión:

$$\ln L = \sum_i \sum_{j=0}^m Y_{ij} \ln P_{i(j)}, \quad (1.1)$$

donde, para cada individuo i ($i=1, \dots, N$), $Y_{ij}=1$ si la alternativa j es elegida, y 0 si no lo es; ello para las $m+1$ posibles alternativas. De esta forma, para cada i , una y sólo una de las Y_{ij} toma el valor uno, para el resto el valor es cero. Igualmente se define Y_i como la variable que toma valores entre 0 y m que indica la categoría finalmente elegida.

La probabilidad de que el individuo i elija la alternativa j , $\text{Prob}(Y_i=j)$ o bien $P_{i(j)}$, definida en el Capítulo 4, es:

$$P(Y_{ij} = 1) = \frac{e^{\beta'_j X_i}}{\sum_{r=0}^m e^{\beta'_r X_i}} \quad \forall j = 0, 1, \dots, m ; i = 1, \dots, N. \quad (1.2)$$

En esta última expresión, X_i indica el vector de características explicativas del individuo i -ésimo y β_r corresponde al vector de parámetros desconocidos para la alternativa r .

Las primeras y segundas derivadas del logaritmo de la función de verosimilitud son:

$$\frac{\partial \ln L}{\partial \beta_j} = \sum_i (Y_{ij} - P_{i(j)}) X_i \quad ; \forall j=1, \dots, m \quad (1.3)$$

$$\frac{\partial^2 \ln L}{\partial \beta_j \partial \beta_k} = \begin{cases} \sum_i -P_{i(j)}(1 - P_{i(j)}) X_i X_i' & \text{si } j = k \\ \sum_i (P_{i(j)} P_{i(k)}) X_i X_i' & \text{si } j \neq k \end{cases} \quad (1.4)$$

El estimador Máximo Verosímil puede obtenerse por cualquiera de los métodos usuales de iteración.

En el Método de Newton, el proceso iterativo es:

$$\hat{\beta}_{t+1} = \hat{\beta}_t - H_t g_t \quad (1.5)$$

donde t indica el número de iteración, H_t es la inversa de la Hessiana evaluada en las estimaciones obtenidas en la iteración t (define un estimador de la matriz de covarianzas asintótica de los coeficientes estimados) y g_t es el gradiente estimado en la t -ésima iteración.

A.2.2 Estimación del Modelo Lógit Multinomial bajo la existencia de estratificación endógena en la muestra.

Sea N el número de individuos que componen la muestra, teniendo en cuenta un proceso de muestreo basado en la elección.

Podemos denotar la alternativa y el vector de variables exógenas observadas para el i -ésimo individuo j_i y x_i , respectivamente.

La función de Verosimilitud asociada al “muestreo basado en la elección” viene dada por:

$$L = \prod_{i=1}^N P_{i(j)} f(x_i) Q(j_i | \beta)^{-1} H(j_i), \quad (2.1)$$

donde, $Q(j_i | \beta)$ recoge la proporción de individuos en la población que eligen la alternativa j y $H(j_i)$ recoge la proporción análoga en la muestra estratificada. Ver

Amemiya (1985) para una explicación. Este autor introduce $f(x_i)$, la densidad conjunta de las variables explicativas.

El estimador Máximo Verosímil del vector de parámetros β propuesto por Manski y Lerman (1977) será el que maximice la expresión:

$$S_n = \sum_{i=1}^n w(j_i) \log P_{i(j)} \quad (2.2)$$

donde $w(j) = Q_0(j) / H(j)$ y $Q_0(j)$ indica la verdadera proporción de individuos de la población que eligen la alternativa j . En este caso, este valor se considera conocido.

Diferenciando con respecto a β :

$$\frac{\partial S_n}{\partial \beta} = \sum_{i=1}^n w(j_i) \frac{1}{P_{i(j)}} \frac{\partial P_{i(j)}}{\partial \beta} \quad (2.3)$$

e igualando a cero se obtienen las ecuaciones normales.

Diferenciando (2.3) con respecto a β' se obtiene la matriz a utilizar en el proceso de estimación iterativo.

$$\frac{\partial^2 S_n}{\partial \beta \partial \beta'} = - \sum_{i=1}^n w(j_i) \frac{1}{P_{i(j)}^2} \frac{\partial P_{i(j)}}{\partial \beta} \frac{\partial P_{i(j)}}{\partial \beta'} \quad (2.4)$$

$$+ \sum_{i=1}^n w(j_i) \frac{1}{P_{i(j)}} \frac{\partial^2 P_{i(j)}}{\partial \beta \partial \beta'}$$

El estimador del vector de parámetros obtenido mediante este método es consistente y asintóticamente normal (Amemiya, 1985).

ANEXO 3

La muestra disponible contiene información de 1995 reclamaciones para un total de 88 variables. La definición de las mismas, teniendo en cuenta su agrupación en nueve bloques identificadores, es la siguiente (presentamos, entre paréntesis, el nombre dado a cada variable):

A. Variables relacionadas directamente con la identificación del siniestro

- "Número de expediente" (*expedte*).
- "Antigüedad del vehículo" (*aavehic*).
- "Año de ocurrencia del siniestro" (*aaoc*).
- "Mes de ocurrencia del siniestro" (*mmoc*).
- "Día de ocurrencia del siniestro" (*ddoc*).
- "Número de expedientes por daños propios abiertos en el siniestro" (*numdp*).
- "Número de expedientes por responsabilidad civil abiertos en el siniestro" (*numrc*).
- "Número de expedientes por recobros abiertos en el siniestro" (*numre*).
- "Número de expedientes por reclamación de daños abiertos en siniestro" (*numrd*).
- "Número de expedientes por lesiones abiertos en el siniestro" (*numle*).
- "Número de siniestros asociados a la póliza" (*numstr*).
- "Culpa del siniestro" (*culpa*).

B. Variables relacionadas con la póliza

- "Código del ramo (coberturas)" (*ramo*).
- "Uso del vehículo" (*uso*).
- "Agente de la compañía (código)" (*agent*).
- "Situación de la póliza en el momento del siniestro" (*sitpol*).
- "Fecha de vencimiento del recibo" (*vctre*).
- "Año de efecto de la póliza" (*aaefe*).
- "Mes de efecto de la póliza" (*mmefe*).
- "Día de efecto de la póliza" (*ddefe*).
- "Hora de efecto de la póliza" (*hrefe*).
- "Minuto de efecto de la póliza" (*miefe*).
- "Año de emisión de la póliza" (*aemis*).
- "Mes de emisión de la póliza" (*memis*).
- "Día de emisión de la póliza" (*demis*).
- "Hora de emisión de la póliza" (*hemis*).
- "Minuto de emisión de la póliza" (*miemis*).

- "Suplementos que ha tenido la póliza desde su emisión" (*suple*).
- "Cantidad de franquicia contratada" (*franq*).
- "Cantidad en accesorios contratada" (*acces*).
- "Cobertura en responsabilidad civil contratada" (*rc*).
- "Forma de pago de la póliza" (*fopgo*).
- "Código del vehículo" (*coveh*).
- "Color del vehículo asegurado" (*colveh*).

C. Variables relacionadas con el siniestro

- "Hora del siniestro" (*hstro*).
- "Minuto del siniestro" (*mstro*).
- "Provincia de ocurrencia del siniestro" (*proac*).
- "Intervino autoridad (tráfico, policía)" (*autdad*).
- "Hubo testigos" (*test*).
- "Naturaleza del siniestro (urbana o carretera)" (*nasin*).
- "Relato del accidente por el asegurado" (*relatoa*).
- "Daños al vehículo asegurado" (*danosa*).
- "Código taller (del asegurado)" (*cotaa*).
- "Código postal taller asegurado" (*cptaa*).
- "Año de comunicación del siniestro" (*acomu*).
- "Mes de comunicación del siniestro" (*mcomu*).
- "Día de comunicación del siniestro" (*dcomu*).
- "Garantía afectada (resultado)" (*gar*).

D. Variables relacionadas con el asegurado-tomador del seguro

- "Código postal del asegurado" (*cpa*).
- "Sexo del asegurado" (*sexa*).
- "Estado civil del asegurado" (*eca*).
- "Año de nacimiento del asegurado" (*anaa*).
- "Mes de nacimiento del asegurado" (*mnaa*).
- "Día de nacimiento del asegurado" (*dnaa*).
- "Año de carnet del asegurado" (*acara*).
- "Mes de carnet del asegurado" (*mcara*).

E. Variables relacionadas con el propietario del vehículo asegurado

- "Código postal del propietario" (*cpp*).
- "Sexo del propietario" (*sexp*).
- "Estado civil del propietario" (*ecp*).
- "Año de nacimiento del propietario" (*anap*).
- "Mes de nacimiento del propietario" (*mnap*).
- "Día de nacimiento del propietario" (*dnap*).
- "Año de carnet del propietario" (*acarp*).
- "Mes de carnet del propietario" (*mcarp*).

F. Variables relacionadas con el conductor del vehículo asegurado

- "Código postal del conductor" (*cpc*).
- "Sexo del conductor" (*sexc*).
- "Estado civil del conductor" (*ecc*).
- "Año de nacimiento del conductor" (*anac*).
- "Mes de nacimiento del conductor" (*mnac*).
- "Día de nacimiento del conductor" (*dnac*).
- "Año de carnet del conductor" (*acac*).
- "Mes de carnet del conductor" (*mcac*).

G. Variables relacionadas con el conductor contrario

- "Código de la compañía contraria" (*ciact*).
- "Marca del vehículo contrario" (*marct*).
- "Identificador apellidos parte asegurada y parte contraria" (*apellct*).
- "Código postal del contrario" (*cpct*).
- "Sexo del conductor contrario" (*sexct*).
- "Estado civil del conductor contrario" (*ecct*).
- "Año de nacimiento del contrario" (*anact*).
- "Mes de nacimiento del contrario" (*mnact*).
- "Día de nacimiento del contrario" (*dnact*).
- "Año de carnet del contrario" (*acact*).
- "Mes de carnet del contrario" (*mcact*).
- "Color vehículo contrario" (*colct*).
- "Daños vehículo contrario" (*danct*).
- "Taller vehículo contrario" (*tallerct*).

H. Variable identificadora del tipo de fraude existente

- "Clase de fraude" (*subfrau*).

I. variable que mide el coste final del siniestro

- "Coste del siniestro ("-", recobrado; "+", pagado)" (*costfrau*).