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.

número de votos a favor de la sospecha de fraude en un siniestro, seleccionando un conjunto de indicadores que pueden incidir en la fundamentación de la misma.

Las variables explicativas introducidas en ambos modelos vienen definidas por los diferentes aspectos y circunstancias de un siniestro. En concreto recogen información relativa al accidente, al demandante, al conductor asegurado, al daño, al tratamiento médico y a la pérdida salarial experimentada[31]. En relación al accidente, las variables consideradas hacen referencia, entre otras cosas, a la presencia o no de informe policial y/o de testigos, a la existencia de un relato del siniestro no plausible o al hecho de que el demandante posea un vehículo antiguo y de escaso valor. Las características del demandante revelan si reside en una ciudad de elevado riesgo, si en las declaraciones que realiza después del siniestro reitera un "exceso de prudencia" y si tiene un historial de accidentes anteriores. Para el conductor asegurado, Weisberg y Derrig únicamente tienen en cuenta la facilidad de cooperación con la entidad, mientras que, en relación al daño producido, el número de indicadores considerado es más elevado (presencia de una daño inusual para el tipo de accidente producido, negativa del demandante a someterse a un examen médico, daños inconsistentes con el informe policial,...). El tratamiento médico derivado del siniestro es igualmente analizado (elevado número de visitas, desmesurado número de prescripciones médicas,...). Por último, los modelos incorporan información relativa a la influencia del siniestro en la capacidad laboral del demandante debido a las repercusiones que puede tener la declaración de larga incapacidad.

La validación de los modelos se realiza a partir del análisis del coeficiente de determinación (R^2) o bondad del ajuste obtenida. No se presenta el poder explicativo de cada una de las variables (en términos de significación individual) ni los resultados de los contrastes de significación global de cada modelo. El modelo que tiene como variable dependiente el índice de sospecha de fraude (de 0 a 10) para los tramitadores presenta una bondad del ajuste del 65%; este coeficiente disminuye al 56% en el modelo con variable dependiente el índice de sospecha de fraude para los investigadores. En el modelo con variable dependiente el número de votos a favor de la presencia de fraude, la bondad del ajuste es del 46%.

Se produce una notable variación en relación a los indicadores de fraude incluidos en la especificación de cada modelo, hecho que según los autores viene provocado por la diferente perspectiva que subyace en la definición de la variable dependiente para tramitadores e

[31] Se dispone de un total de 65 indicadores de fraude relativos a los 127 siniestros (dichos indicadores aparecen definidos en el Anexo 1). Los modelos que presentan Weisberg y Derrig se basan en la utilización de únicamente 10 indicadores debido a consideraciones prácticas asociadas a la modelización.

investigadores. Ello se hace evidente al considerar la clasificación, por tipos de fraude, establecida por ambos en relación a los 127 siniestros estudiados: mientras que los tramitadores clasifican un 39.4% de los siniestros como legales, un 38.6% como *build-up*, un 15.7% como *opportunistic fraud* y un 6.3% como *planned fraud*, los investigadores del I.F.B. clasifican un 44.9% como legales, un 2.4% como *build-up*, un 40.9% como *opportunistic fraud* y un 8.7% como *planned fraud* (el 3.1% restante lo engloban en la categoría genérica de *otros*). La diferencia más notable, tal y como puede observarse, se encuentra en la clasificación del fraude como *build-up* o como *opportunistic* ya que la dificultad existente en muchos casos para diferenciar ambos conceptos puede derivar en la concentración de los casos en un único tipo.

El pequeño tamaño de la muestra y el hecho de que no sea aleatoria aparecen como dos de las principales razones para justificar la obtención de unos resultados no satisfactorios en términos de calidad del ajuste. En relación a la segunda razón cabe destacar que el hecho de trabajar con una muestra estratificada implica realizar una ponderación de las observaciones, siendo a veces complicado establecer una medida adecuada para la corrección.

El estudio realizado por Weisberg y Derrig muestra también que cuanto mayor es el número de indicadores considerados más fuerte es la sospecha de fraude y más elevada es la frecuencia de votos a favor de la presencia del mismo. Sin embargo, la consideración de todos ellos en la tramitación de siniestros resultaría ineficiente, fundamentalmente por dos razones. En primer lugar sería demasiado costoso para la compañía recoger de forma habitual toda la información y, en segundo lugar, se obviaría el hecho de que algún indicador (o combinación de indicadores) puede tener especial relevancia en la explicación de un determinado tipo de fraude. Notemos que estas dos conclusiones apoyan la necesidad de aplicar tratamientos econométricos adecuados para realizar una correcta selección de las variables explicativas e identificadoras del fraude.

En 1995 y en 1996 Weisberg y Derrig modelizan la sospecha de fraude, esta vez, utilizando una nueva base de datos, con siniestros ocurridos en 1993.

Sin embargo, estos modelos (de características análogas a los utilizados en estudios anteriores) continúan estando fundamentados en una muestra de siniestros cuyos expedientes ya han sido cerrados por las compañías. La información disponible hace referencia no sólo a las circunstancias fundamentales que rodean la ocurrencia del accidente sino también a la tramitación del mismo por la entidad.

La elaboración de una gran base de datos, (conocida como la *"Detailed Claim Database-D.C.D."*) regulada legalmente, con información de todos los siniestros cerrados en o a partir del 1 de enero de 1994 por las entidades aseguradoras de Massachusetts, supone para los autores mencionados la posibilidad de continuar con sus estudios teniendo a su disposición un elevado número de casos. Ello supone un importante avance de cara a validar el poder explicativo de los principales indicadores de fraude tradicionalmente utilizados. Las compañías están obligadas a facilitar toda la información relativa al siniestro y, por tanto, las conclusiones obtenidas podrán ser generalizadas al mercado asegurador global. A modo de referencia cabe señalar que en julio de 1997 la *D.C.D.* contenía información para un total de 514583 siniestros de autos.

Sin embargo, el valor potencial de este nuevo sistema de recogida de información parece ligado también a otro aspecto: la posibilidad de analizar la efectividad en coste de un mecanismo automático de control de fraude.

En su artículo más reciente "A.I.B.-P.I.P. Claim Screening Experiment Interim Report. Understanding and Improving the Claim Investigation Process" (1997), Derrig y Weisberg presentan los primeros resultados derivados de la creación de un sistema inteligente de detección de fraudes que ayude a mejorar el proceso de investigación de los siniestros.

Si hasta ahora todos sus estudios estaban relacionados con el tratamiento de expedientes de siniestros ya cerrados, el objetivo perseguido bajo lo que los autores denominan el *"Claim Screening Experiment, C.S.E."* es modelizar la sospecha existente de fraude en diferentes etapas de la "vida" del siniestro. Una vez acaecido el accidente y declarado a la compañía, la información relativa a diferentes aspectos del mismo (fundamentalmente relacionada con el tratamiento médico derivado) puede llegar a la entidad cuando ha transcurrido ya un determinado periodo de tiempo. En base a ello, resulta adecuado generar índices de sospecha que alerten a los tramitadores sobre la conveniencia o no de investigar un siniestro a medida que se va recibiendo información. Normalmente, durante los 30 primeros días se conoce como ha sido el accidente, quién es el demandante y el asegurado. Posteriormente, en los dos meses siguientes, se tienen datos sobre la naturaleza de las lesiones, del abogado del lesionado y de la fase inicial de gestión del siniestro. En el periodo que resta hasta llegar al medio año, se recibe documentación más extensa sobre el tratamiento médico de las lesiones.

Derrig y Weisberg presentan en su estudio una triple clasificación para los siniestros con daños personales (atendiendo a su tramitación). De esta forma diferencian entre:

1. Siniestros con tramitación rápida (*"express"*). Son aquéllos en los que no hay ninguna prueba aparente de sospecha de fraude y las cuantías reclamadas no son elevadas. Por ello, los mismos pueden ser gestionados con rapidez.

2. Siniestros con tramitación no activada (*"duds"*). Para éstos hay una declaración del daño pero no se puede realizar un seguimiento del accidente todavía. Son expedientes que quedan archivados sin que la compañía realice ningún pago por el momento.

3. Siniestros con tramitación prolongada (*"target"*). Son los que despiertan sospecha de fraude y suelen estar caracterizados por elevadas cuantías demandadas (normalmente se incrementan durante el proceso debido, por ejemplo, al alargamiento del tratamiento médico aplicado).

El porcentaje más elevado de siniestros se encuentra en la tercera categoría (60%), según los resultados del *C.S.E.* La utilización de este sistema de detección puede ayudar a los tramitadores a decidirse sobre la conveniencia o no de investigar un siniestro (diferenciando incluso sobre la necesidad de realizar una investigación ordinaria o especializada).

Los autores modelizan la sospecha de fraude utilizando la técnica de regresión lineal, como ya venían haciendo, pero incorporando ciertas interacciones entre las variables.

Los resultados obtenidos de la aplicación, puesta en práctica en cuatro compañías, permiten señalar que en determinados momentos de la tramitación estaría justificada una mayor investigación. Además, y con carácter preliminar, confirman que existe una relación positiva entre el índice de sospecha y la ejecución de investigaciones especiales (*S.I.U.s*,....). Sin embargo, en relación a la efectividad del sistema, las conclusiones ponen de manifiesto un hecho importante. El buen funcionamiento del sistema implica la necesidad de incorporarlo automáticamente en las compañías puesto que los tramitadores se oponen a realizar manualmente la recogida de información requerida para implementar completamente el sistema.

El análisis coste-beneficio del diseño del sistema *C.S.E.* y de la investigación de siniestros derivada de su aplicación aparece como uno de los principales objetivos de estudio a corto plazo. La incorporación de los siniestros experimentados con el *C.S.E.*, que aún permanecen abiertos, al *D.C.D.*, permitirá comparar los costes iniciales y finales de los mismos, considerando además los asociados a la investigación.

El diseño de sistemas inteligentes que ayuden a las entidades a realizar una correcta recogida de datos puede contribuir decisivamente al buen funcionamiento de los sistemas automáticos de control y detección de fraude.

- **Aproximación al tratamiento del fraude mediante análisis cluster**

La aplicación de técnicas multivariantes cluster en el campo asegurador ha ido ganando importancia a lo largo de los últimos años (Lemaire, 1990; Derrig y Ostaszewski, 1994a y 1994b; Cummins y Derrig, 1993, 1997) apareciendo como una metodología susceptible de utilización en la modelización de la incertidumbre.

Conceptuado como un análisis enfocado principalmente a la definición de grupos homogéneos (clasificación de individuos en grupos atendiendo a su similitud en determinadas características), su aplicación en la clasificación de riesgos goza de especial importancia y, desde este punto de vista, ha sido utilizado en diferentes estudios propios del mercado de seguros.

Derrig y Ostaszewski (1994b) seleccionan la técnica cluster para realizar una clasificación de siniestros en términos de sospecha de fraude. Para ello utilizan la misma base de datos formada por 127 siniestros que había sido tratada por Weisberg y Derrig (1993) y que ya ha sido comentada en páginas anteriores. Su objetivo es plantear una solución al problema básico con el que se enfrenta el tramitador de siniestros (o en su caso, el investigador): clasificar o no el siniestro como fraudulento. La técnica utilizada sustituye el usual "cero-uno"[32] por una función de medida cuyos valores oscilarán precisamente entre dichos extremos. Si la valoración obtenida tras la aplicación del análisis es 0 ó 1, la asignación a una de las categorías es clara. Si el valor se encuentra a lo largo del intervalo, la técnica generará una determinada clasificación en relación a la sospecha existente de fraude.

Los autores desarrollan un doble estudio. Por un lado, realizan una agrupación de siniestros teniendo en cuenta los niveles existentes de sospecha y, por otro, presentan una clasificación atendiendo a la valoración de fraude realizada, es decir, considerando los diferentes tipos de fraude posibles.

En el primero de los casos, la sospecha de fraude es cuantificada mediante la creación de una escala de cero a diez (ya había sido utilizada en estudios anteriores). La respuestas son clasificadas en cinco clusters iniciales teniendo en cuenta los niveles de sospecha (según la

[32] No fraude-fraude.

valoración de los tramitadores de siniestros): ausencia de sospecha (0), sospecha débil (1-3), sospecha moderada (4-6), sospecha fuerte (7-9) y sospecha cierta (10). La agrupación posterior se realiza atendiendo a la información suplementaria aportada por otras dos medidas: el nivel de sospecha de los investigadores[33] (en base a la misma escala que los tramitadores) y el número de "votos" a favor de la presencia de fraude (escalado de cero a tres)[34]. El algoritmo utilizado provoca una clasificación final de siniestros en la que las diferencias entre los centros de los diferentes clusters pone de manifiesto una divergencia entre el enfoque utilizado por tramitadores e investigadores. No obstante, dicha diferencia se suaviza al considerar sólo dos clusters: el asociado a la "no percepción de fraude" y el que incluye "percepción moderada de fraude" más "percepción fuerte de la existencia de fraude".

En el segundo caso, la clasificación de siniestros se realiza atendiendo a la tipología de fraude existente. Las categorías consideradas son: "*planned fraud*", "*opportunistic fraud*", "*build-up*" y "*no fraud/build-up*". Cada siniestro es codificado según el nivel de sospecha existente (en una escala de cero a diez) en seis componentes del mismo: el accidente, el daño, el asegurado, el demandante, el tratamiento médico y la pérdida de salario. Los resultados obtenidos muestran una agrupación en cinco clusters[35]. El centro de cada uno de ellos indica, en base al nivel de sospecha existente para cada una de las seis componentes, la posible presencia de uno u otro tipo de fraude[36].

Las conclusiones obtenidas por Derrig y Ostaszewski ponen de manifiesto la susceptibilidad de aplicación del análisis cluster al tratamiento del fraude. En su opinión, su principal conclusión es que hay que otorgar más peso a las componentes asociadas al propio siniestro que a la percepción subjetiva de los tramitadores (y/o de los investigadores).

- **Aproximación al tratamiento del fraude mediante redes neuronales**

Partiendo de la base de datos utilizada por Weisberg y Derrig (1993) y por Derrig y Ostaszewski (1994b), Brockett, Xia y Derrig (1995) presentan una aproximación al estudio del fraude en el seguro del automóvil mediante la aplicación de redes neuronales.

---

[33] Cabe recordar que los 127 siniestros fueron codificados paralelamente por tramitadores de siniestros y por investigadores del I.F.B. de Massachusetts.

[34] La codificación se realizó por dos tramitadores y dos investigadores (el hecho de que en la escala no aparezca el máximo, es decir, cuatro, se debe a que ninguno de los siniestros fue clasificado como fraudulento por todos los codificadores).

[35] La categoría de build-up aparece dividida en dos partes en función del nivel de sospecha existente.

[36] A modo de ejemplo, el cluster con centro (0,0,0,0,0,0) recoge siniestros sin sospecha de fraude. En el cluster con centro (7,8,7,8,8,0) aparecen siniestros con alta sospecha de "*planned fraud*" o "*fraude planeado*".

El objetivo es crear un sistema de detección de fraude teniendo en cuenta el vector de características de cada uno de los siniestros. Para ello, se seleccionan los indicadores de fraude (un total de 65 variables) que ya habían sido utilizados por Weisberg y Derrig (1993), y que como sabemos están relacionados con el accidente, el demandante, el asegurado, el daño, el tratamiento médico y la pérdida salarial derivada.

Seleccionando 77 siniestros del total de la muestra (formada por 127 siniestros) para realizar la modelización y dejando 50 siniestros para llevar a cabo predicción ex-post[37], la aplicación de redes supone la representación de cada uno de los casos mediante un vector de atributos compuesto, en este caso, por los 65 indicadores de fraude utilizados. El método parte de dos premisas básicas: en primer lugar, modelos de siniestros parecidos tendrán niveles de sospecha similares y, en segundo lugar, cada uno de los indicadores considerados tendrá igual importancia en la explicación de la existencia de fraude.

Así, teniendo en cuenta un input formado por 77 vectores (uno para cada siniestro), el output resultante se centra en la clasificación del suceso en base a cuatro categorías: "siniestro válido", "siniestro con débil sospecha de fraude", "siniestro con sospecha de fraude moderada" y "siniestro con una elevada sospecha de fraude". El hecho de que los siniestros con vectores de indicadores parecidos (y, por tanto, con pequeñas "distancias" entre ellos) queden más o menos juntos y alejados del resto, dentro del "mapa de representación", permite crear zonas ("regiones de decisión") que se identificarán con las cuatro posibilidades (outputs) comentados en relación a la sospecha de fraude.

Introduciendo en el proceso las valoraciones subjetivas de sospecha de fraude de tramitadores e investigadores (ya consideradas en estudios anteriores[38]), los resultados obtenidos perfeccionan, según el criterio de los autores, la clasificación de siniestros obtenida a partir del uso, únicamente, de una variable dependiente asociada a la sospecha de fraude y de determinados indicadores. No obstante, problemas relacionados con la propia metodología y, en su caso, con los datos, pueden limitar las conclusiones obtenidas del estudio. Así, el hecho de que todos los indicadores sean igualmente ponderados implica no considerar la posible existencia de variables más significativas que otras.

Los resultados del análisis y su comparación con la clasificación de siniestros presentada por los tramitadores y los investigadores, reflejan una elevada calidad en términos de predicción *ex-post*. Por lo que respecta a la predicción *ex-ante*, las conclusiones no son tan favorables.

[37] Se utiliza una parte de la muestra para validar los resultados obtenidos.
[38] Weisberg y Derrig (1993), Derrig y Ostaszewski (1994b).

Los autores remiten a una ampliación de la muestra, a una mejora de la información o a una eliminación de los indicadores no significativos como posibles soluciones a la búsqueda de resultados más positivos.

- **Aproximación al tratamiento del fraude mediante el uso de modelos de elección probabilística**

Los modelos de elección discreta ofrecen la posibilidad de cuantificar la probabilidad de aparición de comportamientos fraudulentos en los siniestros cuando la variable dependiente ha sido adecuadamente categorizada para recoger la presencia/ausencia de fraude.

La aplicación de modelos lógit y próbit simples queda patente en la literatura existente sobre el tema. Si bien en Ayuso (1995) y en Artís, Ayuso y Guillén (1998) es posible encontrar una aplicación de modelos lógit sencillos al estudio de la dicotomía planteada, en Belhadji y Dionne (1997) el estudio realizado parte de la aplicación de un modelo próbit.

En el trabajo de investigación "El Fraude en el Seguro del Automóvil" (Ayuso, 1995) se especifica un modelo de regresión logística simple para cuantificar la probabilidad de existencia de fraude. Para ello se utiliza la base de datos facilitada por el Insurance Fraud Bureau de Massachusetts, comentada en páginas anteriores, formada por 127 expedientes de siniestros. Teniendo en cuenta la categorización de fraude presentada por Weisberg y Derrig (1993), el objetivo del estudio se centra en modelizar dos situaciones alternativas:

1. ausencia de fraude *versus* fraude *"a priori"* y,
2. ausencia de fraude *versus* fraude *"a posteriori"*.

En el primero de los casos, bajo la denominación de fraude *"a priori"*, se recogen todos aquellos siniestros clasificados en la base de datos como sospechosos de fraude planeado (siniestro totalmente construido). En el segundo, bajo el concepto de fraude *"a posteriori"*, se consideran todos los casos en los que se sospecha de la existencia de fraude tras la ocurrencia real del accidente (suma de aquéllos en los que existe sospecha de *"opportunistic fraud"* y *"build-up"*).

Las variables explicativas utilizadas son seleccionadas del conjunto de indicadores de fraude presentado por Weisberg y Derrig (1993). Están relacionadas con características del accidente y del vehículo (ausencia de atestado policial, ausencia de testigos,...), del demandante (situación laboral), del daño (no existe evidencia objetiva del mismo) y del tratamiento

médico derivado del siniestro (elevado número de visitas médicas,...). Los resultados obtenidos tras la modelización son satisfactorios, tanto en términos de significación individual y global de los modelos, como en capacidad predictiva (el porcentaje de aciertos en la clasificación es del 87.9% para el modelo que contiene como variable dependiente la sospecha de existencia de fraude *a priori* y del 80.6% para el que posee como dependiente la existencia de fraude *a posteriori*). No debe olvidarse, sin embargo, el escaso número de casos analizados en la muestra a la hora de extrapolar resultados.

En Artís, Ayuso y Guillén (1998) puede encontrarse otra aplicación de modelización logística simple, esta vez, con ponderaciones. En este caso, la información contenida en la misma muestra utilizada en la presente Tesis Doctoral es usada para cuantificar la probabilidad de existencia de fraude *versus* no fraude (no se distingue, por tanto, la existencia de diferentes tipos de actuación fraudulenta al alcance del asegurado).

Siguiendo un proceso muy similar al desarrollado en el trabajo mencionado en el párrafo anterior, Belhadji y Dionne (1997) utilizan un modelo próbit para estimar la probabilidad de existencia de fraude (detectado o sospechado) frente a la de no fraude. El objetivo planteado por ambos autores es doble. Por un lado, diseñar una herramienta de detección que ayude a los tramitadores de siniestros en el estudio del posible comportamiento fraudulento de los asegurados (implementación automática de los resultados derivados del modelo próbit previa selección de los indicadores más significativos de fraude). Por otro, desarrollar un sistema que evalúe la conveniencia o no de investigar los siniestros con elevada sospecha de fraude teniendo en cuenta el enfoque coste-beneficio.

La representatividad de la muestra utilizada por Belhadji y Dionne queda suficientemente justificada: formada por 2068 siniestros (1937 clasificados como no fraudulentos, 113 con sospecha de fraude y 18 con fraude detectado), ha sido diseñada en base a la información facilitada por 18 de las más grandes compañías que operan en el mercado asegurador automovilístico de Quebec (Canadá). Los expedientes han sido seleccionados aleatoriamente por las entidades entre todos los cerrados durante el periodo que va del 1 de abril de 1994 al 31 de marzo de 1995 (la participación de cada compañía ha sido proporcional a su cuota de mercado).

La selección adecuada de los indicadores de fraude tradicionalmente utilizados goza de especial importancia para los autores mencionados. El elevado número de circunstancias posiblemente relacionadas con la existencia de fraude (definen un total de 50 indicadores[39]),

[39] Algunos recopilados de la literatura existente y otros elaborados según el criterio de profesionales.

deriva en la aplicación de un criterio que determine las variables más significativas a la hora de explicar la aparición de comportamientos deshonestos. Aunque plantean la posibilidad de realizar la selección entre indicadores atendiendo al cálculo de las probabilidades condicionales de fraude para cada uno de ellos, la aplicación de un modelo próbit permite determinar un conjunto de 18 variables como estadísticamente significativas. Éstas aparecen relacionadas, entre otras cosas, con aspectos del accidente y el daño (coste excesivo para un daño menor, existencia de un daño anterior no relacionado con el siniestro y declaración del coche como robado siendo encontrado posteriormente con daños importantes), del demandante y/o del asegurado (conoce el proceso de tramitación y la jerga empleada en seguros y reparaciones, acepta rápidamente su culpa, presenta un elevado número de facturas por daños corporales,...), de su situación financiera, de la rapidez con que pretende llegar a un acuerdo con la entidad, del nerviosismo que presenta durante la investigación, etc.

La selección de criterios probabilísticos alternativos (probabilidad estimada a partir de la cual el siniestro es clasificado como fraudulento) permite establecer, a partir de la muestra utilizada, reglas de actuación para las entidades frente al fraude. Si se elige como criterio de decisión un nivel probabilístico elevado, el número de siniestros a investigar será muy bajo, el nivel de precisión del modelo en relación a los casos que la compañía tiene clasificados como fraudulentos será muy elevado pero la entidad dejará de detectar un elevado número de fraudes. Si el criterio probabilístico es bajo ocurrirá lo contrario: el número de siniestros a investigar será elevado, el modelo clasificará como sospechosos un número de siniestros muy superior al observado en la muestra pero, tras su aplicación, aumentará el número de casos detectados por la compañía.

En relación al deseo de realizar un análisis coste-beneficio de la investigación de siniestros con elevada sospecha de fraude, Belhadji y Dionne muestran un estudio preliminar, ante la insuficiencia de mayor riqueza en los datos. La regla de decisión que plantean es lógica: si el coste de llegar a un acuerdo con el asegurado es más bajo que el coste de la investigación no resulta rentable investigar el siniestro; de otro modo, la investigación será realizada. En su aproximación hacen depender el coste de la misma de cuatro factores: la probabilidad de fraude, la experiencia de los investigadores, la formación de éstos y la existencia de unidades especiales de investigación dentro de la compañía. El tratamiento que realizan es muy sencillo y únicamente les permite extraer como conclusión la conveniencia de investigar el siniestro cuando las ganancias derivadas del proceso cubran, al menos, el coste de realizarlo.

Ambos autores enfatizan la necesidad de crear mecanismos automáticos de detección que indiquen a las entidades la probabilidad de existencia de fraude y que señalen la conveniencia de investigar o no los siniestros.

Frente a una aproximación del fraude que considere la dicotomía asociada a su existencia o no existencia (en muchos casos y, como hemos ido comentando, en relación a la sospecha de la misma), los trabajos realizados en el seno del Departamento de Econometría, Estadística y Economía Española de la Universidad del Barcelona, han estado dirigidos, desde el principio, a la modelización del fraude teniendo en cuenta sus diferentes formas de manifestarse. Desde este punto vista, las técnicas econométricas planteadas como susceptibles de aplicación han estado ligadas a modelos logísticos multinomiales y anidados para los que la Tesis que presentamos constituye el máximo exponente.

- **Otras aproximaciones al tratamiento del fraude en el seguro del automóvil**

En otro orden de tratamiento, Cummins y Tennyson (1996) presentan una aproximación al fraude considerándolo como una determinada manifestación de azar moral. En su estudio presentan un enfoque sectorial o agregado utilizando variables explicativas relacionadas con el sistema económico en general. Teniendo en cuenta que el asegurado o el demandante pueden realizar acciones después de la ocurrencia de un siniestro que afecten a la distribución de las pérdidas derivadas, es posible hablar de la presencia de "azar moral *ex-post*". Esta situación irá ligada a aquellos casos en los que el demandante tiene más información que el asegurador sobre la situación posterior al accidente.

Teniendo en cuenta que este comportamiento puede afectar tanto al número de siniestros declarados como a las indemnizaciones reclamadas, los autores realizan la modelización de la presencia de fraude (agregadamente a nivel de estado) midiendo su incidencia en la frecuencia de siniestros declarados por daños corporales. Para ello, llevan a cabo una estimación por mínimos cuadrados de un modelo lineal con variable dependiente el ratio entre la frecuencia de siniestros por daños corporales y la frecuencia por daños materiales[40]. Las variables explicativas están relacionadas con características económicas, demográficas y legales de los estados. Como indicadores de azar moral aparecen diferentes actitudes o valoraciones de los individuos hacia el fraude (permitir al médico o al abogado extender facturas por servicios no prestados, mentir sobre la indemnización reclamada para recuperar la franquicia, mentir sobre las pérdidas para recuperar las primas pagadas, declarar que se han ocasionado daños

[40] En realidad, la variable dependiente es el logaritmo de dicho ratio. La muestra está formada por observaciones recogidas en 29 estados norteamericanos en 1991 y 1992.

corporales en individuos no relacionados con el accidente, participar en cadenas de fraude o aprovechar la existencia del seguro para cubrir un siniestro no relacionado con la póliza).

Adicionalmente, Cummins y Tennyson modelizan la frecuencia de siniestros por daños corporales con mayor probabilidad de estar sujetos a la presencia de azar moral. De esta forma, y considerando trabajos anteriores (Weisberg y Derrig, 1991)[41] utilizan modelos de regresión clásicos para estudiar el comportamiento del porcentaje de siniestros por daños corporales leves (variable dependiente). De nuevo introducen variables demográficas, económicas y legales en la especificación del modelo, teniendo en cuenta, asimismo, variables de actitud relativas a diferentes comportamientos fraudulentos.

A pesar de que en esta segunda modelización la bondad del ajuste (medida por el coeficiente de determinación corregido o ajustado) no presenta valores tan elevados como en la primera, los resultados obtenidos en una y otra ponen de manifiesto dos importantes conclusiones: la existencia de azar moral tiene una incidencia importante en la declaración de siniestros por daños corporales y, además, influye significativamente en la proporción esperada de los mismos asociados a lesiones leves. Todo ello sirve para justificar la influencia del fraude y, de su valoración por parte de la sociedad, en el mercado asegurador.

Una vez analizadas las principales aportaciones realizadas hasta el momento en el estudio del fraude, podemos establecer una serie de consideraciones finales en relación a las mismas.

Los trabajos presentados quedan enmarcados dentro de dos grupos diferenciados: aquéllos de contenido puramente teórico y aquéllos de contenido básicamente aplicado.

En relación a los primeros el enfoque es económico y se centra en determinar conceptualmente la utilidad esperada por el individuo al actuar fraudulentamente. En relación a los segundos, el fundamento teórico utilizado aparece poco desarrollado y se centra en presentar resultados derivados de trabajar, en la mayoría de los casos, con muestras de pequeñas dimensiones o de representatividad cuestionable. En la modelización se incluyen variables dotadas, muchas veces, de una gran subjetividad y para las que la compañía no dispone de información de manera inmediata a la ocurrencia del siniestro. La variable dependiente suele ser la sospecha de fraude al no disponerse de muestras representativas con fraude efectivamente detectado.

[41] Weisberg y Derrig (1991) identifican las lesiones leves (torceduras,...) como las más frecuentemente vinculadas con el fraude en daños corporales.

Todas las aportaciones presentan un enfoque microeconómico a la existencia de comportamientos deshonestos, salvo la de Cummins y Tennyson que enfatiza el estudio del fraude teniendo en cuenta una visión sectorial o agregada.

La modelización que presentaremos a partir del próximo Capítulo pretenderá combinar, de forma adecuada, el marco teórico y el aplicado. El objetivo será estudiar el fraude considerando el comportamiento que sigue el asegurado en la búsqueda de una utilidad máxima. Ello quedará patente en la técnica econométrica seleccionada, para la que se presentará un desarrollo teórico extenso (Capítulo 4). La aplicación empírica se centrará en aplicar los modelos desarrollados a una muestra con información, por primera vez, para fraudes detectados. El tamaño de esta última no es elevado, pero permitirá obtener, tal y como veremos, importantes conclusiones (Capítulos 5, 6 y 7).

4. PRESENTACIÓN DEL MODELO DE ELECCIÓN DE FRAUDE CON MÚLTIPLES ALTERNATIVAS EN EL SEGURO DEL AUTOMÓVIL

4.1 Modelo de elección de fraude: una decisión jerárquica entre alternativas

La aproximación teórica al modelo de elección que recoja el proceso seguido por el asegurado a la hora de actuar o no fraudulentamente puede realizarse atendiendo al diseño de un árbol de decisión. Bajo este planteamiento, la idea principal se centra en el hecho de que, en numerosas ocasiones, la elección entre alternativas no tiene por qué realizarse de forma directa; con frecuencia es posible considerar diferentes etapas en el proceso, que permitirán seleccionar la alternativa final, después de valorar el resto de posibilidades presentes en cada paso. En todo momento, el esquema presentado permitirá describir la elección realizada como el resultado de un proceso de decisión unificado, en el que la selección entre alternativas se llevará a cabo, bien bajo el criterio de maximizar la función de utilidad aleatoria, bien bajo el de minimizar los costes de elección (para nosotros, generalmente, costes de oportunidad en tiempo[1]).

La elección entre no cometer fraude o cometerlo (atendiendo a la existencia de diferentes formas de defraudar) puede recogerse mediante el siguiente árbol de decisión:

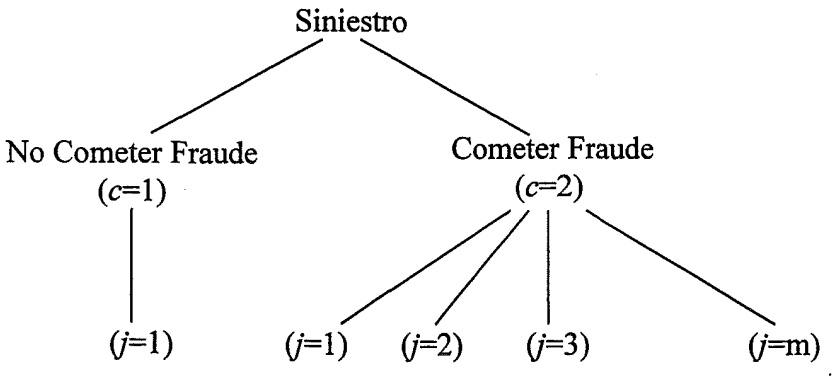

Figura 6

[1] En Koujianou (1995), se habla, por ejemplo, de la minimización de los costes de búsqueda en la elección de un determinado tipo de vehículo.

con c=1,2,...C , alternativas iniciales y, j=1,2,...,m_c , alternativas finales indexadas en cada opción previa c.

La estimación de un modelo de decisión jerárquico, como el planteado en la Figura 6, puede realizarse atendiendo al criterio de maximización de utilidad aleatoria (McFadden, 1978). Las posibilidades de desarrollar el proceso son fundamentalmente dos.

En primer lugar, el cálculo de las probabilidades asociadas a cada una de las situaciones finales (nivel inferior del árbol) puede realizarse de forma directa, es decir, sin considerar la existencia de niveles intermedios de decisión. Bajo este plantemiento, el uso de un modelo lógit multinomial permitirá realizar adecuadamente el proceso, previa adopción de determinadas hipótesis para los términos de error que aparecerán en la definición de la función de utilidad aleatoria.

En segundo lugar, el uso secuencial de modelos lógit multinomiales ha de permitir realizar el cálculo de los estimadores teniendo en cuenta la existencia de niveles intermedios. La consideración de la posible existencia de correlaciones entre las alternativas planteadas derivará en la formulación de un nuevo modelo, resultado de la aplicación de regresión logística jerárquica (modelos lógit anidados).

Al igual que comentábamos en páginas anteriores, la introducción del criterio de actuación del individuo en base a la maximización de la utilidad implicará considerar la premisa de que, entre varias alternativas, el asegurado siempre elegirá aquélla que le reporte mayor utilidad. Según el modelo de utilidad aleatoria (McFadden, 1978), la utilidad que se deriva del comportamiento del individuo puede descomponerse en dos partes, una determinista (que genera la denominada *utilidad estricta*) y una aleatoria:

$$U_{cj} = V_{cj} + e_{cj} \quad \text{con } c=1,2;\ j=1,2,...m_c. \tag{4.1}$$

La utilidad estricta, V_{cj}, recoge una combinación lineal entre parámetros y variables explicativas propias del individuo y/o de la elección realizada (características del asegurado, del contrario, del siniestro, del vehículo,...). El término aleatorio o de error, e_{cj}, recoge los efectos asociados a variables no consideradas en la parte determinista que pueden influir en la elección, así como las imperfecciones en la percepción de la maximización de utilidad (McFadden, 1978; Maddala, 1983; Greene, 1997).

La presencia de variables no observables supone la adopción de una determinada distribución poblacional para las mismas. La estimación de la probabilidad de que se realice una determinada elección *cj* vendrá dada por:

$$P_{cj} = Prob(U_{cj} > U_{c'j'}) \quad \forall cj \neq c'j'. \tag{4.2}$$

Teniendo en cuenta las consideraciones anteriores en relación a la definición de la función de utilidad y a la existencia de una distribución poblacional para las variables no observadas, la probabilidad anterior puede obtenerse como[2]:

$$P_{cj} = \int_{e_{cj}=-\infty}^{+\infty} F_{cj}(<V_{cj} + e_{cj} - V_{c'j'}>)\, de_{cj}. \tag{4.3}$$

donde, e_{cj} recoge la componente *cj-ésima* del vector $\vec{e}$, formado por los términos aleatorios de cada una de las posibles elecciones $(e_{11}, e_{21}, e_{22}, \ldots, e_{2m})$ y F_{cj} es la derivada marginal de la función de distribución acumulada de $\vec{e}$, $(F(\vec{e}))$, con respecto al argumento (elección realizada).

A modo de ejemplo, la cuantificación de la probabilidad de que el asegurado cometa fraude en beneficio propio (ante tres posibilidades de actuación: no defraudar, defraudar en beneficio propio y defraudar en beneficio de un tercero) supondrá tener en cuenta el siguiente árbol de decisión[3]:

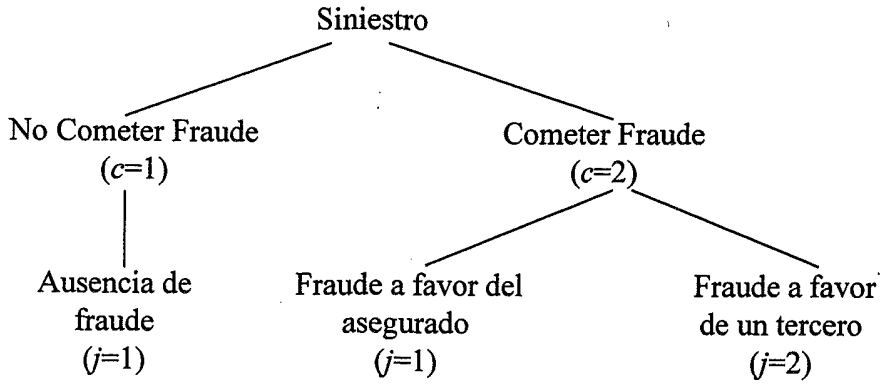

Figura 7

[2] Ver McFadden (1978).

[3] En Artís, Ayuso y Guillén (1997) puede encontrarse el análisis detallado de este árbol de elección.

Atendiendo a la definición de una variable dependiente que recoja la elección final observada de la alternativa, de la forma:

$$Y_{21} = f(U_{21}) = \begin{cases} 1, & \text{si } U_{21} = \text{Max } (U_{11}, U_{21}, U_{22}) \\ 0, & \text{de otra forma} \end{cases} \tag{4.4}$$

y teniendo en cuenta que, de forma análoga, se podrían definir variables dependientes asociadas a la elección de no defraudar y de defraudar en beneficio de un tercero, la modelización de la probabilidad correspondiente supondría determinar,

$$P_{21} = P(U_{21} > U_{11}, U_{21} > U_{22}). \tag{4.5}$$

Dado que $U_{cj} = V_{cj} + e_{cj}$, de forma que $\vec{e} = (e_{11}, e_{21}, e_{22})$, y suponiendo una determinada función de distribución para el vector de residuos ($F(\vec{e})$), tendremos,

$$\begin{aligned} P_{21} &= P(V_{21} + e_{21} > V_{11} + e_{11} \ , \ V_{21} + e_{21} > V_{22} + e_{22}) = \\ &= P(V_{21} + e_{21} - V_{11} > e_{11} \ , \ V_{21} + e_{21} - V_{22} > e_{22}) \ = \\ &= F_{e_{11}, e_{22}} (V_{21} + e_{21} - V_{11} \ , \ V_{21} + e_{21} - V_{22}) \ = \\ &= \int_{e_{21}=-\infty}^{+\infty} \frac{\partial F}{\partial e_{21}} (V_{21} + e_{21} - V_{11} \ , \ e_{21} \ , \ V_{21} + e_{21} - V_{22}) de_{21}. \end{aligned} \tag{4.6}$$

Dejemos el ejemplo y pasemos a continuación a determinar la expresión de la probabilidad de elegir una determinada alternativa, P_{cj}.

4.2 El Modelo Lógit Multinomial

La generación del modelo lógit multinomial supone una particularización de la expresión general presentada para la probabilidad de elegir una determinada opción (4.3), asumiendo que los residuos, e_{cj}, se distribuyen idéntica e independientemente según una distribución *valor extremo tipo I*[4] (McFadden, 1978; Maddala, 1983). De esta forma, la probabilidad de elegir una determinada alternativa *cj* atiende a la expresión,

[4] La función de distribución acumulada para una distribución valor extremo tipo I es $F(e_{cj} < e) = exp(-exp^{-e})$ y la función de densidad probabilística es $f(e_{cj}) = exp(-e_{cj} - exp^{-e_{cj}})$.

$$P_{cj} = P(Y_{cj} = 1) = \int_{-\infty}^{+\infty} \prod_{cj \neq c'j'} F(e_{cj} + V_{cj} - V_{c'j'}).f(e_{cj})de_{cj}\,, \qquad (4.7)$$

expresión que, una vez realizadas las operaciones oportunas (ver Maddala, 1983), puede escribirse como:

$$P_{cj} = \frac{e^{V_{cj}}}{\sum_{b=1}^{C}\sum_{J=1}^{m_b} e^{V_{bJ}}}\,, \qquad (4.8)$$

donde b y J recogen el conjunto total de alternativas intermedias (defraudar o no) y finales (tipo de fraude), respectivamente.

Tomemos dicha expresión y adaptémosla a la situación objeto de estudio.

Como ya comentábamos en páginas anteriores, la función de utilidad estricta recoge una combinación lineal entre parámetros y variables, identificativas estas últimas de determinados atributos o características del individuo y/o de la elección.

La generación del modelo lógit multinomial, estrictamente hablando, implica introducir únicamente como variables explicativas ciertas características del individuo que pueden incidir sobre su elección final. En este sentido, se contrapone al modelo lógit condicional, en el que los atributos considerados son, preferentemente, los asociados a la elección (Greene, 1997).

Bajo esta perspectiva, y realizando la modelización como si de un proceso de decisión único se tratase (estimación de las probabilidades del último nivel del árbol sin considerar la existencia de etapas intermedias), podemos formular la función de utilidad aleatoria de la forma:

$$U_{i(cj)} = V_{i(cj)} + e_{i(cj)} = \beta'_{cj} X_i + e_{i(cj)}\,, \qquad (4.9)$$

donde, *cj* recoge el conjunto total de alternativas finales entre las que ha de elegir el individuo[5], i indica el conjunto de individuos considerados en la muestra, X_i es el vector de características

[5] En el ejemplo presentado en páginas anteriores relacionado con la elección entre no defraudar, cometer fraude en beneficio del asegurado o en beneficio de un tercero, el conjunto de alternativas vendría referenciado por los subíndices *"11"*, *"21"* y *"22"*, modelizándose la decisión del individuo entre tres opciones finales, es decir, sin considerar la posible existencia de un nivel intermedio (éste queda representado por el primer dígito del subíndice, si bien su presencia queda justificada, únicamente, por el deseo de mantener uniformidad en la notación).

asociadas a los mismos y β_{cj} es el vector de parámetros a estimar para cada una de las posibles elecciones.

Atendiendo a esta notación, el modelo lógit multinomial vendrá especificado por:

$$P_{i(cj)} = P(Y_{i(cj)} = 1 \mid X_i) = \frac{e^{V_{i(cj)}}}{\sum_{b=1}^{C} \sum_{J=1}^{m_b} e^{V_{i(bJ)}}} = \frac{e^{\beta'_{cj} X_i}}{\sum_{b=1}^{C} \sum_{J=1}^{m_b} e^{\beta'_{bJ} X_i}} \qquad \forall i = 1,\ldots,N. \qquad (4.10)$$

En esta última expresión N denota el número total de individuos de la muestra disponible.

El árbol decisión presentado en la Figura 6 pone de manifiesto la existencia de un conjunto formado por $m+1$ elecciones finales (podemos, por tanto, reflejar la alternativa seleccionada mediante el subíndice j(=0,1,...,m)[6], donde j=0 refleja la elección de no defraudar). De esta forma, el sistema de ecuaciones a estimar para determinar las probabilidades asociadas a cada alternativa final, teniendo en cuenta la expresión (4.10), será[7]:

$$P(Y_{ij} = 1) = \frac{e^{\beta'_j X_i}}{1 + \sum_{r=1}^{m} e^{\beta'_r X_i}} \qquad \forall j = 1,\ldots,m$$

$$P(Y_{i0} = 1) = \frac{1}{1 + \sum_{r=1}^{m} e^{\beta'_r X_i}}, \qquad (4.11)$$

expresiones obtenidas al imponer la condición de normalización β_0=0.

De esta forma, como puede observarse, el modelo contiene un vector de parámetros (β_r) asociado a cada alternativa (excepto para la primera, debido a la condición de identificación impuesta).

---

[6] El subíndice c ha sido obviado a partir de este momento para simplificar la notación.

[7] El condicionante en X_i ha sido eliminado dado que asumimos que las variables explicativas son deterministas. La expresión obtenida para el lógit multinomial (4.10) podrá escribirse, en base a las consideraciones realizadas, como:

$$P(Y_{ij} = 1) = \frac{e^{\beta'_j X_i}}{\sum_{r=0}^{m} e^{\beta'_r X_i}} \qquad \forall j = 0,1,\ldots,m.$$

La estimación de los parámetros se realiza habitualmente por máxima verosimilitud. La aplicación de un método iterativo (normalmente, el método de Newton) permite obtener valores para los estimadores a partir de la log-verosimilitud del modelo (ver el Anexo 2 del presente trabajo, donde se presenta de forma detallada el proceso de estimación).

La interpretación de los coeficientes no es tan inmediata como en el modelo de regresión lineal clásico. En este último, los parámetros estimados miden la variación esperada que se produce en la variable dependiente cuando la variable explicativa correspondiente aumenta en una unidad[8]. En el modelo lógit, la variación esperada que se produce en la probabilidad de elegir una determinada alternativa al aumentar en una unidad la variable explicativa correspondiente no es constante, sino que depende de los valores que tome dicha variable y el resto de las variables explicativas (Maddala, 1983; Greene, 1997). No obstante, las estimaciones obtenidas para los parámetros son susceptibles de interpretación en términos de dirección en la variación de la probabilidad esperada. Asimismo, fijando el valor de todas las variables explicativas y dejando variar en una unidad únicamente una variable, la exponencial del parámetro estimado se interpreta como la variación esperada que se producirá en el cociente de riesgos.

La propiedad que cumple este modelo en relación a la "Independencia de Alternativas Irrelevantes - *I.I.A*[9].-" (Luce 1959), y que garantiza que el cociente de probabilidades entre la *j-ésima* y la *j'-ésima* elección se mantenga constante independientemente del número de alternativas consideradas, simplifica el proceso de estimación del modelo y de predicción (McFadden, 1978; Maddala, 1983).

No obstante, y considerando el cumplimiento de la *I.I.A.*, es posible profundizar en la modelización del esquema o árbol de decisión jerárquico, avanzando un poco más en el uso de los modelos lógit multinomiales, con el objetivo de considerar etapas de decisión intermedia.

Atendiendo a este nuevo objetivo, la formulación de la función de utilidad estricta implicará la inclusión de atributos o características propios de cada etapa o nivel de decisión. De este modo, en la función de utilidad (4.9),

$$U_{i(cj)} = V_{i(cj)} + e_{i(cj)},$$

la parte determinista (utilidad estricta) adoptará la siguiente forma:

[8] De esta forma, la variación esperada en la variable dependiente es constante para cualquier incremento unitario en el valor de la variable explicativa, *ceteris paribus*.
[9] *Independence of Irrelevant Alternatives*.

$$V_{i(cj)} = \beta' X_{i(cj)} + \alpha' Z_{i(c)}, \tag{4.12}$$

siendo $X_{i(cj)}$ un vector de atributos específicos de cada elección final para el individuo i, $Z_{i(c)}$ el vector de variables observadas que varían sólo con la elección intermedia (decisión de defraudar o no defraudar, sin detallar el tipo de fraude que se realiza) y β y α los vectores de parámetros correspondientes. Es posible hablar en términos de estos nuevos vectores de parámetros simplemente considerando que constituyen una concatenación de los parámetros referidos a cada alternativa, redefiniendo convenientemente el vector de explicativas.

En base a esta nueva formulación, la estimación de la probabilidad de elegir una determinada alternativa *cj*, será ahora el resultado de multiplicar dos probabilidades, una para cada nivel del árbol en el que nos situemos. Así,

$$P_{i(cj)} = P_{i(j|c)} P_{i(c)}, \tag{4.13}$$

donde:

$P_{i(cj)}$ es la probabilidad de que el individuo i elija la alternativa *(cj)*,

$P_{i(c)}$ es la probabilidad de que i elija la alternativa intermedia c, y

$P_{i(j|c)}$ es la probabilidad condicionada de elegir la alternativa j una vez el individuo ya se ha decidido por la alternativa c.

Teniendo en cuenta la definición del lógit multinomial, la probabilidad de que el individuo i elija una determinada opción final j (atendiendo a la elección previa de c) podrá definirse como:

$$P_{i(j|c)} = \frac{e^{V_{i(cj)}}}{\sum_{J=1}^{m_c} e^{V_{i(cJ)}}} = \frac{e^{\beta' X_{i(cj)}}}{\sum_{J=1}^{m_c} e^{\beta' X_{i(cJ)}}}, \tag{4.14}$$

donde J recoge el conjunto de elecciones posibles en la alternativa intermedia c. Por otro lado, la probabilidad asociada a la elección c atenderá a la expresión:

$$P_{i(c)} = \frac{\sum_{J=1}^{m_c} e^{V_{i(cJ)}}}{\sum_{b=1}^{C} \sum_{J'=1}^{m_b} e^{V_{i(bJ')}}} = \frac{e^{\alpha' Z_{i(c)}} \left[\sum_{J=1}^{m_c} e^{\beta' X_{i(cJ)}} \right]}{\sum_{b=1}^{C} \left[e^{\alpha' Z_{i(b)}} \left[\sum_{J'=1}^{m_b} e^{\beta' X_{i(bJ')}} \right] \right]}, \tag{4.15}$$

donde b recoge el conjunto de alternativas intermedias y J' el de alternativas dentro de cada opción intermedia. Además el subíndice i que indica el individuo especificado varía entre 1 y N.

Teniendo en cuenta la expresión anterior, el valor definido por:

$$I_{i(c)} = ln\left[\sum_{J=1}^{m_c} e^{\beta' X_{i(cJ)}}\right], \tag{4.16}$$

denominado *valor inclusivo*, recogerá la utilidad esperada agregada para un subconjunto de elección o conjunto de alternativas finales asociadas a la intermedia.

Teniendo en cuenta esta especificación y, a partir de las expresiones (4.13), (4.14) y (4.15) podemos reescribir la probabilidad de elegir una determinada alternativa *cj* en un modelo logístico anidado, como:

$$P_{i(cj)} = P_{i(j|c)} P_{i(c)} = \frac{e^{\beta' X_{i(cj)}}}{e^{I_{i(c)}}} \frac{e^{\alpha' Z_{i(c)} + I_{i(c)}}}{\sum_{b=1}^{C} e^{\alpha' Z_{i(b)} + I_{i(b)}}} = \frac{e^{\beta' X_{i(cj)} + \alpha' Z_{i(c)}}}{\sum_{b=1}^{C} e^{\alpha' Z_{i(b)} + I_{i(b)}}}. \tag{4.17}$$

Continuando con el desarrollo presentado, y siguiendo a McFadden (1978), cabe señalar que la estimación de la probabilidad de elegir una determinada alternativa puede realizarse estimando los parámetros β a partir del modelo condicional (4.14), determinado el valor inclusivo (4.16) y estimando el vector α a partir de la probabilidad marginal definida en (4.15). Esta aproximación no es sino una forma alternativa de modelizar la secuencia jerárquica de decisión, frente a lo que sería la estimación del modelo de forma completa (sin considerar la existencia de alternativas intermedias).

Sin embargo, el desarrollo de los modelos lógit multinomiales, tal y como hemos visto en páginas anteriores, se basa en la adopción de una determinada función de distribución para los términos aleatorios que definen la función de utilidad. Partiendo de la misma, los términos de error aleatorio se suponen idéntica e independientemente distribuidos, sin que se refleje, por lo tanto, la posible existencia de correlaciones entre los mismos. Ello queda de manifiesto en el hecho de que el coeficiente que acompaña al valor inclusivo es igual a la unidad.

La modelización de un proceso jerárquico de decisión, como el que venimos tratando, hace necesario analizar qué ocurre cuándo la hipótesis adoptada para los términos de error es violada, es decir, cuando es posible hablar de existencia de correlación entre las alternativas. Ante esta situación, el proceso econométrico a utilizar está fundamentado en el uso de los modelos lógit anidados, que pasamos a describir a continuación.

4.3 El Modelo Lógit Anidado

La introducción de un coeficiente diferente de la unidad para el valor inclusivo, que aparece en la probabilidad $P_{i(c)}$ de elegir una determinada opción c, dará lugar a un modelo lógit anidado (McFadden, (1978); Maddala (1983)). Este modelo es una generalización del modelo multinomial presentado, de la forma:

$$P_{i(c)} = \frac{e^{\alpha' Z_{i(c)} + (1-\sigma) I_{i(c)}}}{\sum_{b=1}^{C} e^{\alpha' Z_{i(b)} + (1-\sigma) I_{i(b)}}}, \qquad (4.18)$$

donde σ, $0 \leq \sigma \leq 1$, es un coeficiente que en breve interpretaremos.

Siguiendo a Maddala (1983), el modelo lógit multinomial anidado puede derivarse a partir de la adopción, para el vector de residuos $\vec{e}$, de una distribución *Valor Extremo Generalizado* (*G.E.V.*). Su característica principal radica en la posibilidad de estimación del modelo a pesar de la existencia de correlación entre aquellas alternativas que, estando en el mismo nivel de decisión, forman parte de conjuntos de elección diferentes. De esta forma, se relaja la hipótesis de independencia para los errores y, por tanto, no se exige el cumplimiento de la propiedad de Independencia entre Alternativas Irrelevantes (*I.I.A.*).

El modelo lógit anidado puede derivarse, al igual que hemos visto para el lógit multinomial, a partir de la Teoría de maximización de la Utilidad Aleatoria (McFadden, 1978).

Debido a la relación existente entre los modelos lógit anidados y los modelos de valor extremo generalizado[10] (los primeros pueden considerarse un caso particular de los segundos), es posible demostrar como los lógit anidados son también consistentes con la maximización de la utilidad aleatoria. No obstante, dicha consistencia viene marcada por el cumplimiento de unas ciertas condiciones para el coeficiente del valor inclusivo.

Este parámetro recoge una estimación de la similitud entre los términos no observados (errores aleatorios) del nivel inferior del árbol de decisión. Así, el campo de variación para el mismo oscilará entre cero y uno, según exista ausencia de correlación (σ=0) o correlación máxima (σ=1). Lógicamente, el grado de independencia (1-σ) oscilará también entre cero y uno.

[10] Un desarrollo exhaustivo de estos modelos puede encontrarse, asimismo, en McFadden (1978) y Maddala (1983).

El valor inclusivo, tal y como lo hemos definido en páginas anteriores, mide la utilidad agregada esperada asociada a un conjunto de alternativas. Los coeficientes para dichos valores, $(1-\sigma)$, se estiman con el resto de coeficientes del modelo y registran la disimilitud entre alternativas que pertenecen a un subconjunto en particular.

Siguiendo a McFadden (1978), la estructura anidada es consistente con la maximización de la utilidad aleatoria si y sólo si los coeficientes de los valores inclusivos se encuentran dentro del intervalo unitario[11]. Cuando la disimilitud $(1-\sigma)$ tiende a uno, σ tiende a cero y, por tanto, la correlación es prácticamente nula (se cumple la hipótesis de *I.I.A.*) y podemos modelizar utilizando un lógit multinomial simple[12]. Si el coeficiente tiende a 0, la correlación de los términos de error tiende a uno y, por tanto, los consumidores eligirán aquella alternativa que les proporcione una mayor utilidad estricta (atendiendo a la parte determinista de su función de utilidad y, por tanto, en base a sus atributos individuales y/o a los de elección), aunque las alternativas pueden no ser independientes.

La condición establecida en relación al intervalo de variación para $(1-\sigma)$ o para σ, es necesaria y suficiente para que el modelo lógit anidado sea consistente con la maximización de la utilidad (McFadden, 1978).

Si el coeficiente del valor inclusivo es mayor que uno ($(1-\sigma) >1$), se produce un seguimiento en el proceso de decisión a lo largo de los nudos del árbol, de forma que el resultado obtenido no es consistente con la maximización de la utilidad (el anidamiento no es consistente con la maximización de la utilidad).

En base a ello (Koujianou, 1995), cabe destacar que la jerarquía en el toma de decisiones no implica necesariamente que los consumidores (asegurados) elijan secuencialmente. En este caso, el árbol más que plasmar el proceso de selección del individuo, permite reflejar la existencia de posibles correlaciones entre factores no observados. Ello posibilita una modelización econométrica más general.

Así, por ejemplo, la interpretación puede realizarse teniendo en cuenta que el esquema de decisión seguido por el asegurado a la hora de optar por cometer o no fraude puede llevar asociados unos determinados costes de oportunidad (tiempo empleado en la preparación material del siniestro, tiempo invertido en diseñar la forma más verosímil posible del relato a presentar

[11] Ver en McFadden (1978) la demostración de la consistencia de los modelos lógit anidados con la maximización de la Utilidad Aleatoria.

[12] Los términos aleatorios se distribuyen idéntica e independientemente según una distribución *valor extremo tipo I* (McFadden, 1983).

por el propio asegurado y el contrario,...), de forma que, bajo este enfoque, el uso del árbol de decisión se fundamenta más en una forma de eliminar alternativas del conjunto de elección, que en una maximización de la utilidad esperada. Además cabe reseñar (Koujianou, 1995, pág. 914) que la presencia de un valor inclusivo con parámetro negativo no es consistente con la maximización de la utilidad.

El modelo lógit anidado puede ser estimado secuencialmente[13] (*"estimación por etapas"*). En el primer paso, se realiza la estimación de los parámetros de la log-verosimilitud condicional (nivel inferior). El modelo de elección discreta simple provee estimadores de β. El vector de parámetros estimados y las observaciones muestrales son utilizados para calcular los valores inclusivos. Después, en un segundo paso, se realiza la estimación en relación al nivel superior del árbol de decisión, siendo el valor inclusivo una de las variables exógenas del modelo.

El proceso de estimación secuencial permite obtener estimadores de los parámetros consistentes, pero ineficientes.

Además del método de estimación secuencial, es posible calcular los estimadores de los parámetros a partir del método de Máxima Verosimilitud Completa[14], realizándose la estimación del modelo de forma unificada.

4.4 Optimización del punto de corte en modelos con múltiples alternativas

La determinación del criterio probabilístico óptimo que permita efectuar la clasificación de una observación en una categoría de elección es habitual en las aplicaciones asociadas a modelos lógit sencillos, cuya finalidad sea predictiva.

Cuando la variable dependiente toma únicamente dos valores, recogiendo la ausencia/presencia de una determinada situación, el criterio para asignar un individuo a uno u otro grupo es, por defecto el de una probabilidad estimada superior o no a 0.5. Sin embargo dicho criterio es discutible desde el punto de vista de que, tomando a modo de ejemplo los resultados que obtendríamos al cuantificar la probabilidad de no existencia de fraude (Y=0) *versus* fraude (Y=1), un siniestro con probabilidad estimada de 0.51 sería clasificado como fraudulento mientras que otro con probabilidad estimada de 0.49 lo sería como no fraudulento. La solución a este problema radica en determinar a partir de qué nivel de probabilidad se debe

[13] Una aplicación importante puede encontrarse en Koujianou (1995). En este caso, la autora aplica un método secuencial para estimar un modelo anidado con cinco niveles de decisión.
[14] Full Information Maximum Likelihood (FIML).

considerar el siniestro como sospechoso de fraude teniendo en cuenta los resultados obtenidos (análisis del grupo de pertenencia observado y del predicho por el modelo).

Ahora bien, ¿qué ocurre cuando modelizamos teniendo en cuenta la posibilidad de múltiples respuestas de elección?. ¿Cómo determinamos el criterio probabilístico que nos permita optimizar la asignación de una observación a una determinada categoría?.

La asignación de una observación a una categoría de elección se fundamenta en el criterio de máxima probabilidad. De esta forma, una vez realizado el proceso de estimación, el grupo de pertenencia para la misma (categoría en que queda clasificado cada caso) será aquel para el que se obtiene una mayor probabilidad predicha. La suma de las probabilidades obtenidas ha de ser la unidad, de forma que teniendo en cuenta tres alternativas finales de elección, el modelo será incapaz de decidir cuando las probabilidades predichas sean iguales (Prob(Y=0)= 1/3; Prob(Y=1)= 1/3; Prob(Y=2)= 1/3). Este será el criterio seguido por defecto para delimitar las zonas de clasificación, sin, embargo, no tiene por qué ser el óptimo, tal y como veremos a continuación.

Sean P(0), P(1) y P(2) las probabilidades estimadas de clasificación en el primer, segundo y tercer grupo de elección, respectivamente. Sabiendo que la suma de las misma es igual a uno, es posible representar gráficamente las zonas de clasificación para cada una de las observaciones de la siguiente forma,

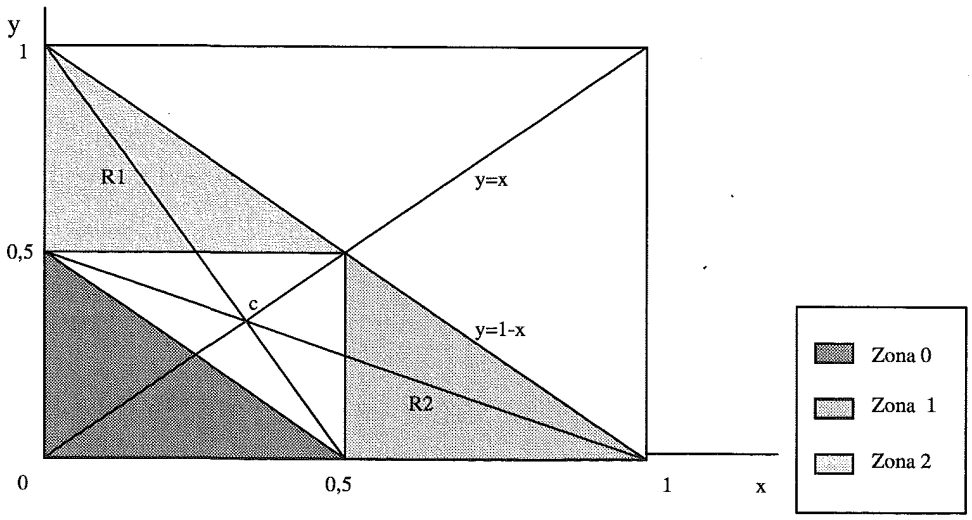

Gráfico 1

Por simplificación notacional hemos realizado un cambio de variables, siendo x= P(1) e y= P(2). Lógicamente, la utilización de complementarios, P(0)= 1-(P(1)+P(2)), permitirá deducir el valor correspondiente a la probabilidad para la primera categoría de elección.

Analicemos, paso a paso, la construcción del gráfico 1.

La hipotenusa del triángulo, representada por la recta y=1-x, permite delimitar las zonas de clara aceptación de la alternativa 1 (zona marcada con un 1) y de la alternativa 2 (zona 2), sin más que tener en cuenta, que siempre que la probabilidad estimada para una de ellas sea mayor a 0.5, la observación será clasificada en el grupo correspondiente.

Análogamente, la recta y=0.5-x delimita la zona de aceptación de la alternativa 0 (zona 0), teniendo en cuenta que todos los puntos situados por debajo de la misma tendrían una probabilidad estimada P(0)>0.5.

Sin embargo, como se desprende de la representación, queda una zona en la que *a priori* no sabríamos como clasificar la observación, siendo necesario fijar un criterio que nos ayude a incluirla dentro de cualquiera de las tres categorías iniciales, teniendo en cuenta el criterio de máxima probabilidad, es decir, verificando que con el mismo la observación que inicialmente está en zona de incertidumbre ha quedado clasificada en la categoría para la que el modelo otorga una mayor probabilidad estimada.

La representación sobre el triángulo de las rectas,

$R1 \Rightarrow y = 1-2x$
$R2 \Rightarrow y = (1-x)/2$

y la utilización de la diagonal (y=x), permite delimitar, teniendo en cuenta el punto de corte de las mismas (c), el criterio general de asignación de las observaciones en la zona 0, en la zona 1 o en la zona 2, respectivamente, siguiendo el siguiente proceso:

$$\text{Si} \quad \begin{cases} y < \dfrac{1-x}{2} \\ y < 1-2x \end{cases}, \tag{4.19}$$

la observación será clasificada dentro de la primera categoría de elección, al ser P(0) la probabilidad estimada máxima.

$$\text{Si} \quad \begin{cases} y < x \\ y > 1 - 2x \end{cases}, \tag{4.20}$$

la observación se clasificará dentro de la segunda categoría, lógicamente, al ser P(1) (es decir, x) la probabilidad máxima y,

$$\text{Si} \quad \begin{cases} y > x \\ y > \dfrac{1-x}{2} \end{cases}, \tag{4.21}$$

se clasificará en la tercera categoría atendiendo a la explicación anterior pero para P(2) (es decir, y).

La demostración de que la probabilidad estimada es máxima para cada una de las situaciones anteriores es sencilla.

Veamos en primer lugar como, si se cumplen las condiciones marcadas en (4.19), P(0)>P(1) y P(0)>P(2).

**Demostración 1.*

Siempre se verifica, por definición del modelo que $P(0)+P(1)+P(2)=1$

Sea (0) la categoría de asignación si $\begin{cases} \text{P(2)} < 1\text{-}2\text{P(1)} & \text{(a.1)} \\ \text{P(2)} < \dfrac{1\text{-P(1)}}{2} & \text{(b.1)} \end{cases}$.

Aplicando (b.1) $\Rightarrow 2\text{P(2)} < 1\text{-P(1)}$

$$1 - P(1) > 2P(2)$$

podemos demostrar $P(0) = 1 - P(1) - P(2) > 2P(2) - P(2) > P(2)$.

Aplicando (a.1) $\Rightarrow -P(2) > 2P(1) - 1$

podemos demostrar $P(0) = 1 - P(1) - P(2) > 1 - P(1) + 2P(1) - 1 > P(1)$.

Haciendo lo propio con P(1), teniendo en cuenta las condiciones que aparecen en la expresión (4.20):

**Demostración 2*

Como siempre, se cumple que $P(0)+P(1)+P(2)=1$

Sea (1) la categoría de asignación si $\begin{cases} P(2) < P(1) & (a.2) \\ P(2) > 1-2P(1) & (b.2) \end{cases}$.

Aplicando (b.2) $\Rightarrow P(2) > 1-2P(1)$

$$-P(2) < 2P(1)-1$$

podemos demostrar $P(1) = 1-P(0)-P(2) < 1-P(0)+2P(1)-1;$

$$P(1)-2P(1) < -P(0) \Rightarrow -P(1) < -P(0) \Rightarrow P(1) > P(0),$$

y $P(1) > P(2)$ por la propia condición (a.2) .

Por último, considerando las condiciones que aparecen en la expresión (4.21), podemos demostrar para P(2),

**Demostración 3*

Análogamente, como en los casos anteriores, se cumple que P(0)+P(1)+P(2)=1.

Sea (2) la categoría de asignación si $\begin{cases} P(2) > P(1) & (a.3) \\ P(2) > \dfrac{1-P(1)}{2} & (b.3) \end{cases}$.

Aplicando (b.3) $\Rightarrow 2P(2) > 1-P(1)$

$$-P(1) < 2P(2)-1$$

podemos demostrar $P(2) = 1-P(1)-P(0) < 1+2P(2)-1-P(0) = 2P(2)-P(0)$

$$P(2) < 2P(2)-P(0) \Rightarrow P(2)-2P(2) < -P(0) \Rightarrow -P(2) < -P(0) \Rightarrow P(2) > P(0),$$

y $P(2) > P(1)$ por la condición (a.3).

Finalmente, tal como queríamos comprobar, P(2)>P(0) y P(2)>P(1).

Una vez delimitadas las zonas de aceptación, ¿cómo podemos optimizar la clasificación de las observaciones, ampliando, por ejemplo, alguna de las áreas?. Parece lógico pensar que el desplazamiento del punto de corte *c* (variación del criterio probabilístico) a lo largo de las

rectas utilizadas o en otras direcciones permitirá modificar (en el sentido de ampliar o reducir) las zonas de clasificación.

Fijemos la variación de c a lo largo de la recta y=x y determinemos las ecuaciones para las rectas R1 y R2 teniendo en cuenta que ambas han de pasar por el punto seleccionado. Al establecer este cambio en el punto c estamos actuando de manera simétrica respecto a las zonas 1 y 2. De manera más general se podría proceder a cambiar el criterio para favorecer alguna de las tres elecciones posibles.

La solución al sistema de ecuaciones que pasan por un punto da lugar al siguiente resultado,

$$R1 \Rightarrow \left\{ \frac{x-0}{c} = \frac{y-1}{c-1}; \quad \frac{(c-1)x}{c} = y-1; \quad y = 1 + \frac{(c-1)}{c}x \right.$$

y, (4.22)

$$R2 \Rightarrow \left\{ \frac{x-1}{c-1} = \frac{y-0}{c}; \quad y = \frac{c(x-1)}{c-1} = \frac{c}{c-1}(x-1) \right.$$

De esta forma y teniendo en cuenta las condiciones marcadas anteriormente para delimitar las zonas de clasificación, podemos establecer un criterio para optimizar el punto de corte en un modelo logístico multinomial con tres alternativas, de manera que si,

$$\left\{ \begin{array}{l} y < \frac{c}{c-1}(x-1) \\ y < 1 + \frac{c-1}{c}x \end{array} \right\} \text{ ó } \left\{ \begin{array}{l} P(2) < \frac{c}{c-1}[P(1)-1] \\ P(2) < 1 + \frac{c-1}{c}P(1) \end{array} \right\} \tag{4.23}$$

clasificaremos la observación dentro de la primera categoría (zona 0); si

$$\left\{ \begin{array}{l} y < x \\ y > 1 + \frac{(c-1)}{c}x \end{array} \right\} \text{ ó } \left\{ \begin{array}{l} P(2) < P(1) \\ P(2) > 1 + \frac{c-1}{c}P(1) \end{array} \right\} \tag{4.24}$$

la clasificaremos en la segunda (zona 1) y, si

$$\left\{ \begin{array}{l} y > x \\ y > \frac{c}{c-1}(x-1) \end{array} \right\} \text{ ó } \left\{ \begin{array}{l} P(2) > P(1) \\ P(2) > \frac{c}{c-1}[P(1)-1] \end{array} \right\} \tag{4.25}$$

la clasificaremos en la tercera (zona 2).

La variación de c permitirá, por lo tanto, optimizar los ratios de clasificación.

Terminamos este punto señalando que el criterio establecido por defecto en los programas informáticos econométricos, en el caso de lógit multinomiales, es el de c=1/3, baricentro del triángulo utilizado en la demostración. Para los modelos anidados, la aplicación del proceso desarrollado se realizaría teniendo en cuenta las probabilidades finales (calculadas según la expresión 4.17).

El desarrollo presentado en la modelización de elección entre alternativas nos ha de permitir obtener conclusiones en orden al primero de los objetivos fijados al final del epígrafe 3.4: la cuantificación de la probabilidad de que un determinado asegurado cometa fraude[15]. Además, se ha establecido un mecanismo de causalidad a través del cual dicha probabilidad depende de un conjunto de factores explicativos.

Con el procedimiento que acabamos de generar, podemos centrarnos ya en el segundo de los objetivos: cuantificar la probabilidad de que el fraude sea detectado por la entidad[16]. La generación de una tabla de predicción (en base al uso de técnicas de predicción *ex-post*[17]) permite determinar el porcentaje de casos que son correctamente clasificados por el modelo (porcentaje de casos fraudulentos y no fraudulentos correctamente clasificados). En base a la relativización al total de observaciones, estaremos en condiciones de obtener una estimación agregada para la probabilidad de detectar el fraude y, por lo tanto, podremos aproximar, dado que conocemos todos los parámetros, la utilidad que el individuo asegurado espera obtener.

No obstante, los resultados obtenidos en relación a la probabilidad de detección de fraude deberán ser objeto de una serie de matizaciones, puesto que vendrán condicionados por el modelo que acabemos seleccionando para cuantificar la probabilidad de existencia de fraude. Sobre ello, realizaremos un análisis detallado en los Capítulos 6 y 7.

[15] Estimación, por tanto, del parámetro t que aparece en la formulación de Picard (1996).
[16] Estimar, por tanto, el parámetro q de la formulación de Picard (1996).
[17] Esta técnica de predicción consiste en seleccionar, del total de la muestra, un conjunto de observaciones para realizar la estimación, utilizando el resto de casos para validar el modelo.

5. APLICACIÓN EMPÍRICA A LA CUANTIFICACIÓN Y DETECCIÓN DE FRAUDE: DESCRIPCIÓN DE LA MUESTRA

5.1 Comentarios generales

La fundamentación teórica presentada en el Capítulo anterior, en relación al tratamiento del fraude, será materializada, a lo largo de los Capítulos 5, 6 y 7, en lo que constituye la aplicación empírica del trabajo realizado.

Los objetivos presentados en las páginas anteriores en relación a la cuantificación de la probabilidad de presencia de fraude en el seguro del automóvil y de su detección serán sujetos a modelización, en base a las técnicas econométricas comentadas en el Capítulo 4, bajo la aplicación de una muestra, obtenida con la colaboración de una de las principales aseguradoras del mercado asegurador español.

La generación de un *"modelo de detección y control de fraude en el seguro automovilístico español"* gozará, por lo tanto, de la particularidad que le confiere el trabajar con datos de una única compañía. No obstante, el funcionamiento más o menos análogo, observado entre entidades, podría permitirnos generalizar los resultados obtenidos, de cara a presentar el comportamiento observado para los principales indicadores de fraude a nivel español.

La disponibilidad de un amplio número de variables en relación a la ocurrencia del siniestro, a las características de la póliza y a determinados atributos de la parte asegurada y de la parte contraria ha permitido realizar un estudio exhaustivo de cuáles son los principales condicionantes que pueden incrementar la probabilidad de aparición de fraude. Asimismo, la creación de variables adicionales, definidas en base a las anteriores, ha contribuido a detectar situaciones directamente relacionadas con la aparición de comportamientos fraudulentos, tales

como, la ocurrencia del siniestro en fin de semana o con anterioridad a la fecha de emisión de la póliza.

La aplicación empírica realizada atiende al siguiente esquema de presentación. A continuación, dentro del Capítulo 5, se detalla, de forma exhaustiva, el contenido de la muestra de siniestros utilizada en la modelización. Además de realizar los comentarios oportunos sobre el tamaño y el carácter aleatorio o no de la misma, se presentan las variables que la componen y se describen estadísticamente. En especial aquéllas que se utilizarán directamente en los modelos presentados (en la mayoría de los casos, las variables seleccionadas no aparecerán directamente en la base de datos sino que habrán sido creadas a partir de la información contenida en la misma).

En el Capítulo 6 se presentan los resultados obtenidos al realizar la modelización bajo la aplicación de modelos lógit multinomiales. En el Capítulo 7 se presentan los resultados obtenidos al utilizar modelos logísticos anidados. En ambos casos, el objetivo es cuantificar la probabilidad de aparición de fraude y la probabilidad de que la compañía lo detecte. Asimismo, se detallan aquellas variables que resultan estadísticamente significativas de la aparición de comportamientos fraudulentos y se analiza la calidad del ajuste y capacidad predictiva de los modelos presentados.

La necesidad de realizar una preparación diferente de la base de datos para aplicar los métodos de estimación de los modelos lógit multinomiales y anidados implicará presentar un breve apartado de explicación, antes de ejecutar ambas modelizaciones, sobre la configuración de la misma.

Por último, en el Capítulo 8 se detallará el resumen de las principales conclusiones extraídas a lo largo de la Tesis, tanto desde un punto de vista teórico como desde un punto de vista aplicado.

5.2 Descripción de la base de datos utilizada en la modelización

5.2.1 Naturaleza de los datos

La muestra utilizada en la estimación de los modelos ha sido obtenida a partir de una selección aleatoria de siniestros de una de las principales entidades del mercado asegurador español.

La información proporcionada por la misma hace referencia a un total de 1995 expedientes de siniestros, todos ellos acaecidos entre 1993 y 1996, habiendo sido la mitad (998 expedientes) clasificados como legítimos y la otra mitad (997 expedientes) clasificados como fraudulentos por la compañía. Los expedientes fraudulentos son en nuestro caso siniestros donde se ha comprobado la existencia real de fraude (por ejemplo, por propia admisión) y no se reducen a expedientes sospechosos como en otros trabajos. En relación a estos últimos, el tipo de fraude identificado por la entidad atiende a la siguiente clasificación:

- Falsa declaración del asegurado para eludir casos excluidos en la póliza (246 casos).
- Falsa declaración del asegurado para obtener un beneficio sin intervención de un tercero (53 casos).
- Versiones cruzadas para cobrar ambos implicados (9 casos).
- Contratación de la póliza después de ocurrido el accidente (30 casos).
- Ocultación de alcoholemia (16 casos).
- Falso conductor habitual para eludir los recargos (10 casos).
- Fraude del taller (12 casos).
- Falsa declaración del asegurado para favorecer a un tercero (314 casos).
- Identificación de fraude sin disponibilidad de tipo (307 casos).

La información proporcionada posee cobertura nacional, en el sentido de que recoge siniestros ocurridos en prácticamente todas las provincias españolas. Asimismo, hace referencia a una amplia gama de coberturas aunque para su tratamiento éstas han sido englobadas en tres categorías genéricas[1]: "a terceros", "a terceros más complementarios" y "todo riesgo" (con o sin franquicia). Los diferentes tipos de vehículos quedan también suficientemente

[1] La clasificación inicial permite diferenciar entre un amplio conjunto de coberturas relacionadas con alternativas variadas de contratación de la póliza (normalmente relacionadas con la contratación del seguro obligatorio de responsabilidad civil y de coberturas adicionales).

representados y han sido clasificados en cuatro grupos fundamentales[2]: "turismos uso particular", "turismos otros usos", "motocicletas" y "camiones".

Los comentarios presentados ponen de manifiesto la disponibilidad de información muestral de elevada calidad y queda suficientemente justificada su aplicación para realizar un correcto tratamiento del fraude.

El análisis de la naturaleza de los datos revela, sin embargo, un hecho significativo: la existencia en la muestra de una sobre-representación para los siniestros fraudulentos. La corrección de este problema se realizará mediante la adecuada ponderación de las observaciones. La disponibilidad de información para los diferentes tipos de fraude nos permitirá modelizar la probabilidad de aparición de éste atendiendo a las diversas formas de comportamiento al alcance del asegurado defraudador. Además, seremos capaces de analizar la significación estadística de determinadas variables en relación a cada uno de los tipos de fraude presentados (normalmente, y con el objetivo de evitar una excesiva categorización de los tipos de fraude, éstos serán agrupados en clases más genéricas, siguiendo la clasificación planteada en los árboles de decisión presentados en el Capítulo 2, epígrafe 2.6).

5.2.2 Descripción de variables

La información contenida en la base de datos hace referencia a un total de 88 variables (en el Anexo 3 puede encontrarse la definición de las mismas), agrupadas, básicamente, en 9 bloques identificadores (atendiendo a la información suministrada):

- **Variables relacionadas directamente con la identificación del siniestro** (número de expediente, año de fabricación del vehículo[3], fecha de ocurrencia del siniestro, número de expedientes[4] que la compañía decide abrir para tramitar las diferentes consecuencias de un siniestro, número de siniestros asociados a la póliza, y culpa del accidente).

- **Variables relacionadas con la póliza** (coberturas contratadas, tipo-uso del vehículo, código del agente de seguros, situación de la póliza en el momento del siniestro, fecha de

[2] La clasificación inicial permitía diferenciar una gran variedad de posibilidades, atendiendo al uso del vehículo (particular o no), al tonelaje de carga para vehículos comerciales, al número de plazas para vehículos de transporte de personas, a la distancia en kilómetros recorrida de forma habitual por vehículos de transporte en general,...

[3] O de matriculación, si no se conociera la fecha de fabricación del vehículo.

[4] Asociados a diferentes conceptos, como daños propios, responsabilidad civil, recobros, reclamación de daños y/o lesiones.

vencimiento de recibo, fecha de efecto y de emisión de la póliza, número de suplementos[5] que ha tenido la póliza desde su emisión, existencia y cantidad de franquicia contratada, existencia y cantidad de cobertura por accesorios, clase de responsabilidad civil contratada, forma de pago de los recibos, marca del vehículo asegurado y color del mismo).

- **Variables relacionadas con el siniestro** (hora de ocurrencia del siniestro, provincia de ocurrencia del mismo, presencia o no de testigos, intervención o no de la policía, naturaleza del siniestro[6], relato del accidente por el asegurado, daños al vehículo asegurado, código del taller del asegurado y fecha de comunicación del siniestro).

- **Variables relacionadas con el asegurado-tomador del seguro** (lugar de residencia[7], sexo, estado civil, fecha de nacimiento y antigüedad de carnet).

- **Variables relacionadas con el propietario del vehículo asegurado** (información análoga a la presentada para el tomador del seguro).

- **Variables relacionadas con el conductor del vehículo asegurado** (información análoga a la presentada para el tomador y para el propietario del vehículo asegurado).

- **Variables relacionadas con el conductor contrario** (código de la compañía contraria, marca y color del vehículo contrario, daños ocasionados al mismo, taller de reparación, identificador de coincidencia entre apellidos de la parte asegurada y la contraria, lugar de residencia[8], sexo, estado civil, fecha de nacimiento y antigüedad de carnet).

- **Variable identificadora del tipo de fraude existente.**

- **Variable que mide el coste final del siniestro** (indicándose, en cada caso, si se trata de pagos o de recobros[9]).

El estudio descriptivo de la información contenida en la muestra disponible pone de manifiesto la presencia de un elevado número de variables de carácter cualitativo que han sido

[5] Se entiende por "suplemento" cualquier cambio que afecte a las condiciones de la póliza (coberturas, vehículo, conductor, forma de pago, accesorios,...).
[6] Ocurrencia del siniestro en zona urbana o en carretera.
[7] Únicamente se dispone del código postal.
[8] De nuevo, indicado por el código postal.
[9] La aplicación de los Convenios C.I.D.E. y/o A.S.C.I.D.E. implica que la compañía pague a su asegurado los daños derivados del siniestro y recobre de la compañía contraria un módulo fijo (que puede superar lo pagado).

objeto de la codificación oportuna. Las variables de carácter cuantitativo son menos numerosas dentro de la base de datos aunque quizá el propio funcionamiento del sistema asegurador, en términos de la información normalmente recogida, ayude a justificar este hecho. Es muy frecuente la aparición de variables que recogen información sobre fechas (fecha de ocurrencia del siniestro, fecha de comunicación del mismo,...) siendo su tratamiento una de las partes más elaboradas dentro del estudio que presentamos, dado que ha permitido examinar cada uno de los siniestros dentro de las reglas, que en términos de plazos, rigen en el mercado de seguros español.

Los grupos de edad y de antigüedad de carnet para los conductores asegurados, en el momento del siniestro, quedan representados de forma correcta dentro de la muestra utilizada, observándose intervalos de variación para los mismos suficientemente amplios. De esta forma, la edad del conductor asegurado oscila entre los 16 y los 83 años y la antigüedad de carnet entre los 0 y los 56 años.

En relación a la antigüedad del vehículo, los datos disponibles reflejan la ocurrencia de siniestros que involucran tanto a vehículos nuevos como a vehículos antiguos (la antigüedad del vehículo en el momento del siniestro oscila entre 0 y 37 años). También y, en relación al siniestro, es posible observar la existencia de culpa admitida tanto por parte del asegurado como por parte del contrario.

El intervalo horario y las fechas de ocurrencia del accidente (considerando el día de la semana en que se produjo) permiten modelizar en términos de horas de acaecimiento de los sucesos y de la mayor o menor relación de los mismos con el fin de semana.

El relato del siniestro por parte del asegurado, aunque puede ser breve, es muy informativo. Por ello, es posible realizar un análisis estadístico textual que permita señalar cuáles son las situaciones o circunstancias que tienen un mayor grado de asociación con las declaraciones fraudulentas.

Todas las pólizas se encontraban dadas de alta en el momento del accidente, y las diferentes formas de pago de los recibos (anual, semestral o trimestral) quedan reflejadas en la información facilitada.

En relación al sexo del conductor asegurado, es necesario hablar de una mayor proporción de hombres (1672 hombres frente a 323 mujeres) aunque no cabe esperar una sobre-

representación de los mismos puesto que ello ya se espera a nivel poblacional. En relación al estado civil del asegurado, la muestra evidencia casos para las cuatro categorías posibles (casado, soltero, separado y viudo). Para el conductor contrario y, en relación a estas dos variables, se observa una clara ausencia de información.

Atendiendo a las consideraciones que aparecen en la literatura existente (Cobo, 1993; Weisberg y Derrig, 1993; Belhadji y Dionne, 1997) sobre la presencia de determinadas circunstancias (o indicadores) que alertan sobre la existencia de fraude y en base a los resultados que hemos obtenido al realizar un estudio exhaustivo de las características propias de los siniestros con fraude detectado, las variables originales han sido debidamente seleccionadas con el objetivo de crear otras nuevas[10] que permitan modelizar, de forma adecuada, la presencia o no de comportamientos fraudulentos dentro de los expedientes analizados.

Las variables creadas son fundamentalmente dicotómicas y quedan definidas a continuación (omitimos el subíndice *i* que indicaría su referencia a cada individuo de la muestra),

- **Variables relacionadas con la póliza y el vehículo en el momento del contrato**

DRAMO: variable dicotómica que toma valor 1 si la cobertura contratada es a terceros (con o sin complementarios) y valor 0 si es a todo riesgo (con o sin franquicia).

DUSO: variable dicotómica que toma valor 1 si el vehículo asegurado es un turismo de uso particular y valor 0 en caso contrario (si se encuentra dentro de las categorías de "turismo otros usos", "motocicletas" o "camiones").

SFRANQUI: variable dicotómica que toma valor 1 si existe franquicia en la póliza; toma valor 0 en caso contrario.

ACCESORI: variable dicotómica que toma valor 1 si en la póliza existe cobertura de accesorios; toma valor 0 en caso contrario.

DPAGO1: variable dicotómica que toma valor 1 si el pago de los recibos es semestral; toma valor 0 si es anual o trimestral.

[10] Variables que recogen situaciones comúnmente asociadas a la presencia de fraude.

DPAGO2: variable dicotómica que toma valor 1 si el pago de los recibos es trimestral; toma valor 0 si es anual o semestral.

- **Variables relacionadas con el asegurado en el momento del siniestro**

HISTSTR: número de siniestros asociados a la póliza anteriores al estudiado.

EDAD: edad del conductor asegurado en el momento del siniestro.

CARNET: años de antigüedad de carnet del conductor asegurado en el momento del siniestro.

DSEXC: variable dicotómica que toma valor 1 si el conductor asegurado es hombre.

DECC: variable dicotómica que toma valor 1 si el conductor asegurado está casado; toma valor 0 en caso contrario (soltero, separado o viudo).

- **Variables basadas en la relación con el conductor contrario**

DESTAF: 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.

DDOMICIL: variable dicotómica que toma valor 1 si existe proximidad de domicilios[11] entre el conductor asegurado y el contrario; vale 0 en caso contrario.

- **Variables relacionadas con el siniestro y el vehículo en el momento del accidente**

DCULPA: variable dicotómica que toma valor 1 si la culpa es del asegurado y valor 0 si la culpa es del contrario.

ANTIGUO: antigüedad (en años) del vehículo en el momento del siniestro.

[11] La ausencia de información en relación a domicilios exactos (calle,...) ha supuesto la necesidad de analizar dicha proximidad atendiendo únicamente a los códigos postales de residencia declarados por el conductor asegurado y por el contrario, con todas las matizaciones que ello puede llevar asociado.

DCOMOCUR: variable dicotómica que toma valor 1 si la comunicación del siniestro se produce con posterioridad al séptimo día de su ocurrencia[12]; toma valor 0 en caso contrario.

ALTAHORA: variable dicotómica que toma valor 1 si el siniestro ha ocurrido a altas horas (entre las 11 de la noche y las 5 de la madrugada); vale 0 en caso contrario.

DZONA1: variable dicotómica que toma valor 1 si el siniestro ha ocurrido en zonas de siniestralidad alta[13]; toma valor 0 si ha ocurrido en zonas de siniestralidad media o baja.

DZONA3: variable dicotómica que toma valor 1 si el siniestro ha ocurrido en zonas de siniestralidad baja; toma valor 0 si ha ocurrido en zonas de siniestralidad alta o media.

CARRET: variable dicotómica que toma valor 1 si el siniestro ha ocurrido en carretera; vale 0 si el siniestro ha ocurrido en zona urbana.

SABDOM: variable dicotómica que toma valor 1 si el siniestro ha ocurrido en sábado o domingo; vale 0 en caso contrario.

DEFOCEMI: 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.

STEST: variable dicotómica que toma valor 1 si hay testigos del siniestro; vale 0 en caso contrario.

DRELATO: variable dicotómica que toma valor 1 si en la declaración del asegurado aparecen alguno de los relatos relacionados con: estacionamiento,

---

[12] La ley de Contrato del Seguro en su Art. 16 señala: *"el tomador del seguro o el asegurado o el beneficiario deberán comunicar al asegurador el acaecimiento del siniestro dentro del plazo máximo de siete días de haberlo conocido, ..."*.

[13] Atendiendo a la clasificación de zonas de siniestralidad en España presentada por U.N.E.S.P.A. (1993a). Teniendo en cuenta, entre otras cosas, factores climáticos y de infraestructuras, se diferencia entre:

- *Zona 1* (elevada siniestralidad): Galicia, Asturias, Cantabria, País Vasco y Navarra.
- *Zona 2* (siniestralidad media): Madrid y Cataluña.
- *Zona 3* (siniestralidad baja): Resto de Comunidades, Ceuta y Melilla.

aparcamiento, marcha atrás, adelantamiento y/o cruce; toma valor 0 en caso contrario.

SAUTDAD: variable dicotómica que toma valor 1 si hay intervención de policía en el siniestro; toma valor 0 en caso contrario.

5.3 Estadísticos descriptivos básicos

Los estadísticos de resumen para las variables presentadas en el epígrafe anterior quedan recogidos en las tablas que aparecen a continuación:

Tabla 1. Estadísticos descriptivos[14] para las variables cuantitativas
(tamaño muestral: 1995 casos)

Variable	Media		Desv. Standar		Mínimo		Máximo	
	Total	F. NF.	Total	F. NF.	Total	F. NF.	Total	F. NF.
HISTSTR	1.42		1.80		0		19	
		1.62		1.92		0		19
		1.22		1.64		0		13
EDAD	38		12.32		16		83	
		37		11.51		17		75
		39		12.00		16		83
CARNET	14		9.10		0		56	
		13		8.64		0		51
		15		9.45		0		56
ANTIGUO	6.17		4.49		0		37	
		6.26		4.55		0		37
		6.08		4.42		0		25

F: fraude (997 casos); NF: no fraude (998 casos)

[14] Estadísticos muestrales no ponderados.

En la Tabla 2 se presentan las frecuencias relativas para las variables dicotómicas:

Tabla 2. Frecuencias relativas[15] de las variables dicotómicas (en porcentaje) (tamaño muestral: 1995 casos)

	Frec. (%)		
Variable	Total	Fraude	No Fraude
DRAMO	90.6	94.1	87.1
DUSO	88.4	84.5	92.3
SFRANQUI	2.6	1.2	4.0
ACCESORI	6.8	5.9	7.6
DPAGO1	26.1	26.0	26.3
DPAGO2	24.3	26.0	22.5
DSEXC	83.8	85.2	82.5
DECC	65.8	63.7	67.9
DESTAF	6.3	9.7	2.8
DDOMICIL	25.0	27.5	22.4
DCULPA	32.0	47.1	16.9
DCOMOCUR	24.2	36.5	11.8
ALTAHORA	13.4	21.6	5.3
DZONA1	13.5	9.6	17.4
DZONA3	49.1	58.8	39.4
CARRET	7.2	8.9	5.4
SABDOM	27.1	30.7	23.5
DEFOCEMI	1.7	2.7	0.6
STEST	0.7	0.7	0.7
DRELATO	59.5	63.6	55.4
SAUTDAD	11.1	3.2	19.0

F: fraude (997 casos); NF: no fraude (998 casos)

El análisis de los datos que aparecen en las tablas anteriores refleja la existencia de variables en las que no se observan grandes diferencias, para los estadísticos planteados, entre las submuestras de fraude y no fraude. De este modo, por ejemplo, la media de la variable edad (Tabla 1) para los dos grupos considerados no muestra diferencias significativas (37 y 39 años). Lo mismo podría señalarse para las variables que recogen los años de antigüedad de carnet del conductor asegurado o la antigüedad del vehículo en el momento del siniestro. Las frecuencias relativas que, en la Tabla 2, aparecen para variables como la existencia de testigos o la forma de pago semestral de la póliza reflejan un comportamiento análogo. De esta forma y, aunque deberíamos ejecutar algún tipo de contraste (por ejemplo, de medias o proporciones) que nos permitiese ratificar lo anterior, parece lógico pensar que el estudio de la

[15] Estadísticos muestrales no ponderados.

influencia que las variables presentadas tienen sobre la elección de defraudar se ha de realizar considerándolas de forma conjunta. La aplicación de técnicas univariantes que recogiesen el efecto de las variables una a una, impediría obtener resultados significativos para aquellos indicadores con comportamientos similares en ambas submuestras. Por lo tanto, la modelización a aplicar ha de estar ligada a técnicas multivariantes que recojan el efecto conjunto entre las variables consideradas.

El análisis detallado de los descriptivos presentados revela como determinadas situaciones, asociadas, por ejemplo, a la ocurrencia del siniestro en fecha anterior o igual a la fecha de emisión de la póliza y posterior o igual a la fecha de efecto (variable *DEFOCEMI*), la comunicación del siniestro fuera del plazo establecido en la Ley de Contrato del Seguro (variable *DCOMOCUR*) y/o la ocurrencia del accidente a altas horas de la noche/madrugada (variable *ALTAHORA*) son más frecuentes en los casos en los que se ha detectado fraude. El mismo comportamiento se observa para la variable que recoge la aparición en la declaración del asegurado de relatos típicos como estacionamiento, marcha atrás,... (variable *DRELATO*). La coincidencia de apellidos entre la parte asegurada y la contraria (variable *DESTAF*)y la aceptación de culpa por parte del asegurado (variable *DCULPA*) presentan el mismo comportamiento.

A diferencia de las anteriores la intervención de policía en el siniestro (variable *SAUTDAD*) muestra una mayor frecuencia relativa para los casos de ausencia de fraude.

No queremos relegar a un segundo plano el tratamiento textual de la variable que explica cómo ocurrió el siniestro. Aunque es una información que aporta el propio asegurado y por tanto su fiabilidad puede quedar cuestionada, tiene una característica esencial. Es una de las pocas magnitudes sobre las que el asegurado puede ejercer algún tipo de manipulación. Muchas variables son observables directamente por la compañía o son más fácilmente comprobables o, simplemente, no permiten la flexibilidad que aporta una descripción verbal o escrita de lo ocurrido.

Para abordar el análisis de las respuestas textuales que aparecen en los expedientes hemos realizado una aproximación estadística básica al estudio de este tipo de datos, si bien somos conscientes de que un tratamiento textual más avanzado tendría mucho interés. Con él sería posible buscar asociaciones entre palabras e incluso poder plasmar distancias entre conceptos y proximidades de ciertos relatos a la presencia y ausencia de fraude. Nuestro análisis, si bien preliminar, se ha reducido a tener en cuenta la frecuencia de aparición de ciertos términos.

Para ello, se ha considerado la raíz semántica de la palabra y el contexto de su aparición. Este tratamiento es ciertamente laborioso, y ha permitido identificar ciertos relatos que se asocian con mayor frecuencia a la presencia de fraude. Esta aproximación es muy novedosa puesto que los estudios empíricos conocidos hasta la fecha no realizan ningún análisis de este tipo. Sin embargo, sí incluyen variables que intentan recoger el mismo tipo de información.

Es usual en otros estudios valorar la actitud del asegurado al realizar la reclamación y suele ser normal encontrar variables que indican si el mismo coopera con la entidad, es excesivamente explícito e incluye muchos detalles en su declaración o incluso, muestra mucha habilidad en el uso de la terminología utilizada en la tramitación, dando a entender que conoce a la perfección el funcionamiento del mercado asegurador (y, por tanto, es más fácil que sepa como actuar deshonestamente). La principal crítica que han recibido estas variables es su elevado grado de subjetividad puesto que dependen de la valoración que realiza el tramitador. En nuestro caso, no hemos podido disponer de ninguna valoración respecto a actitudes y comportamientos del asegurado al efectuar la declaración (excepto la dilación entre la ocurrencia del siniestro y su comunicación). Sin embargo, creemos que la utilización de los relatos del asegurado en la propia declaración aporta elementos informativos sobre su forma de proceder.

Igualmente, aunque no reviste el tratamiento estadístico anterior, se ha incluido el indicador de coincidencia entre los apellidos de las partes implicadas. Cabía esperar que la coincidencia de algún apellido y, más aún, la de los dos (sobre todo, si son apellidos poco frecuentes) debiera implicar una investigación automática, por parte de la compañía, de las circunstancias del accidente. Ello no es así (debido a la falta de automatización en la tramitación) y, por tanto, también la hemos considerado como variable potencialmente explicativa de circunstancias sospechosas.

El paquete estadístico utilizado en el tratamiento estadístico inicial de la base de datos ha sido el SPSS para Windows, versión 6.1.3. Para terminar este capítulo queremos señalar que en el Anexo 1 presentamos una serie de variables[16] que aportarían información adicional a la actualmente recogida por las entidades españolas. En realidad, al analizarlas una a una, se observan apectos muchas veces conocidos por los tramitadores, pero para los que no se suele hacer, en muchos casos, una recogida de datos explícita.

[16] Teniendo en cuenta variables sugeridas por otros investigadores (Weisberg y Derrig, 1993; Belhadji y Dionne, 1997).

6. APLICACIÓN EMPÍRICA I: MODELIZACIÓN LOGÍSTICA MULTINOMIAL

6.1 Introducción

Las consideraciones que hemos ido estableciendo a lo largo de la Tesis en relación al estudio y tratamiento del fraude en el seguro del automóvil, son materializadas a lo largo de este Capítulo y del siguiente en la determinación de un conjunto de modelos que permitan cuantificar su probabilidad de existencia.

Como hemos establecido, el asegurado realiza una elección entre alternativas discretas, persiguiendo, en cualquiera de los casos, maximizar su utilidad esperada. Sin embargo, el tratamiento econométrico susceptible de aplicación es variado y está sujeto a la disponibilidad de información adecuada.

En la generación de un modelo de detección y control del comportamiento fraudulento es necesario tener en cuenta un hecho importante: poblacionalmente hablando, los niveles existentes de fraude son de difícil cuantificación. Las divergencias presentadas en el Capítulo 2 (apartado 2.4) en relación a los porcentajes estimados de fraude (para diferentes países y ramos) y la poca claridad con que éstos son presentados por los diferentes autores (a veces es difícil diferenciar entre si los porcentajes son sobre fraude detectado o sobre sospecha de fraude) dificulta la obtención de valores poblacionales fiables.

Atendiendo a lo anterior, la determinación del tamaño muestral óptimo que asegure la representatividad de la población puede estar sujeta a matizaciones, más aún, cuando nada garantiza que en muchos de los siniestros que las compañías tienen calificados como de no fraudulentos puede existir fraude subyacente[1]. No obstante, la generación de una muestra, como la analizada en el Capítulo anterior, en la que se dispone de información suficiente para expedientes de siniestros sin fraude detectado y para expedientes fraudulentos nos ha

[1] Sólo los siniestros que han sido investigados exhaustivamente por la entidad en los que se ha reconocido la acusación o ha existido condena responden a una clasificación correcta en términos de fraude. Para los que se clasifican como no fraude cabría dudar sobre si la investigación ha sido suficiente.

permitido desarrollar una aproximación innovadora al tratamiento del mismo. El criterio de selección muestral seguido en el diseño y recogida de información y, la introducción de ponderaciones que adecuen, en la medida de lo posible, el comportamiento muestral al poblacional serán factores importantes a tener en cuenta en la estimación de los modelos.

La modelización del fraude por parte del asegurado admite, desde la óptica de los modelos de elección[2], una triple aproximación:

- especificación de modelos de regresión logística simple,
- especificación de modelos lógit multinomiales y
- especificación de modelos lógit anidados.

La primera y más sencilla supone definir como variable dependiente del modelo la dicotomía ausencia/presencia de fraude. El estudio se realiza en sentido genérico (sin diferenciar entre tipologías de fraude) y los resultados, en términos de significación estadística de las variables, permiten determinar qué circunstancias tienen un mayor poder explicativo en la aparición de actitudes deshonestas.

El uso de modelos lógit multinomiales supone no sólo trabajar con una metodología más avanzada sino también, obtener conclusiones en relación a las diferentes formas de comportamiento que puede presentar el asegurado defraudador. Los tipos de fraude son variados (como veíamos en el Capítulo anterior, la variable *SUBFRAU* presenta hasta nueve valores diferentes asociados a comportamientos fraudulentos distintos[3] -fraude a favor del asegurado, a favor del contrario, del taller,...-) y la determinación de las variables estadísticamente significativas en la explicación de cada uno de ellos puede ser decisiva a la hora de establecer patrones de investigación por parte de la compañía. En definitiva se trata de "matizar" y de proponer la metodología correspondiente para distinguir formas de fraude (aspecto, el de la existencia de diversos tipos de comportamientos deshonestos, sobre el que existe un amplio consenso entre los especialistas). Además, como veremos, la capacidad predictiva del modelo (fundamentada en el porcentaje de casos correctamente clasificados por el mismo[4]) ha sido mejorada gracias al diseño, dentro del presente trabajo (Capítulo 4), de un método que permita seleccionar el punto óptimo de corte en modelos multinomiales.

---

[2] Modelos de respuesta cualitativa o de elección individual

[3] No obstante, cabría hablar únicamente de ocho clases de fraude recogidas en la muestra dado que la última categoría recoge casos fraudulentos sin identificación de tipo.

[4] Porcentaje de casos no fraudulentos y fraudulentos (teniendo en cuenta la tipología existente) correctamente clasificados como tales.

El hecho de disponer únicamente de información relativa al individuo (asegurado/contrario), a la póliza y al siniestro y no de la elección (características propias de cada uno de los posibles comportamientos) supondrá la aplicación *a priori* de modelos lógit multinomiales propiamente dichos, en contraposición a los denominados lógit condicionales[5] (en éstos se tienen en cuenta, además, atributos específicos de cada una de las alternativas sobre las que versa el proceso de decisión).

Esta limitación, que puede no ser importante en la estimación de modelos multinomiales, lo es en la especificación de lógit anidados. La utilización de modelos jerárquicos es, sin duda, la aplicación más sofisticada de regresión logística y permite interpretar resultados atendiendo, bien al criterio de maximización de utilidad (McFadden, 1978), bien al deseo de tomar decisiones de forma secuencial eliminando alternativas no deseadas[6] (Koujianou, 1995). La preparación de la base de datos (nuevo diseño de la misma con el objetivo de hacerla susceptible de aplicación de este tipo de modelos) y la creación de variables asociadas a las elecciones ("atributos") permitirá obtener resultados atrayentes con una técnica de modelización caracterizada, fundamentalmente, por solventar el problema de la posible existencia de correlación entre las alternativas.

En este Capítulo nos centraremos en el análisis de resultados obtenidos tras la aplicación de un modelo lógit multinomial mientras, que en el Capítulo 7 hacemos lo propio ampliando las perspectivas al uso de modelos anidados. Las principales conclusiones derivadas de la aplicación de un lógit simple pueden encontrarse en Artís, Ayuso y Guillén (1998).

6.2 Criterio de selección muestral. Determinación de las ponderaciones a introducir en los modelos

La información muestral que hemos utilizado en la estimación de los modelos logísticos, descrita ampliamente en el Capítulo anterior, proviene de una selección de siniestros declarados aleatoria estratificada con estratificación endógena ("muestreo basado en la elección"). Desde este punto de vista, su diseño no atiende a los criterios propios de muestreo aleatorio simple en sentido estricto, sino que viene fundamentado en técnicas de recogida de

[5] *Modelo lógit condicional de McFadden* (Maddala, 1983).

[6] Al final del apartado 4.3, dentro del Capítulo 4, hacíamos referencia a ambas interpretaciones teniendo en cuenta el intervalo de variación para los coeficientes de los valores inclusivos. Si éstos oscilan entre 0 y 1 podremos interpretar los resultados atendiendo al criterio de maximización de utilidad.

información basadas en la variable dependiente del modelo[7]. De esta forma, la probabilidad de que un individuo esté en la muestra depende de la alternativa que haya elegido.

El muestreo basado en la elección es útil cuando la realización de una recogida de información completamente aleatoria o basada en alguna de las variables explicativas (o todas) incluidas en el modelo, sólo consigue una escasa representatividad, ya que deriva en la obtención de un número insuficiente de individuos eligiendo una alternativa particular. En la literatura existente (Maddala, 1983; Amemiya, 1985; Pudney, 1989) se hace referencia a esta técnica de muestreo aludiendo al diseño de una muestra relacionada con decisiones de transporte. A pesar de que una de las ventajas manifestada por los autores es la económica (recoger información en las paradas de autobús o de metro es menos costoso que hacer entrevistas en las casas), el hecho de que gracias a la misma es posible alcanzar una cuota adecuada de representación de todas las alternativas la hace más atractiva.

A pesar de la dificultad comentada en la introducción de este Capítulo en relación a trabajar con porcentajes poblacionales de fraude fiables, parece lógico pensar que el número de expedientes de siniestros sin fraude sea notablemente superior al de fraudes. Las compañías, hasta el momento, no han realizado políticas agresivas de detección de fraude y, por ello, la obtención de información en base a muestreo aleatorio hubiera derivado, probablemente, en una escasa presencia de expedientes de este tipo. Es por ello que la recogida de datos se realizó mediante una estratificación endógena: se seleccionaron aleatoriamente entre los expedientes sin fraude (o al menos sin fraude detectado) un total de 1000 casos y de los expedientes clasificados como fraudulentos y, también aleatoriamente, un total de 1000. La selección aleatoria de estos últimos ha supuesto una notable variación en el número de casos clasificados en los diferentes tipos de fraude. Finalmente se rechazaron 2 expedientes sin fraude y 3 con fraude, debido a la incongruencia en la información que suministraban.

Ahora bien, ¿qué implicación tiene el hecho de trabajar con una muestra formada por un 50% de siniestros de cada clase?. Si retomamos lo mencionado al inicio del párrafo anterior, dando como válido el hecho de que poblacionalmente cabe esperar que el número de no-fraudes sea muy superior al de fraudes, debemos hacer algún tipo de corrección con el objetivo de ajustar los resultados obtenidos tras la modelización al comportamiento de la población. La introducción de pesos o ponderaciones dentro del modelo permitirá corregir la infra-representación de los no fraudes y la sobre-representación de los fraudes.

[7] Amemiya (1985) diferencia básicamente entre tres tipos de técnicas de muestreo: *muestreo aleatorio, muestreo exógeno* (realizado en base a las variables independientes o explicativas) y *muestreo endógeno* (según la variable dependiente). Esta última técnica se conoce, dentro de los modelos de respuesta cualitativa, como "muestreo basado en la elección (*choice-based sampling*)".

¿Cómo determinaremos los pesos a introducir en el modelo?.

La muestra ha sido diseñada deliberadamente para contener un 50% de casos sin fraude y un 50% de casos fraudulentos mientras que las proporciones en la población (tomando como referencia las cifras presentadas por Cobo, 1993) son del 78% y 22%, respectivamente. De esta forma, los casos de no fraude están infra-representados por un factor de 0.64 (=0.5/0.78) mientras que los de fraude aparecen sobre-reperesentados por un factor de 2.27 (=0.5/0.22). Para obtener una representación adecuada en la población será necesario, por tanto, ponderar las observaciones de no fraude por un factor de 1.56 (=0.78/0.50) y las de fraude por 0.44 (=0.22/0.50).

Lógicamente, la consideración de diferentes tipos de fraude supondrá modificar adecuadamente los factores de ponderación. Si bien los pesos anteriores serán los utilizados en la especificación y estimación del modelo lógit simple (modelizando directamente la probabilidad de presencia de fraude *versus* no fraude), la agrupación de las diferentes clases de comportamientos fraudulentos en categorías más genéricas supondrá trabajar con diferentes tamaños muestrales y por lo tanto con diferentes múltiplos del factor de ponderación (construidos, no obstante, bajo el criterio presentado en el párrafo anterior).

La aplicación de modelos lógit multinomiales se ha realizado teniendo en cuenta el diseño previo de dos esquemas de elección que se presentan en las Figuras 8 y 9. En el primero de ellos, el asegurado ha de elegir entre tres posibles alternativas: no cometer fraude; cometer fraude a favor de sí mismo o hacerlo a favor del contrario.

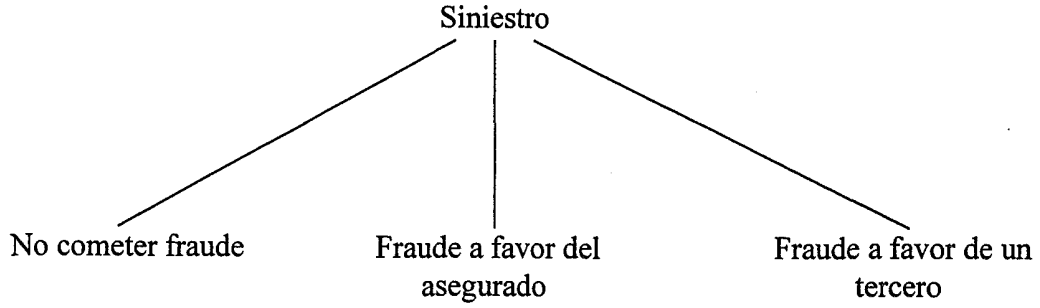

Figura 8

En el segundo, y bajo el objetivo de recoger de algún modo la posible intencionalidad en la creación del siniestro, las alternativas modelizadas son: no cometer fraude; cometer fraude para conseguir una indemnización o hacerlo para incrementar la indemnización que corresponde en función de la póliza contratada.

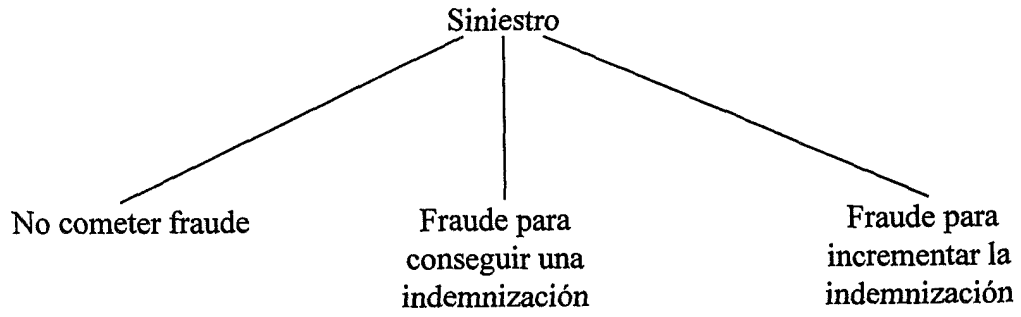

Figura 9

Lógicamente, el objetivo perseguido es modelizar la probabilidad de que el individuo cometa fraude atendiendo, no a todas y cada una de las formas posibles que tiene de hacerlo, sino a criterios generales fundamentados en el propio sistema asegurador.

El análisis de la información disponible para los tipos de fraude pone de manifiesto, tal y como hemos analizado en el Capítulo anterior, la existencia de nueve formas diferentes de clasificar los expedientes en los que la compañía ha detectado existencia de comportamiento fraudulento. Modelizar por separado los posibles comportamientos detectados hubiera derivado en un modelo con un elevado número de parámetros, más aún cuando los pocos casos englobados en algunas de las categorías hubiera impedido extrapolar poblacionalmente algunos de los resultados obtenidos. El objetivo perseguido se centra en justificar determinados patrones de comportamiento, analizando qué circunstancias los favorecen. Englobar categorías en otras más genéricas no supone ningún impedimento, sobre todo, cuando a veces resulta difícil comprender cuál es la razón que lleva al tramitador a englobar el siniestro dentro de un tipo u otro de fraude.

¿Cuál ha sido el criterio seguido a la hora de diseñar las alternativas finales?.

El seguro del automóvil español se fundamenta en dos aspectos de gran importancia a la hora de intentar dar un enfoque al estudio del fraude. Por un lado, las coberturas contratadas, si bien pueden estar relacionadas con la protección de daños sobre el propio vehículo, siempre tienen en cuenta la reparación de daños a un tercero[8]. Por otro, el hecho de que la mayoría de entidades españolas "estén en convenio[9]" facilita la posible colaboración entre las partes (asegurado y contrario). La eliminación de procesos judiciales y la rapidez con que se han de

[8] Siempre teniendo en cuenta la calificación que de "tercero" aparece en el condicionado de la póliza. Por ejemplo, es posible que en el marco de la cobertura de responsabilidad civil de suscripción voluntaria no tengan la consideración de terceros los familiares hasta tercer grado de consanguinidad. Una explicación análoga puede aplicarse, cuando el asegurado sea una persona jurídica, para sus representantes o los familiares de los mismos.
[9] Aplican convenios (C.I.D.E./A.S.C.I.D.E).

pagar las indemnizaciones son, sin duda, factores que favorecen los comportamientos deshonestos o al menos dificultan su detección.

En base a lo anterior parece lógico realizar una agrupación de tipos de fraude como la presentada en el primer esquema (Figura 8). La clasificación presentada en el segundo (Figura 9) pretende diferenciar, básicamente, entre aquellos casos en los que se persiguen evitar exclusiones de cobertura (el siniestro se produce bajo determinadas circunstancias que quedan fuera de las condiciones de la póliza) y aquéllos en los que el objetivo es aumentar la indemnización correspondiente (el siniestro está cubierto pero el asegurado quiere obtener un beneficio no justificado). En cierto modo, esta clasificación guardará analogías con la utilizada por Weisberg y Derrig (1993) y que diferencia entre fraude *a priori* (intencionalidad en la creación del accidente) y fraude *a posteriori* (uso indebido de la ocurrencia real del siniestro). No obstante, el hecho de que utilizando la información disponible no podamos diferenciar con exactitud en qué casos el siniestro ha sido creado premeditadamente o en qué casos se ha producido bajo situaciones excluidas impide establecer comparaciones con los resultados obtenidos por los autores mencionados.

La agrupación de tipos se podría haber realizado bajo la premisa de otras consideraciones. Por ejemplo, sería posible analizar si el fraude ha sido cometido de forma individual o en connivencia. Aún más, si la connivencia se ha realizado con la parte contraria u otros estamentos (talleres, médicos, abogados,...). Sin embargo, para este último caso, no disponemos de información completa en la muestra utilizada.

La categorización del fraude en dos grupos según su pretensión de beneficiar al asegurado o beneficiar a un tercero se ha realizado del siguiente modo. Aquellos casos en los que la compañía ha detectado falsas declaraciones del asegurado para eludir casos excluidos en la póliza (246 expedientes) o para obtener un beneficio sin intervención de un tercero (53 expedientes) han sido utilizados como submuestra de la categoría *fraude a favor del asegurado*. Aquellos en los que queda de manifiesto, de forma expresa, que el asegurado ha realizado una declaración falsa para favorecer a un tercero (314 casos) determinan la submuestra de *fraude a favor de un tercero*. El tamaño muestral utilizado en la estimación del primer modelo es, por tanto, de 1611 observaciones (teniendo en cuenta, lógicamente, los 998 casos en los que no se ha detectado comportamiento fraudulento) perdiéndose un total de 384 observaciones.

Para realizar la agrupación de casos teniendo en cuenta si el objetivo del fraude es conseguir una indemnización o incrementar la correspondiente según la póliza contratada hemos seguido

el siguiente criterio. El conjunto de expedientes clasificados por la entidad como fraudulentos por poseer falsa declaración del asegurado para eludir casos excluidos en la póliza (246 casos), por haberse contratado la póliza después de haber ocurrido el accidente (30 casos), por detectarse ocultación de alcoholemia (16 casos) o porque se haya declarado un falso conductor habitual para eludir recargos (10 casos) han sido agrupados bajo la categoría genérica de *fraude para conseguir una indemnización.* Los expedientes en los que se han detectado comportamientos del tipo de falsa declaración del asegurado para obtener un beneficio sin intervención de un tercero (53 casos), presentación de versiones cruzadas para cobrar el asegurado y el contrario (9 casos) o declaración falsa del asegurado para favorecer a un tercero (314 casos) han determinado la clase de *fraude para incrementar la indemnización.* Finalmente, la muestra utilizada para realizar la estimación este segundo modelo está constituida por un total de 1676 casos (319 menos que en la inicial).

Una vez diseñadas las muestras a utilizar en cada uno de los casos, las ponderaciones introducidas en cada uno de los modelos para corregir la sobre-representación de expedientes con fraude han sido las siguientes. En el primero, las observaciones de *no fraude* han sido ponderadas por un factor de 1.26 (=0.78/0.62), las de *fraude a favor del asegurado* por 0.58 (=0.11/0.19) y las de *fraude a favor de un tercero* también por 0.58 (=0.11/0.19). Para los casos fraudulentos y dado que poblacionalmente aceptamos que en un 22% de los casos existe sospecha[10] de fraude (Cobo, 1993), adoptamos la hipótesis de que un 11% serán de un tipo y el 11% restante del otro. Esta hipótesis está fundamentada en una ausencia de información sobre el comportamiento de ambas clases en la población aunque, en este caso, la gran similitud entre el número de observaciones obtenidas a partir de la muestra para cada una de ellas nos aporta elementos para justificarla

Para el segundo modelo, las ponderaciones son muy similares. Los casos no fraudulentos han sido ponderados por un factor de 1.3 (=0.78/0.60); los casos en los que el motivo del fraude es conseguir una indemnización por 0.61 (=0.11/0.18) y aquéllos en los que se persigue incrementar la cantidad a percibir por la ocurrencia del siniestro por 0.5 (=0.11/0.22).

[10] Lógicamente, este porcentaje puede estar sujeto a matizaciones dado que puede ser que exista sospecha de fraude y sin embargo no sea posible demostrar su existencia (Cobo, 1993). No obstante, teniendo en cuenta que a nivel de mercado y a nivel científico es comúnmente aceptada la existencia de un elevado porcentaje de fraudes no detectados (a modo de ejemplo Derrig, Weisberg y Chen (1994) estiman que el porcentaje de siniestros con sospecha de fraude estaría entre el 40 y el 50%) y dado que el porcentaje presentado surge como resultado de un proceso de muestreo creemos justificada su aplicación.

Una vez se ha puesto de manifiesto cuál ha sido el criterio seguido en el diseño de la muestra y se ha justificado el uso de ponderaciones en los modelos, pasamos a continuación a presentar el método de estimación utilizado y los principales resultados obtenidos.

6.3 Estimación de los modelos logísticos multinomiales

La estimación de los modelos lógit multinomiales suele realizarse, tal y como comentamos en el Capítulo 4, por el método de Máxima Verosimilitud[11]. Su aplicación permite la obtención de estimadores que cumplen propiedades estadísticas deseables al ser posible demostrar la normalidad, consistencia y eficiencia asintótica de los mismos (Amemiya, 1985).

Sin embargo, el logaritmo neperiano de la función de verosimilitud definido como

$$\ln L = \sum_{i} \sum_{j=0}^{m} Y_{ij} \ln P_{i(j)} \qquad (6.1)$$

ha de ser corregido cuando la base de datos utilizada en la modelización presenta estratificación endógena. En la expresión (6.1) Y_{ij} indica con un 1 si el individuo i elige la alternativa j (ó 0 en caso contrario) con probabilidad $P_{i(j)}$.

El criterio seguido en el diseño de la muestra y la necesidad de modelizar introduciendo pesos ha de estar presente en el proceso de estimación. Tratar una base de datos estratificada como si fuera aleatoria conduciría a estimadores inconsistentes (Manski y Lerman, 1977; Manski y McFadden, 1981; Amemiya, 1985), de manera que la función de verosimilitud que hemos de maximizar para obtener el vector de coeficientes estimados de la regresión[12] y la matriz de varianzas y covarianzas asintótica de los mismos, ha de ser una función ponderada[13].

En este sentido, la contribución de cada una de las observaciones al logaritmo neperiano de la verosimilitud del modelo ha de estar multiplicada proporcionalmente por el factor $w(j)=Q(j)/H(j)$, siendo $Q(j)$ la proporción de población que elige la alternativa j y $H(j)$ la proporción análoga teniendo en cuenta el muestreo basado en la elección (Maddala, 1983). $H(j)$ está influida por el criterio de estratificación seguido en el diseño de la muestra y es

[11] En el Anexo 2 pueden encontrarse resumidas las principales etapas en el proceso de estimación por Máxima Verosimilitud de un modelo lógit multinomial.

[12] En Manski y Lerman (1977) y en Amemiya (1985) queda demostrada la consistencia y normalidad asintótica del vector de estimadores máximo verosímiles obtenido a partir de la función de log-verosimilitud ponderada.

[13] En el Anexo 2 aparece resumido el proceso de estimación de la función de verosimilitud ponderada.

posible que no tome el valor óptimo de cara a minimizar la matriz de varianzas y covarianzas asintótica de los estimadores. La corrección de esta matriz con el objetivo de recoger la sobre-representación en la muestra de algunas de las alternativas (Manski y Lerman, 1977; Manski y McFadden, 1981) garantizará la eficiencia asintótica de los estimadores.

La estimación de los modelos lógit multinomiales y también de los anidados, se ha realizado con el programa econométrico LIMDEP, versión 7.0. Se ha incorporado una variable de ponderación durante el proceso para corregir el efecto del diseño muestral y el cálculo de la matriz de covarianzas asintótica de los estimadores se ha realizado a partir del siguiente producto matricial, $V=A\Lambda A$, siendo A la inversa negativa de la Hessiana del logaritmo neperiano de la función de verosimilitud ponderada y Λ la suma de los productos externos de las primeras derivadas de dicha función.

6.3.1 Estimación logística multinomial de la probabilidad de existencia de *fraude a favor del asegurado* o *a favor de un tercero* (MODELO 1)

Presentamos a continuación el análisis detallado de los principales resultados obtenidos para el primer modelo de elección (Figura 8).

6.3.1.1 Análisis de la significación individual y global de modelo

La especificación del modelo que nos permita cuantificar la probabilidad de que el asegurado no cometa fraude, de que lo haga a favor de sí mismo o para beneficiar a un tercero (Figura 8) se ha realizado teniendo en cuenta una amplio conjunto de variables explicativas, todas ellas presentadas en el Capítulo anterior, y directamente relacionadas con aspectos del propio sistema asegurador (póliza, siniestro,...). El hecho de no disponer en la muestra de atributos específicos para cada una de las elecciones ha sugerido la aplicación de modelos lógit multinomiales. Así, el árbol que presentábamos para este mismo esquema de decisión en el Capítulo 2, en el que teníamos en cuenta una etapa intermedia de decisión,

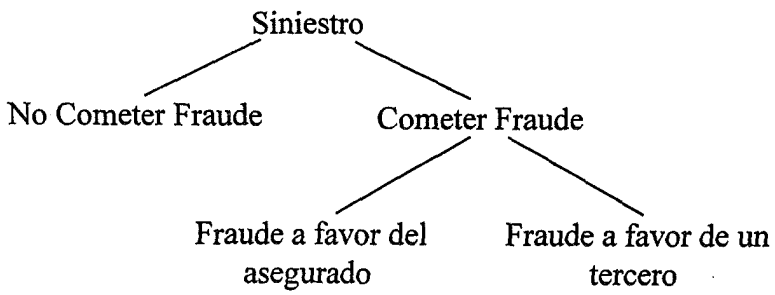

Figura 10

ha sido simplificado, teniéndose en cuenta únicamente las alternativas finales (Figura 8).

Al no disponer de ninguna variable que tome valores diferentes para las dos categorías de fraude presentadas parece lógico pensar en una modelización de este tipo, siendo necesario, no obstante, demostrar que ambas alternativas son independientes desde un punto de vista econométrico (en realidad el esquema de decisión subyacente es el presentado en la Figura 10, y por lo tanto podría suceder que los términos de error de las funciones de utilidad para las alternativas estuvieran correlacionados). El test de Hausman-McFadden nos permitirá, más adelante, demostrar el cumplimiento de la hipótesis de Independencia de Alternativas Irrelevantes.

La variable dependiente es el tipo de siniestro que hemos codificado por: Y=0, si el siniestro no es fraudulento; Y=1, si el asegurado ha defraudado en beneficio propio e, Y=2, si existe fraude para beneficiar a un tercero. Las observaciones de dicha variable han sido ponderadas durante el proceso de estimación para tener en cuenta el diseño muestral. Consideramos la categoría de los siniestros no fraudulentos como la de referencia.

En la especificación del modelo se han introducido diecinueve variables explicativas (incluido el término independiente). Las variables, definidas en el capítulo anterior no se refieren a las elecciones finales, es decir, no cambian con la alternativa elegida. Todas ellas están relacionadas con aspectos de la póliza, del automóvil y del siniestro, siendo la inmediatez su principal característica (esta conclusión podría matizarse para la variable que recoge la culpa del siniestro). La compañía dispone de la información que las caracteriza en un breve periodo de tiempo, lo que sin duda constituye un factor importante a considerar en la agilización de la investigación de siniestros.

Los parámetros estimados aparecen en la página siguiente:

Tabla 3. Resultados de la estimación de un modelo lógit multinomial [MODELO 1]

variable	coeficiente	t-test	P-valor
Fraude para beneficio del asegurado:			
CONSTANTE	-3.403	-5.357	0.000
DRAMO	1.067	2.516	0.012
DUSO	-0.720	-2.664	0.008
HISTSTR	0.196	4.132	0.000
EDAD	-0.028	-1.987	0.047
CARNET	0.003	0.171	0.864
DESTAF	1.504	4.793	0.000
DCULPA	0.229	1.104	0.270
ANTIGUO	0.029	1.421	0.155
DCOMOCUR	1.324	6.590	0.000
ALTAHORA	1.366	5.199	0.000
CARRET	0.607	1.855	0.064
SABDOM	0.298	1.500	0.134
DEFOCEMI	0.729	0.696	0.486
STEST	-9.862	-15.206	0.000
DZONA1	0.751	2.286	0.022
DZONA3	1.036	4.763	0.000
DRELATO	0.623	3.154	0.002
SAUTDAD	-1.675	-3.600	0.000
Fraude a favor de un tercero:			
CONSTANTE	-4.010	-6.297	0.000
DRAMO	0.801	2.135	0.033
DUSO	-0.384	-1.325	0.185
HISTSTR	0.171	2.981	0.003
EDAD	-0.021	-1.486	0.137
CARNET	0.001	0.022	0.982
DESTAF	0.747	1.785	0.074
DCULPA	2.670	12.760	0.000
ANTIGUO	0.006	0.267	0.790
DCOMOCUR	1.335	6.511	0.000
ALTAHORA	1.084	3.618	0.000
CARRET	0.208	0.542	0.588
SABDOM	0.274	1.286	0.198
DEFOCEMI	0.163	0.166	0.868
STEST	2.315	2.160	0.031
DZONA1	-0.525	-1.277	0.202
DZONA3	0.491	2.347	0.019
DRELATO	0.486	2.466	0.014
SAUTDAD	-1.245	-2.834	0.004
número de observaciones	1611	Chi-cuadrado	573.399
ln función verosimilitud	-808.878	grados de libertad	36
ln verosimilitud restringida	-1095.577	nivel signif.	0.000

Los primeros diecinueve parámetros definen el vector de coeficientes estimados para la probabilidad de que el asegurado cometa fraude en beneficio propio. El segundo grupo de parámetros estimados se refiere a la probabilidad de que exista fraude en beneficio de un tercero. Ambas especificaciones usan las mismas variables, presentándose, para todas ellas, los estadísticos de significación individual y los valores de la probabilidad de la cola.

El test de la razón de verosimilitud rechaza la hipótesis de que todos los coeficientes (excepto las constantes) en las dos ecuaciones estimadas sean iguales a cero; el logaritmo neperiano de la verosimilitud decrece de -808.9 a -1095.6, mientras que se ganan 36 grados de libertad al imponer las restricciones de igualdad. El estadístico utilizado, comparable al clásico test de la F en el contexto del modelo de regresión, se calcula como:

$$2\left[\ln L\left(\hat{\beta}_g\right) - \ln L\left(\hat{\beta}_r\right)\right] \sim \chi_q^2$$

siendo, $\ln L\left(\hat{\beta}_g\right)$ el logaritmo neperiano de la verosimilitud del modelo general (modelo con todas las variables explicativas) y $\ln L\left(\hat{\beta}_r\right)$ el logaritmo neperiano de la verosimilitud del modelo restringido (modelo que contiene únicamente los términos independientes), y se distribuye asintóticamente según una distribución Chi-cuadrado con tantos grados de libertad como restricciones estemos contrastando (diferencia entre los grados de libertad de los dos modelos comparados). En nuestro caso, el estadístico (también conocido como estadístico Chi-cuadrado), toma un valor igual a 573.4 y su comparación con el valor en tablas para una distribución χ^2 con 36 grados de libertad permite aceptar la hipótesis alternativa de significación global del modelo al 1% de significación.

Los valores obtenidos muestran hechos importantes en relación a la probabilidad estimada de fraude. Como se desprende de la Tabla 3, la significación estadística de las variables es notablemente distinta al tener en cuenta ambos tipos de fraude. Mientras que en el primer grupo encontramos catorce coeficientes significativos (trece al 5% y uno al 10%) en el segundo sólo once lo son (diez al 5% y uno al 10 %). El hecho de que haya variables que no tengan poder para explicar la existencia de un determinado tipo de fraude y lo tengan para explicar el otro es, sin duda, un punto importante a tener en cuenta en la investigación de siniestros y la evidencia empírica a nuestro entender más deseable.

En la interpretación de los resultados, para ambos vectores de parámetros y por separado, debe recordarse que la categoría de referencia de la variable dependiente es la de los siniestros no fraudulentos.

Para aquellas variables creadas con el objetivo de recoger situaciones claramente sospechosas de fraude, el signo de la estimación máximo verosímil de los parámetros asociados es el esperado. Para el resto de variables, y sin ser capaces en algunos casos de extraer conclusiones en cuanto a signos esperados (el propio mecanismo asegurador dificulta, en algunos casos, obtener conclusiones sobre el comportamiento que se espera para algunos regresores en relación al fraude), los resultados son claramente significativos.

Comencemos por los términos independientes del modelo. La aparición de un coeficiente estadísticamente significativo (al 1%) y con signo negativo en el vector de parámetros asociado a la categoría de fraude a favor del asegurado permite señalar que, si el resto de variables introducidas en el modelo son nulas, es inferior la probabilidad de aparición de este tipo de comportamiento en relación al no fraude. Una conclusión análoga se extrae a partir del coeficiente para el término independiente de la categoría de fraude a favor del tercero.

Para las variables relacionadas con la póliza y el vehículo en el momento del contrato, cabe destacar como la existencia en la póliza de cobertura a terceros está positivamente relacionada con la aparición de los dos tipos de fraude considerados, siendo el coeficiente estimado para la variable *DRAMO* estadísticamente significativo (al 5%) en los dos vectores de parámetros. Cuando ocurre el siniestro, el asegurado que posee cobertura de daños propios no necesita cometer fraude para conseguir reparar, por ejemplo, los daños de su vehículo, mientras que con la cobertura a terceros las inquietudes del individuo por conseguir una indemnización se ven incrementadas.

El tipo de vehículo que posee el asegurado tiene diferente poder explicativo en los dos tipos de fraude considerados. Así, mientras que el coeficiente que acompaña a la variable *DUSO* es estadísticamente significativo (al 1%) en la categoría de fraude a favor del asegurado, deja de serlo en la categoría de fraude a favor del tercero. El signo negativo obtenido para el mismo permite señalar que la posesión de un turismo de uso particular disminuye tanto la probabilidad de que el asegurado cometa fraude en beneficio propio como para un tercero.

En relación a la variable que recoge el número de siniestros previos declarados en la compañía (dentro del conjunto de variables relacionadas con el asegurado en el momento del siniestro), los resultados son claramente significativos. El valor del coeficiente estimado, positivo y

estadísticamente significativo al 1% en ambos vectores de parámetros, indica que los asegurados con siniestros previos tienden a cometer fraude. La explicación a este hecho puede ser variada. Quizá el asegurado, ante la no detección por parte de la compañía de fraudes anteriores, decida reincidir creando nuevos siniestros. Quizá esté en la peor clase de *bonus-malus* y por lo tanto piense que su situación, aunque tenga un nuevo siniestro, no puede ser peor.

La edad del conductor asegurado sólo es significativa en la explicación del fraude cometido en beneficio propio. La aparición de un coeficiente con signo negativo permite señalar que la probabilidad de fraude para beneficio propio disminuye a medida que aumenta la edad del individuo. Los años de antigüedad de carnet del conductor asegurado ha resultado ser una variable no significativa en la explicación del fraude.

La coincidencia de apellidos entre la parte asegurada y la contraria (variable basada en la relación con el conductor contrario) está también directamente asociada con la probabilidad de aparición de ambos tipos de fraude. No obstante el hecho de que el coeficiente goce de mayor significación estadística en el primer vector de parámetros (para el segundo sólo podemos aceptar la significación al 10%), nos lleva a pensar en una mayor influencia de la variable en la aparición de fraude a favor del asegurado. Una justificación para este resultado puede encontrarse, por ejemplo, en aquellos casos en los que el asegurado, que posee cobertura a terceros y ha sufrido un accidente sin involucración de un contrario, conviene con un familiar que sea éste quien acepte la culpa del siniestro para así reparar sus daños.

Para el resto de variables, relacionadas con el siniestro y el vehículo en el momento del accidente, los resultados obtenidos son también relevantes.

La culpa del siniestro es una variable estadísticamente significativa en la explicación de fraude a favor del contrario y, sin embargo, no goza de significación en la categoría de fraude a favor del asegurado. El signo positivo que acompaña al coeficiente estimado en el segundo vector de parámetros permite señalar que si el asegurado admite la culpa del siniestro aumenta la probabilidad de que defraude a favor de un tercero. Si el asegurado posee cobertura a terceros (situación más frecuente) y admite la culpa del siniestro parece lógico pensar que sea para favorecer al contrario ya que no tiene cubiertos sus daños. No obstante, cabe suponer que este último compensará de una u otra forma la pérdida que experimenta el asegurado al entrar en la escala de *malus* (o descender en la de *bonus*). Lógicamente, si el asegurado posee cobertura de daños propios puede admitir su culpa para que, a la vez que repara sus daños, beneficie a un contrario que posee únicamente cobertura a terceros.

La antigüedad del vehículo en el momento del siniestro no tiene poder explicativo en la aparición de ninguno de los dos tipos de fraude considerados. Así lo indica el coeficiente estimado para dicha variable, que aún presentando signo positivo en ambos vectores de parámetros no reviste significación estadística. No obstante, el hecho de que el nivel de significación al que aceptaríamos la hipótesis alternativa (coeficiente individualmente significativo) sea notablemente inferior para el coeficiente que aparece en el primer vector, nos permite justificar lo que sería en principio un comportamiento esperado: de tener influencia, la antigüedad del vehículo del asegurado aparecerá relacionada con la existencia de fraude en beneficio propio.

El hecho de que el siniestro sea comunicado a la compañía con posterioridad a la primera semana desde su ocurrencia y el que haya ocurrido a altas horas de la noche son variables con coeficientes estadísticamente significativos y relacionados positivamente con la probabilidad de aparición de ambos tipos de fraude.

El primero de los anteriores coeficientes (relativo a la dilación en la comunicación del accidente a la entidad) puede interpretarse atendiendo al hecho de que, cuando no existe fraude, cabe esperar que el asegurado comunique el siniestro a su compañía con la mayor brevedad posible. La preparación de un fraude puede suponer una demora en la comunicación del accidente (indecisión sobre la posibilidad de comportarse fraudulentamente, diseño del plan a seguir con el contrario,...).

Cuando el siniestro ocurre a altas horas de la noche puede resultar más fácil cometer fraude. La ausencia de testigos y la posibilidad de que en el siniestro intervengan asegurados jóvenes (más propensos, como hemos visto, a cometer fraude) son algunos de los factores que pueden explicar el comportamiento observado para esta variable.

La ocurrencia del accidente en una zona de alta siniestralidad respecto a otras zonas (siniestralidad media o baja) incrementa la probabilidad de fraude en beneficio propio mientras que reduce la probabilidad de fraude a favor de un tercero. No obstante, para este último caso carece de significación estadística. Por otro lado, el acaecimiento del siniestro en zonas de baja siniestralidad (respecto a las de siniestralidad alta o media) presenta coeficientes positivos y estadísticamente significativos en ambos vectores de parámetros.

Los resultados obtenidos para la segunda variable considerada que nos llevan a señalar un incremento de la probabilidad de fraude en las zonas con siniestralidad baja, pueden sugerir

que la definición de zonas de siniestralidad utilizada quizá no sea la adecuada para discriminar comportamientos diferenciales.

La ocurrencia del siniestro en carretera aumenta la probabilidad de fraude a favor del asegurado. El coeficiente que acompaña a la variable *CARRET*, positivo, es significativo (no obstante, a un 10%) en la primera categoría de fraude, mientras que carece de significación estadística en la categoría de fraude para un tercero.

El hecho de que el siniestro suceda en fin de semana no tiene poder explicativo en la aparición de los dos tipos de fraude. Esta conclusión se observa también para el hecho de que el siniestro ocurra entre la fecha de efecto y la fecha de emisión de la póliza. La obtención de coeficientes no significativos para ambas variables justifica tales afirmaciones. No obstante y, a pesar de que los resultados no permiten evidenciar la relación de la segunda variable con la aparición de fraude a favor del asegurado, la suposición desde un punto de vista empírico de que el retroceso en la fecha de efecto puede venir provocado muchas veces por el deseo de beneficiar al asegurado (por parte de personal de la compañía y de los agentes), nos llevará, en el próximo Capítulo, a convertirla en un atributo específico de elección de esta segunda alternativa de fraude.

Algunos parámetros estimados sugieren que la influencia de las variables asociadas es opuesta en las dos ecuaciones. Así, la presencia de testigos disminuye la probabilidad de fraude en beneficio propio, mientras que está positivamente relacionada con la probabilidad de defraudar a favor del contrario. En este último caso cabe pensar en la existencia de testigos falsos que colaboran en la ejecución del fraude.

Por último, la presencia en la declaración del asegurado de los relatos sospechosos en el estudio de los siniestros fraudulentos[14] aumenta la probabilidad de fraude, tanto a favor del asegurado como del tercero. La intervención de policía en el siniestro disminuye la probabilidad para ambos tipos, resultado lógico si tenemos en cuenta que la presencia de autoridad y la elaboración de un atestado reduce notablemente el espacio para cometer fraude. Ambas variables revisten significación estadística en la explicación de los dos tipos de comportamiento.

La interpretación de los coeficientes, tal y como comentamos en el Capítulo 4, no es inmediata. El hecho de que la variación que se espera se produzca en la probabilidad de elegir

[14] Recordemos que éstos estaban relacionados con descripciones que incluían referencias a aparcamientos, adelantamientos, marcha atrás,...

una determinada alternativa al aumentar en una unidad una determinada variable explicativa no sea constante, sino que dependa de los valores que tome dicha variable y el resto de las explicativas incluidas en el modelo, hace que los coeficientes sólo sean directamente interpretables en términos de dirección de variación de la probabilidad.

Una solución se encuentra en el cálculo de los efectos marginales de los regresores en las probabilidades[15]. En nuestro caso el cómputo de los mismos ha derivado en la obtención de valores muy bajos. No obstante, el hecho que la mayoría de las variables explicativas consideradas sean dicotómicas hace que su interpretación sea complicada, teniendo en cuenta que las derivadas parciales de las probabilidades con respecto al vector de características para un determinado individuo son evaluadas en el vector de medias de las variables explicativas consideradas. Es por ello que obviamos su presentación.

Una vez estimados los coeficientes de la regresión logística multinomial, podemos realizar el cómputo de las probabilidades ajustadas para cada elección utilizando la expresión (4.11), presentada en el Capítulo 4.

De esta forma,

$$P(Y_{i0}=1)=\frac{1}{1+B_{i1}+B_{i2}}$$

$$P(Y_{i1}=1)=\frac{B_{i1}}{1+B_{i1}+B_{i2}}$$

$$P(Y_{i2}=1)=\frac{B_{i2}}{1+B_{i1}+B_{i2}}$$

donde,

$$\begin{aligned}B_{i1}=\exp(&-3.403+1.067\,DRAMO-0.720\,DUSO+0.196\,HISTSTR-0.028\,EDAD+0.003\,CARNET\\&+1.504\,DESTAF+0.229\,DCULPA+0.029\,ANTIGUO+1.324\,DCOMOCUR+1.366\,ALTAHORA\\&+0.607\,CARRET+0.298\,SABDOM+0.729\,DEFOCEMI-9.862\,STEST+0.751\,DZONA1\\&-1.036\,DZONA3+0.623\,DRELATO-1.675\,SAUTDAD)\end{aligned}$$

[15] Computados en el vector de medias de las variables explicativas, el cálculo de los efectos marginales se realiza mediante la siguiente expresión,

$$\frac{\partial P_{i(k)}}{\partial X_i}=P_{i(k)}\left[\beta_k-\sum_{r=1}^{2}P_{i(r)}\beta_r\right].$$

$$B_{i2} = \exp(-4.010 + 0.801DRAMO - 0.384DUSO + 0.171HISTSTR - 0.021EDAD + 0.001CARNET + 0.747DESTAF + 2.670DCULPA + 0.006ANTIGUO + 1.335DCOMOCUR + 1.084ALTAHORA + 0.208CARRET + 0.274SABDOM + 0.163DEFOCEMI + 2.315STEST - 0.525DZONA1 - 0.491DZONA3 + 0.486DRELATO - 1.245SAUTDAD).$$

Los parámetros ajustados pueden utilizarse directamente en el cómputo de los log-odds[16] entre probabilidades,

$$\ln\left[\frac{P_{i(1)}}{P_{i(0)}}\right] = \ln(B_{i1})$$

$$\ln\left[\frac{P_{i(2)}}{P_{i(0)}}\right] = \ln(B_{i2})$$

de forma que, a partir de los cocientes entre probabilidades planteados, es posible realizar una interpretación adicional de los coeficientes de la regresión. Así, por ejemplo, el parámetro 0.196 asociado a la variable *HISTSTR* ("número de siniestros asociados a la póliza anteriores al estudiado") indica que cuando el valor de la misma aumenta en una unidad, manteniéndose constantes el resto de variables explicativas, el cociente entre la probabilidad de que exista fraude a favor del asegurado y la probabilidad de que el siniestro sea legítimo aumenta en 1.216 (= exp(0.196)). Del mismo modo, el coeficiente -1.245 asociado a la variable *SAUTDAD* ("existe intervención policial en el siniestro") indica que cuando el valor de dicha variable cambia de 0 a 1, manteniéndose constantes el resto de explicativas, el cociente entre la probabilidad de que exista fraude a favor de un tercero y la probabilidad de que el siniestro sea legítimo se multiplica por 0.288 (= exp(-1.245)), es decir, disminuye.

[16] El modelo lógit multinomial aplicado permite el cálculo de j (=1,2) log-odds según la expresión,

$$\ln\left[\frac{P_{i(j)}}{P_{i(0)}}\right] = \beta_j' X_i .$$

Asimismo,

$$\ln\left[\frac{P_{i(j)}}{P_{i(k)}}\right] = X_i'(\beta_j - \beta_k).$$

6.3.1.2 Independencia de Alternativas Irrelevantes

Como vimos en el Capítulo 4, desde el punto de vista de la estimación del modelo ha de ocurrir que el cociente entre las probabilidades de elegir dos alternativas cualquiera sea independiente de los atributos o de la disponibilidad de una tercera alternativa de elección. Este hecho, comúnmente denominado Independencia de Alternativas Irrelevantes, viene fundamentado en la hipótesis de independencia para los términos de error en el modelo original (independencia entre los términos de error de las funciones de utilidad planteadas[17]), y puede contrastarse mediante la aplicación del test de Hausman-McFadden (1984)[18].

Para nuestro caso, los resultados del test son 0.7 y 1.0, respectivamente, cuando eliminamos cada tipo de fraude. El contraste con una distribución Chi-cuadrado, en este caso con 19 grados de libertad, permite señalar que no podemos rechazar la hipótesis nula de Independencia de Alternativas Irrelevantes, incluso al 1% de significación y por lo tanto, los cocientes planteados, $P_{i(1)}/P_{i(0)}$ y, $P_{i(2)}/P_{i(0)}$ permanecen constantes cuando la tercera y la segunda alternativa no son tenidas en cuenta, respectivamente.

El cumplimiento de la hipótesis planteada se percibe si comparamos el vector de parámetros estimados obtenido al considerar todas las elecciones y el obtenido al eliminar la categoría de fraude a favor del tercero y fraude a favor del asegurado, respectivamente. Tomemos como referencia los resultados presentados en la Tabla 3, donde aparecían los vectores de coeficientes estimados para la categoría de fraude a favor del asegurado y de fraude a favor de un tercero. Al eliminar la categoría de fraude a favor de un tercero (el tamaño muestral pasa a ser de 1297 observaciones) y estimar el modelo, el vector de coeficientes para la categoría de fraude a favor del asegurado es el siguiente (se adjunta el estadístico t asintótico y el P-valor):

---

[17] Ver Capítulo 4.

[18] El estadístico es igual a:

$$T = (\hat{\beta}_r - \hat{\beta}_c)'(\hat{V}_r - \hat{V}_c)^{-1}(\hat{\beta}_r - \hat{\beta}_c).$$

donde $\hat{\beta}_c$ y $\hat{\beta}_r$ son los vectores de parámetros estimados obtenidos al aplicar Máxima Verosimilitud en el conjunto de elección no restingido y restringido, respectivamente, y $\hat{V}_c$ and $\hat{V}_r$ son las matrices de covarianzas asintóticas estimadas correspondientes a ambos conjuntos de elección. Bajo la hipótesis nula, el estadístico se distribuye asintóticamente según una Chi-cuadrado con tantos grados de libertad como rango posea $(\hat{V}_r - \hat{V}_c)$.

Tabla 4. Resultados de la estimación de un modelo logístico
(eliminando la alternativa de fraude a favor de un tercero)

variable	coeficiente	t-test	P-valor
CONSTANTE	-3.383	-4.898	0.000
DRAMO	1.118	2.399	0.016
DUSO	-0.699	-2.446	0.014
HISTSTR	0.190	3.908	0.000
EDAD	-0.028	-1.822	0.068
CARNET	0.002	0.090	0.928
DESTAF	1.453	4.321	0.000
DCULPA	0.194	0.858	0.390
ANTIGUO	0.030	1.316	0.188
DCOMOCUR	1.288	6.046	0.000
ALTAHORA	1.339	4.750	0.000
CARRET	0.545	1.560	0.119
SABDOM	0.285	1.318	0.188
DEFOCEMI	0.880	0.836	0.403
STEST	-9.859	-18.046	0.000
DZONA1	0.726	2.113	0.035
DZONA3	1.016	4.423	0.000
DRELATO	0.586	2.806	0.005
SAUTDAD	-1.671	-3.453	0.000
número de observaciones	1297	Chi-cuadrado	203.016
ln función verosimilitud	-377.400	grados de libertad	18
ln verosimilitud restringida	-478.908	nivel signif.	0.000

Como puede observarse, los resultados no difieren significativamente de los obtenidos al estimar el modelo con las 1611 observaciones y por lo tanto se intuye la independencia de alternativas.

Un resultado análogo puede obtenerse al eliminar la categoría de fraude a favor del asegurado. En este caso los resultados obtenidos de la estimación (realizada con un total de 1312 observaciones), son los siguientes:

Tabla 5. Resultados de la estimación de un modelo logístico
(eliminando la alternativa de fraude a favor del asegurado)

variable	coeficiente	t-test	P-valor
CONSTANTE	-3.970	-5.760	0.000
DRAMO	0.714	1.803	0.071
DUSO	-0.468	-1.467	0.142
HISTSTR	0.159	2.333	0.020
EDAD	-0.019	-1.241	0.214
CARNET	-0.004	-0.190	0.849
DESTAF	0.973	2.173	0.030
DCULPA	2.689	11.965	0.000
ANTIGUO	0.012	0.534	0.593
DCOMOCUR	1.435	6.341	0.000
ALTAHORA	1.485	3.470	0.000
CARRET	0.299	0.704	0.481
SABDOM	0.263	1.131	0.258
DEFOCEMI	0.226	0.258	0.796
STEST	2.351	2.075	0.038
DZONA1	-0.396	-0.869	0.385
DZONA3	0.493	2.170	0.030
DRELATO	0.492	2.263	0.024
SAUTDAD	-1.242	-2.599	0.009
número de observaciones	1312	Chi-cuadrado	360.406
ln función verosimilitud	-317.876	grados de libertad	18
ln verosimilitud restringida	-498.079	nivel signif.	0.000

En este caso y únicamente para el coeficiente que acompaña a la variable *CARNET*, se observa una diferencia en cuanto a comportamiento en signo, aunque éste no tiene significación estadística. El resto de coeficientes son muy similares a los obtenidos con el total de la muestra.

Para terminar, cabe señalar que en la demostración de la independencia de alternativas ha sido necesario realizar una corrección de los pesos introducidos en el modelo, teniendo en cuenta, en cada caso, el nuevo tamaño muestral y el hecho de trabajar con una elección menos.

6.3.1.3 Análisis de la calidad del ajuste

En los modelos de elección probabilística es frecuente utilizar como medida de calidad del ajuste el porcentaje de casos correctamente clasificados por el modelo, teniendo en cuenta las alternativas planteadas. Tras la aplicación del lógit multinomial, el cómputo de las probabilidades predichas por el modelo para cada una de las observaciones supondrá una clasificación del siniestro en aquella categoría para la que se obtiene una mayor

probabilidad[19]. La comparación, para cada observación, entre la clasificación real u observada y la predicha queda sintetizada en la siguiente tabla,

Tabla 6. Frecuencias de clasificación (utilizando el criterio de máxima probabilidad)

	Elección predicha			Total
	Legítimo	Beneficio propio	Beneficio tercero	
Elección Observada				
Legítimo	961	12	25	998
Beneficio propio	222	37	40	299
Beneficio tercero	177	9	128	314
Total	1360	58	193	1611

Los resultados obtenidos permiten señalar que 1126 siniestros han sido clasificados en el grupo correcto. De esta forma el porcentaje de clasificación correcta es del 70%. Sin embargo, y a pesar de que el porcentaje de siniestros sin fraude detectado correctamente clasificados como tales es elevado (96.3%), no podemos decir lo mismo para los siniestros fraudulentos. Así, tan sólo un 12.4 % y un 40.8% de los siniestros han sido correctamente clasificados en las categorías de fraude a favor del asegurado y fraude a favor de un tercero, respectivamente.

En relación a los siniestros clasificados de forma errónea, los resultados no son buenos. Las frecuencias obtenidas reflejan una elevada proporción de casos predichos por el modelo como no fraudulentos. De esta forma 222 expedientes con fraude observado a favor del asegurado y 177 con fraude observado a favor de un tercero son clasificados por el modelo como no fraudulentos. Lógicamente y, desde este punto de vista, la capacidad predictiva del modelo planteado es baja, dado que tras su aplicación, la entidad dejaría de detectar un 74.2% de siniestros de la primera tipología de fraude y un 56.4% de la segunda. El porcentaje de siniestros sin fraude detectado clasificados por el modelo como fraudulentos, y el de siniestros fraudulentos de un tipo clasificados en la categoría alternativa de fraude no son elevados. Además, y para el último caso, las consecuencias de una clasificación incorrecta no son tan preocupantes, pues en ambas se detecta fraude (lógicamente, ésto se podría matizar al tener en cuenta los costes derivados de investigar uno u otro tipo de fraude).

La obtención de unos resultados como los presentados para los siniestros fraudulentos (de uno u otro tipo) puede venir provocada por el hecho de que la muestra utilizada, en relación a los mismos, no sea suficientemente representativa. El hecho de que en la alternativa de no fraude

[19] Lógicamente, la suma de las tres probabilidades predichas por el modelo para cada una de las alternativas ha de sumar la unidad.

dispongamos de un número de observaciones notablemente superior al de las alternativas de fraude puede ser una de las razones para que el modelo cometa un mayor error en la clasificación de los siniestros fraudulentos. Sin embargo, los resultados de clasificación presentados pueden mejorar notablemente al optimizar el criterio probabilístico utilizado para clasificar un siniestro dentro de una determinada categoría.

Como veíamos en el Capítulo 4 (la optimización del punto de corte en un modelo lógit multinomial constituye, sin duda, una de las principales aplicaciones de la Tesis Doctoral y su fundamentación teórica ha sido presentada en el apartado 4.4) el punto de corte o probabilidad utilizada por defecto en la delimitación de las zonas de clasificación de los siniestros en cada una de las tres categorías presentadas es de 1/3. La selección del baricentro presentado en el Gráfico 1 (ver Capítulo 4), delimita áreas de aceptación para cada una de las alternativas con idénticas dimensiones. Sin embargo, tras la aplicación del modelo, los resultados obtenidos en términos de calidad del ajuste pueden sugerir la conveniencia de modificar dicho criterio probabilístico, fundamentalmente, al tener en cuenta la elevada presencia de siniestros de una determinada categoría clasificados dentro de otra.

En un lógit multinomial con tres alternativas las probabilidades predichas por el modelo pueden representarse gráficamente en tres dimensiones. Para el modelo analizado los resultados obtenidos aparecen en los gráficos siguientes,

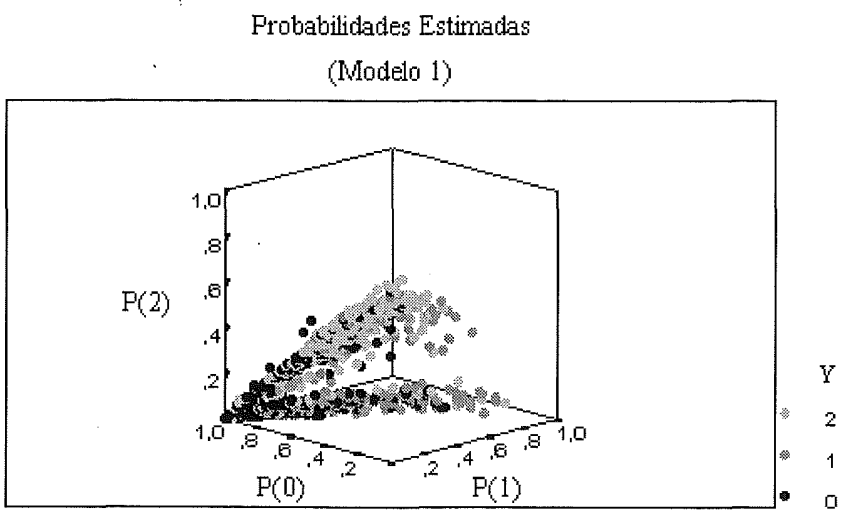

Figura 11

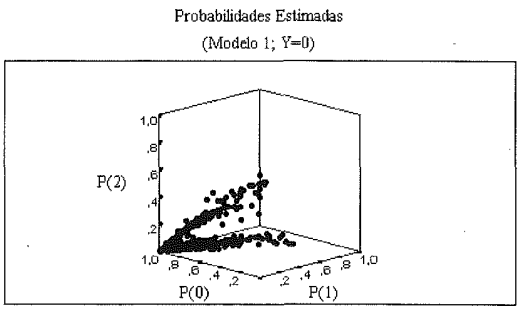

Figura 12

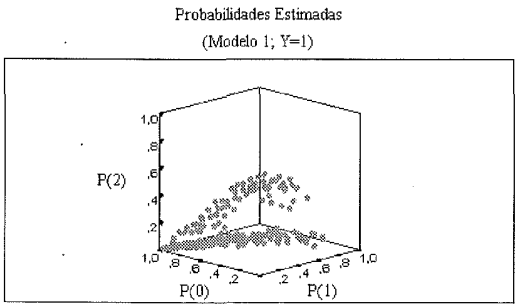

Figura 13

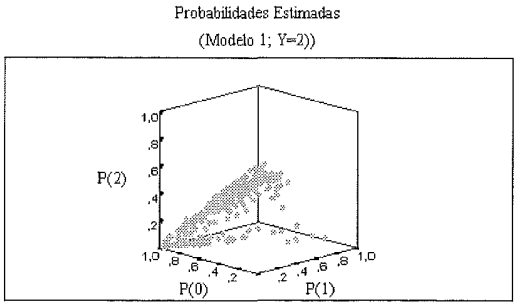

Figura 14

En la Figura 11 se representan de forma conjunta las probabilidades predichas por el modelo para cada observación teniendo en cuenta las tres categorías definidas para la variable dependiente. En las Figuras 12, 13 y 14 y con el objetivo de facilitar la interpretación, se presenta el estudio individualizado de dichas probabilidades para el conjunto de siniestros sin fraude detectado (Y=0); para los siniestros con fraude a favor del asegurado (Y=1) y para los siniestros con fraude a favor del tercero (Y=2), respectivamente.

La distribución observada para las probabilidades ajustadas, aproximadamente en forma de triángulo, está fundamentada en el desarrollo teórico presentado en el apartado 4.4, siendo posible establecer una analogía entre las representaciones gráficas ahora obtenidas y la que aparece en el Gráfico 1 (realizada, esta última, en dos dimensiones). La correspondencia entre las zonas de clasificación de los siniestros permite observar una concentración de probabilidades predichas en la zona de no fraude (área inferior izquierda), no sólo para los casos en los que se observa ausencia de fraude (resultado esperado), sino también para aquellos en los que existe fraude detectado (sobre todo en relación a la primera categoría).

Ante este resultado, que no hace sino confirmar el obtenido en la Tabla 6 , la optimización del punto de corte, destinada fundamentalmente a reducir el porcentaje probabilístico que permite calificar al siniestro como fraudulento (ampliación de las zonas de aceptación de ambos tipos de fraude) ha permitido obtener la siguiente clasificación de siniestros, estando el punto de corte fijado en el vector de probabilidades (0.15, 0.15, 0.15).

Tabla 7. Frecuencias de clasificación (utilizando el criterio c=0.15)

	Elección predicha			
	Legítimo	Beneficio propio	Beneficio tercero	Total
Elección Observada				
Legítimo	753	111	134	998
Beneficio propio	104	129	66	299
Beneficio tercero	48	34	232	314
Total	905	274	432	1611

El porcentaje de siniestros correctamente clasificados dentro de sus categorías respectivas es ahora del 69.1%. Un total de 753 siniestros sin fraude detectado han sido clasificados de forma adecuada por el modelo, de forma que el porcentaje de aciertos para la primera alternativa es del 75.5%. En relación a las alternativas de fraude, para los siniestros en los que se observa fraude a favor del asegurado, el porcentaje de aciertos es del 43.1% mientras que

para aquellos en los que se observa fraude a favor de un tercero dicho porcentaje es superior, del 73.9%.

La comparación con los resultados que aparecen en la Tabla 6 muestran, sin duda, una notable mejora. Mientras que el porcentaje total de siniestros clasificados de forma correcta prácticamente no sufre variación, el incremento de aciertos para los casos de fraude es una señal de la mayor capacidad predictiva del modelo en relación a este tipo de comportamiento.

¿Qué ocurre con los siniestros clasificados de forma errónea?

Como veíamos en páginas anteriores, el error cometido al clasificar como no fraudulentos siniestros en los que existe fraude detectado es especialmente preocupante para la compañía ya que, en esos casos, al no realizarse la investigación pertinente, resulta difícil detectar comportamientos deshonestos. Los porcentajes que para este error de clasificación observábamos en la Tabla 6 son ahora muy inferiores, pasando a ser del 34.8% para el primer tipo de fraude y del 15.3 para el segundo, advirtiéndose por lo tanto una mejora importante. Un 22.1% de los siniestros con fraude a favor del asegurado son clasificados de forma errónea como fraudes a favor del tercero mientras que un 10.8% de los siniestros con fraude a favor del contrario son clasificados como fraudes a favor del asegurado. No obstante, el hecho de que la compañía continúe detectando fraude (lógicamente esta conclusión podría matizarse si tuviéramos en cuenta la existencia de costes de investigación muy diferentes para cada tipo de comportamiento fraudulento) hace que estos resultados no sean *a priori* tan preocupantes. Por último, un 24.5% de siniestros sin fraude detectado han sido clasificados como fraudulentos por el modelo (111 expedientes sin fraude son clasificados dentro de la primera categoría de fraude y 134 dentro de la segunda) siendo el aumento innecesario en los costes de investigación la principal desventaja asociada a este error de clasificación.

Los resultados presentados en la Tabla 7 permiten señalar que el modelo lógit multinomial, tras la optimización del punto de corte, proporciona una técnica de predicción aceptable, fundamentalmente en lo que a siniestros fraudulentos se refiere. Además, y dado que para los siniestros sin fraude el porcentaje de aciertos es también muy elevado podemos ratificarnos en la afirmación de que la aplicación del Modelo 1 ayudará a la compañía a realizar una correcta clasificación de siniestros, teniendo en cuenta la tipología de fraude presentada.

No obstante, para la entidad aseguradora podría resultar interesante el planteamiento de un escenario adicional. Así, si asumimos que los costes de oportunidad en tiempo y los costes monetarios asociados a la investigación de siniestros no son elevados (la compañía está

dispuesta a aceptar un mayor error en la clasificación de siniestros no fraudulentos como fraudulentos), podemos mejorar aún más la capacidad del modelo en términos de predicción de fraude. De esta forma, la selección del punto de corte (0.1, 0.1, 0.1) permite obtener los resultados que aparecen a continuación,

Tabla 8. Frecuencias de clasificación (utilizando el criterio c= 0.10)

	Elección predicha			
	Legítimo	Beneficio propio	Beneficio tercero	Total
Elección Observada				
Legítimo	642	203	153	998
Beneficio propio	63	167	69	299
Beneficio tercero	25	49	240	314
Total	730	419	462	1611

observándose unos porcentajes de clasificación correcta para los siniestros con fraude a favor del asegurado y del tercero, del 56% y 76%, respectivamente. El número de siniestros fraudulentos clasificados erróneamente en la categoría de no fraude es muy bajo (63 casos para el primer tipo de fraude y 25 para el segundo). El porcentaje global de casos correctamente clasificados es ahora del 65.1% demostrándose, de nuevo, una elevada capacidad predictiva del modelo.

Los resultados obtenidos de la optimización del punto de corte evidencian la importancia del proceso dentro de la modelización logística multinomial. Así, hemos podido demostrar como, al optimizar el criterio probabilístico utilizado en la definición de las zonas de aceptación de cada una de las alternativas, el porcentaje de fraudes detectados ha aumentado de forma muy importante. Con todo ello, la compañía puede encontrar en el Modelo 1 una buena herramienta, no sólo para analizar qué variables tienen mayor poder explicativo en la aparición de fraude a favor del asegurado y de fraude a favor del tercero, sino también para saber cuál sería el margen de error cometido en la detección de este tipo de fraudes dentro de su cartera. La comparación entre el coste de investigación de aquellos siniestros que el modelo clasifica como fraudulentos y los beneficios finalmente derivados de la detección de ambos tipos de fraudes permitirá a la entidad extraer las primeras conclusiones en relación a las ventajas o, en su caso desventajas, de instalar un mecanismo de control.

Una vez realizada la validación del Modelo 1, determinada la probabilidad estimada de existencia de cada tipo de comportamiento y cuantificada la probabilidad de clasificación correcta e incorrecta de cada tipo de siniestro, hemos dado respuesta al primer bloque de

objetivos que presentábamos en el Capítulo 3. Ahora bien, los resultados obtenidos pueden utilizarse, tal y como comentábamos en el mencionado Capítulo, en la consecución de dos objetivos finales: por un lado, estimar la utilidad esperada por el individuo al elegir cada una de las alternativas y , por otro, cuantificar el coste por siniestro esperado por la compañía a partir de la aplicación de un modelo, como el presentado, de control de fraude. El planteamiento de escenarios nos ayudará, en los siguientes apartados, a analizar las principales conclusiones para ambos, teniendo siempre en cuenta las alternativas de elección presentadas y el tipo de modelo que hemos utilizado.

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

Como veíamos en el Capítulo 3, la estimación de la utilidad que el asegurado espera de su comportamiento en términos de fraude puede realizarse teniendo en cuenta una serie de parámetros básicos. Estableciendo una analogía con la formulación presentada por Hoyt (1990) y Picard (1996), la obtención de la información necesaria para estimar la utilidad esperada por el individuo supondrá tener en cuenta básicamente dos conceptos. Por un lado, habremos de cuantificar la probabilidad de que el asegurado elija una determinada alternativa y, dado que en su decisión intervendrá la información que posee en relación a la mayor o menor probabilidad de que la compañía audite el siniestro, necesitaremos estimar la probabilidad de que se realice auditoría o investigación de los casos.

Al realizar la estimación del Modelo 1 hemos conseguido cuantificar la probabilidad de que el asegurado decida no defraudar, de que cometa fraude en beneficio propio o de que cometa fraude en beneficio de un tercero. De esta forma disponemos de información relativa al primero de los conceptos mencionados y la podremos utilizar directamente en la estimación de la utilidad esperada.

La probabilidad de que la compañía audite el siniestro estará determinada por los resultados obtenidos tras la aplicación del modelo en relación a la existencia de fraude: sólo si la probabilidad estimada supera el criterio probabilístico establecido como punto de corte se realizará la investigación del siniestro en cuestión. Tras la aplicación del mecanismo de control, la entidad será capaz de detectar un elevado número de siniestros fraudulentos de uno y otro tipo (ver Tabla 7), pero cometerá error en la clasificación de determinados expedientes. Cuando existiendo fraude la compañía no realiza ningún tipo de investigación, la utilidad esperada por el individuo que actúa fraudulentamente se ve incrementada. Para el resto de siniestros erróneamente clasificados la auditoría alcanzará los objetivos deseados, pues tras la

investigación supondremos que se detectará el tipo real de fraude existente o se confirmará la no aparición de comportamiento fraudulento.

La existencia de diferentes alternativas quedará reflejada en la especificación de la función de utilidad. Si el objetivo fuera analizar la utilidad que el individuo espera al actuar fraudulentamente frente a un comportamiento honesto bastaría con agrupar los resultados obtenidos para los dos tipos de fraude considerados. De esta forma, la probabilidad de cometer fraude vendría determinada por la suma de la probabilidad asociada al primer tipo de comportamiento fraudulento más la probabilidad asociada al segundo. Asimismo, la probabilidad de que la compañía clasificase un siniestro fraudulento en cualquiera de los tipos de fraude (probabilidad de detección de fraude) vendría determinada por el cociente entre el número de siniestros fraudulentos correctamente clasificados en sus categorías respectivas más aquellos clasificados incorrectamente en las categorías alternativas y el número total de siniestros fraudulentos observados. Una aplicación de este escenario puede encontrarse en Artís, Ayuso y Guillén (1997).

Sin embargo, cuando el objetivo es determinar la utilidad esperada por el individuo teniendo en cuenta la elección entre tres alternativas (sin realizar, por lo tanto, una agrupación previa entre las dos categorías de fraude) se ha de tener en cuenta un amplio abanico de probabilidades. Éstas estarán relacionadas no sólo con la probabilidad de que el individuo elija una opción determinada sino también, con la probabilidad de que la compañía detecte o no su comportamiento.

¿Cómo podemos determinar la utilidad esperada por el asegurado al cometer fraude en beneficio propio?.

A partir de los resultados presentados en la Tabla 7[20] podemos determinar q_1, definida como la probabilidad de que la compañía detecte la existencia de fraude a favor del asegurado y que estimamos que es igual a 0.431 ya que q_1=129/299. La probabilidad de que la compañía no detecte este tipo de comportamiento, $(1-q_1)$, será por lo tanto de 0.569. Sin embargo, este último porcentaje ha de ser objeto de corrección al tener en cuenta que, en determinadas situaciones, el individuo también será penalizado dado que el siniestro será clasificado erróneamente dentro de la categoría de fraude a favor del tercero (q_1'=66/299= 0.221). La probabilidad de detección será, por tanto, q_1+q_1'= 0.652, siendo (1-0.652) la probabilidad de que el fraude considerado no sea detectado. Suponemos que clasificarlo en una categoría

[20] De forma análoga podrían utilizarse los resultados presentados en la Tabla 8.

errónea de fraude implica que al final se identificará el comportamiento deshonesto debido a que se iniciará una investigación. Es importante destacar que hemos estimado la probabilidad de detección sin tener en cuenta que ésta puede ser diferente para los distintos individuos, ya que no sabemos si los errores de clasificación son mayores o menores en función de ciertas características del expediente. Aunque ésta es una importante objeción al desarrollo presentado, continuamos con esta hipótesis hasta el final del apartado.

Dada W (riqueza inicial del individuo), P (prima pagada por la cobertura) y s (indemnización derivada del siniestro fraudulento[21]) y, suponiendo que la penalización para el asegurado ante la realización de cualquier tipo de fraude es siempre la misma y supone la expulsión de la compañía[22] (cuantificada, por lo tanto, en base a la nueva prima que ha de pagar en otra compañía para gozar de un seguro que es obligatorio, P'), nuestro modelo proporciona una estimación de la utilidad que el asegurado espera al cometer fraude a favor de sí mismo,

$$E[U(1)]= [0.652U(W-P-P') + 0.348U(W-P+s)].$$

En relación a la utilidad esperada por el asegurado que decide cometer fraude en beneficio del tercero, el proceso de cálculo es análogo al anterior. La probabilidad de que la compañía detecte dicho comportamiento es ahora 0.739 ya que q_2=232/314, y a la misma habremos de adicionar la probabilidad de que la compañía detecte, de forma errónea, fraude a favor del asegurado (q_2'= 34/314= 0.108) pues en tal caso supondremos que también penalizará al individuo porque lo detectará. La probabilidad de que el fraude no sea detectado será por lo tanto de 0.153:

$$E[U(2)]= [0.847U(W-P-P') + 0.153U(W-P+s)].$$

Si el objetivo es cuantificar la utilidad que el individuo espera de su comportamiento teniendo en cuenta, de forma conjunta, la existencia de tres posibilidades de elección, será necesario adicionar el resultado proporcionado por el modelo en términos de utilidad esperada para cada una de las situaciones, teniendo en cuenta la probabilidad estimada de que el individuo elija cada una de las alternativas.

---

[21] Dicha indemnización puede ser simplemente la que corresponde por contrato (en caso de la creación de un siniestro o de la falsedad de datos para conseguir cobertura en situaciones excluidas en la póliza) o la misma pero incrementada (la ocurrencia legítima del siniestro es utilizada para conseguir una indemnización mayor a la que corresponde en base al tipo de seguro contratado).

[22] Asumimos que el castigo impuesto no varía de un individuo a otro. Las compañías tienden a aplicar una regla común en orden a proteger los derechos de los demandantes honestos. Normalmente es la expulsión del asegurado de la entidad aunque podría ocurrir que la penalización impuesta consistiera únicamente en la devolución de la parte de indemnización no justificada.

Para cada individuo y dadas sus características, el modelo genera estimaciones puntuales de t_0 (probabilidad estimada de que el individuo se comporte honestamente), de t_1 (probabilidad estimada de que el asegurado cometa fraude en beneficio propio) y de t_2 (probabilidad estimada de que cometa fraude a favor del tercero). Teniendo en cuenta los resultados anteriormente presentados en relación a la utilidad esperada para cada uno de los comportamientos fraudulentos, la utilidad esperada del comportamiento del asegurado vendrá dada por la expresión (el subíndice *i* relativo al individuo no ha sido considerado, para simplificar la notación):

$$E[U]= t_1[0.652U(W-P-P')+0.348U(W-P+s)]+t_2[0.847U(W-P-P')+0.153U(W-P+s)] \\ +t_0\, U(W-P+s_0).$$

siendo s_0 la indemnización a percibir por el asegurado derivada de un comportamiento honesto (normalmente se cumplirá la relación $s_0 \leq s$).

La sustitución, en la expresión anterior, de las probabilidades estimadas por el modelo de elegir cada una de las alternativas (t_j, con $j=0,1,2$), y de los valores fijados para W, P, P', s y s_0, permitirá cuantificar el resultado esperado por el asegurado tras su actuación.

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

Las probabilidades estimadas de elección de cada una de las alternativas y de clasificación de los siniestros pueden utilizarse, asimismo, en la estimación del coste que la compañía espera del siniestro, tras la aplicación de un mecanismo de control de fraude.

Una posible solución a la estimación del coste esperado tras la materialización del riesgo, teniendo en cuenta la existencia de fraude, vendría por agrupar los resultados obtenidos (probabilidades de elección y de clasificación de siniestros) tras la aplicación del lógit multinomial, para las dos categorías de fraude consideradas. De esta forma acabaríamos cuantificando el coste esperado final teniendo en cuenta el coste esperado para los siniestros no fraudulentos y para los fraudulentos, estos últimos, genéricamente considerados.

Sin embargo, la estimación del coste esperado teniendo en cuenta la existencia de tres alternativas finales de elección constituye un proceso más elaborado, tal y como veremos a continuación.

Retomemos 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 a favor del asegurado. Para todas estas probabilidades disponemos de alguna estimación mediante los resultados que proporciona el modelo.

Supongamos que el coste que tiene que asumir la compañía resulta de la adición de dos componentes: la indemnización finalmente pagada más los costes asociados a la investigación del siniestro. De esta forma, y para cada una de las situaciones, el coste esperado del siniestro queda recogido en la Tabla 9.

Tabla 9. Coste esperado del siniestro

	Clasificado como		
	Legítimo	Beneficio propio	Beneficio tercero
Tipo de siniestro			
Legítimo	s_0	c_1+s_0	c_1+s_0
Beneficio propio	s	c_2	c_2
Beneficio tercero	s	c_2	c_2

Donde s es el coste del accidente que debe ser pagado de acuerdo con la póliza contratada (s_0, es el coste real del siniestro), c_1 es el coste de investigar un siniestro sospechoso que no es fraudulento (legítimo) y, c_2 es la cantidad gastada en identificar un fraude.

En relación a la cantidad pagada en concepto de indemnización y, de cara a simplicar la interpretación de la Tabla 9, hemos supuesto que es la misma para los dos tipos de fraude (s). Las situaciones, sin embargo, pueden ser diversas. La aplicación de los convenios C.I.D.E.-A.S.C.I.D.E., implica que en algunos casos el coste del siniestro sea únicamente la cantidad pagada por *módulo*. La diferencia puede ser importante. Así, por ejemplo, si el asegurado, que posee cobertura a terceros, acepta la culpa del siniestro para beneficiar al contrario, el coste del siniestro para la compañía en concepto de indemnización será únicamente de 90000 pts. Si es el contrario el que acepta la culpa para beneficiar al asegurado, el coste para la compañía será el asociado a la reparación de los daños (previo recobro del *módulo* de la compañía contraria). Generalizando, la consideración del tipo de cobertura existente en la póliza, la identificación de la culpa del siniestro, la aplicación o no de convenios,..., puede derivar en indemnizaciones de diferente cuantía para la entidad.

Sin embargo y, dado que nuestro objetivo es presentar solamente una aproximación al tratamiento de los costes[23], hemos creído conveniente simplificar las hipótesis y diferenciar únicamente entre el coste del siniestro sin fraude y del siniestro con fraude. Aunque podrían aparecer situaciones en las que la existencia de comportamiento fraudulento implicase un menor coste para la entidad (en el ejemplo planteado en primer lugar en el párrafo anterior la entidad únicamente ha de pagar el *módulo*), la aparición de fraude suele representar, en la mayoría de los casos, un coste para la compañía superior al esperado ($s \geq s_0$).

En relación a los costes de investigación, la simplificación es también evidente. Así, hemos supuesto que el coste de identificar como legítimo un siniestro con sospecha de fraude a favor del asegurado o con sospecha de fraude a favor de un tercero, es el mismo y toma el valor c_1. Siguiendo el mismo criterio, el coste de detectar cualquiera de los tipos de fraude es siempre c_2. Tomando en consideración ambos costes, podríamos esperar una relación del tipo $c_1 \leq c_2$. Posiblemente, la entidad invertirá mayores cantidades monetarias y su coste de oportunidad en tiempo será superior cuando persiga identificar la existencia de fraude en el siniestro, a diferencia de aquellos casos en los que, durante el proceso de investigación, no tiene suficientes pruebas para sospechar de la existencia de comportamiento fraudulento.

Teniendo en cuenta las consideraciones anteriores, el coste total esperado del siniestro, $E(T)$, puede escribirse como:

$$E(T)= t_1[q_1c_2 + q_1'c_2 + (1-(q_1+q_1'))s] + t_2[q_2c_2 + q_2'c_2 + (1-(q_2+q_2'))s] + t_0[q_0s_0 + q_0'(s_0+c_1) + (1-(q_0+q_0'))(s_0+c_1)].$$

Esta expresión es equivalente a:

$$E(T)= t_1[(q_1+q_1')(c_2-s)+s] + t_2[(q_2+q_2')(c_2-s)+s] + t_0[q_0s_0+(1-q_0)(s_0+c_1)].$$

Nuestro modelo provee un método para estimar la probabilidad de elección de cada tipo de comportamiento (t_1, t_2 y t_0) dadas las características del siniestro y/o del individuo. Como los parámetros de coste (s_0, s, c_1 y c_2) son fijos y conocidos y las probabilidades de la compañía clasifique correcta (q_0, q_1 y q_2) y erróneamente (q_0', q_1' y q_2') a los siniestros en su categoría correspondiente pueden inferirse de los resultados del Modelo, es posible estimar el coste esperado del siniestro.

[23] La no disponibilidad, en la muestra original, de información completa en relación a los costes asociados a los siniestros con fraude detectado (diferencia entre la cuantía del siniestro y la cantidad finalmente pagada, si es que se paga alguna cantidad) ha impedido realizar un estudio más exhaustivo de este tema.

Retomemos, para acabar este apartado, los resultados que aparecen en la Tabla 7 y planteemos un escenario sencillo.

La probabilidad de clasificar un siniestro legítimo como tal es 0.755 ya q_0= 753/998, siendo 0.245 la probabilidad de clasificarlo incorrectamente como fraudulento. Como ya sabemos, para los siniestros con fraude a favor del asegurado el porcentaje de aciertos es 0.431, mientras que la probabilidad de clasificarlos como fraudes a favor del tercero es 0.221. Además la probabilidad de clasificar de forma correcta un siniestro con fraude a favor del tercero es 0.739. La probabilidad de clasificarlo como fraude a favor del asegurado es 0.108.

Sea s_0 el coste real del siniestro y fijemos que s=1.5s_0 es la indemnización pagada por la entidad atendiendo a la declaración presentada por el asegurado. Si asumimos que c_1= 0.10s_0 y, c_2= 0.15s_0, el coste total esperado estimado del siniestro es igual a:

$$E(T)= [t_1(0.62s_0) + t_2(0.36s_0) + t_0(1.02s_0)].$$

Si asumimos que la probabilidad estimada de que un asegurado cometa fraude a favor de sí mismo (t_1) es 0.44 y la de que cometa fraude a favor del tercero (t_2) es 0.30, entonces, en término medio, el coste esperado por la compañía se cifra, aproximadamente, en un 65% de la cantidad monetaria que ésta debería pagar en concepto de indemnización (s_0), en caso de que el siniestro fuese legítimo.

Es decir de no aplicar ningún control del fraude, la compañía pagaría 1.5s_o para este siniestro, mientras que de aplicar el sistema se espera que pague 0.65s_o. En realidad ello supone un ahorro esperado de (1.5-0.65)s_o, es decir, 0.85s_o.

En base a los resultados la instalación de un mecanismo de detección de fraude quedaría justificada.

6.3.2 Estimación logística multinomial de la probabilidad de existencia de fraude *para conseguir una indemnización* o *para incrementar una indemnización* (MODELO 2)

Una vez realizado un estudio exhaustivo en relación al primer modelo planteado, pasamos a continuación a analizar los resultados obtenidos al modelizar el segundo árbol de decisión considerado. Tendremos en cuenta, por tanto, una clasificación alternativa para los tipos de fraude que diferencia entre fraude para conseguir una indemnización y fraude para incrementar una indemnización.

6.3.2.1 Análisis de la significación individual y global del modelo

La cuantificación de la probabilidad de que un asegurado no cometa fraude, de que lo haga para conseguir una indemnización que no le corresponde por contrato o para incrementar la que le correspondía en función del mismo, ha sido realizada, de nuevo, mediante la especificación de un modelo lógit multinomial (Figura 9). La carencia de atributos específicos para cada una de las alternativas finales ha impedido realizar la modelización teniendo en cuenta el esquema de decisión presentado en el Capítulo 2, mediante la aplicación de un lógit condicional.

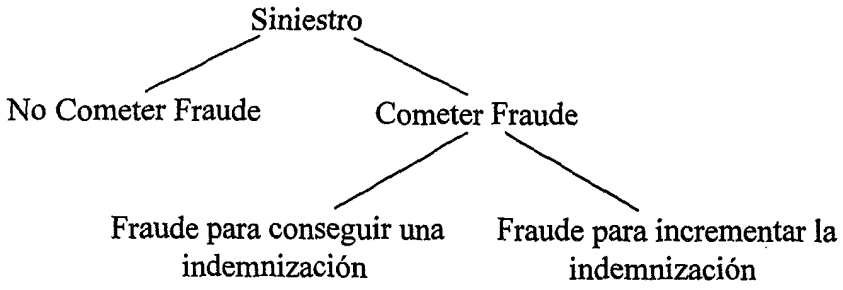

Figura 15

La muestra utilizada en la estimación del modelo está formada, tal y como hemos comentado en el apartado 6.2, por 1676 observaciones. La variable dependiente continúa siendo el tipo de siniestro, lógicamente teniendo en cuenta las categorías de fraude ahora analizadas. Así, hemos utilizado la codificación siguiente: Y=0, si el siniestro no es fraudulento (categoría de referencia); Y=1, si el fraude ha sido cometido para conseguir una indemnización e Y=2, si la ocurrencia del siniestro ha sido utilizada para incrementar la indemnización que correspondería por contrato. Las observaciones han sido ponderadas con la finalidad de reflejar el proceso de obtención de la muestra.

Las variables introducidas en la especificación del modelo son, con algunas variaciones, las mismas que hemos considerado en el Modelo 1. De nuevo utilizamos información que puede ser obtenida por la entidad en un breve periodo de tiempo y que recoge los principales aspectos relacionados con el seguro contratado, las partes intervinientes y el siniestro.

Los estimaciones obtenidas para los coeficientes del modelo, los correspondientes estadísticos de significación individual y los valores de la probabilidad de la cola aparecen en la Tabla siguiente:

Tabla 10. Resultados de la estimación de un modelo lógit multinomial [MODELO 2]

variable	coeficiente	t-test	P-valor
Fraude para conseguir una indemnización:			
CONSTANTE	-3.119	-5.028	0.000
DRAMO	0.426	0.959	0.338
DUSO	-0.658	-2.476	0.013
SFRANQUI	0.015	0.021	0.983
ACCESORI	-0.150	-0.417	0.676
HISTSTR	0.188	4.080	0.000
EDAD	-0.022	-2.985	0.003
DESTAF	1.379	4.459	0.000
DCULPA	0.537	2.761	0.006
ANTIGUO	0.032	1.608	0.108
DCOMOCUR	1.240	6.326	0.000
ALTAHORA	1.489	5.866	0.000
CARRET	0.662	2.169	0.030
SABDOM	0.344	1.733	0.083
DEFOCEMI	2.734	4.805	0.000
STEST	-9.946	-15.261	0.000
DZONA1	0.968	3.009	0.003
DZONA3	1.195	5.371	0.000
DRELATO	0.427	2.288	0.022
SAUTDAD	-1.784	-3.968	0.000
Fraude para incrementar una indemnización:			
CONSTANTE	-3.603	-5.744	0.000
DRAMO	0.886	2.110	0.035
DUSO	-0.403	-1.450	0.147
SFRANQUI	0.100	0.142	0.887
ACCESORI	-0.429	-1.137	0.255
HISTSTR	0.170	3.109	0.002
EDAD	-0.021	-2.710	0.007
DESTAF	0.908	2.333	0.020
DCULPA	2.156	11.332	0.000
ANTIGUO	0.006	0.280	0.779
DCOMOCUR	1.286	6.596	0.000
ALTAHORA	1.146	4.152	0.000
CARRET	0.244	0.673	0.501
SABDOM	0.270	1.339	0.180
DEFOCEMI	0.461	0.519	0.604
STEST	1.646	1.498	0.134
DZONA1	-0.422	-1.126	0.260
DZONA3	0.452	2.277	0.023
DRELATO	0.444	2.322	0.020
SAUTDAD	-1.211	-3.032	0.002
número de observaciones	1676	Chi-cuadrado	535.637
ln función verosimilitud	-880.948	grados de libertad	38
ln verosimilitud restringida	-1148.767	nivel signif.	0.000

El primer vector de parámetros es el asociado a la alternativa de fraude para conseguir una indemnización y el segundo, se refiere la alternativa de fraude para incrementar los costes. Como se observa, en ambas especificaciones se han utilizado las mismas variables.

El modelo es globalmente significativo. El test de la razón de verosimilitud es igual a 535.6, siendo el logaritmo neperiano de la verosimilitud del modelo general (con todas las variables explicativas) igual a -880.9 y el correspondiente a la verosimilitud del modelo restringido igual a -1148.8. La comparación del estadístico con el valor en tablas para una χ^2 con 38 grados de libertad permite rechazar al 1% de significación la hipótesis nula de que todos los coeficientes (excepto las constantes) son iguales a cero en las dos ecuaciones estimadas.

Los resultados obtenidos para los coeficientes estimados permiten destacar la existencia de variables con efectos claramente diferenciados en la explicación de ambos tipos de fraude. Teniendo en cuenta que la categoría de referencia de la variable dependiente es la de los siniestros no fraudulentos y que por tanto la interpretación de los resultados, para los dos vectores de parámetros por separado, debe hacerse en términos de la misma, pasamos a analizar la influencia de las variables consideradas en los comportamientos modelizados.

Los coeficientes que acompañan a las constantes del modelo son estadísticamente significativas (al 1%) en ambos conjuntos de parámetros. Su signo negativo permite interpretar una reducción en la probabilidad de aparición del primer y segundo tipo de fraude, cuando suponemos que el resto de variables incluidas en el modelo son nulas.

Para las variables relacionadas con la póliza y el vehículo en el momento del contrato, los resultados obtenidos evidencian una notable diferencia en su influencia sobre los dos tipos de fraude.

La existencia de cobertura a terceros está directamente relacionada con la probabilidad de aparición de ambos tipos de fraude, pero sólo goza de poder explicativo cuando el fraude ha sido cometido para incrementar una indemnización. La obtención de este resultado puede estar fundamentada en la propia definición de la cobertura, dado que cuando hablamos de conseguir una indemnización parece lógico pensar que sea a favor del asegurado y con la contratación de esta garantía quedan excluidos los daños sufridos por el mismo (o por su vehículo).

Con la variable que recoge el tipo de vehículo que posee el asegurado, el efecto observado es el contrario. La aparición de un coeficiente con signo negativo en ambos vectores de

parámetros indica una disminución en la probabilidad de aparición de fraude cuando el vehículo asegurado es un turismo de uso particular. Sin embargo, únicamente se observa significación estadística (al 5%) para el coeficiente asociado a la categoría de fraude destinado a conseguir una indemnización.

Para los coeficientes asociados a las variables *SFRANQUI* y *ACCESORI*, relacionadas con la existencia de franquicia en la póliza y la contratación de la cobertura de accesorios, respectivamente, los resultados ponen de manifiesto una ausencia de significación estadística en ambos vectores de parámetros. No obstante y, a pesar de carecer de poder explicativo, la aparición de un signo positivo para el primer coeficiente permitiría interpretar que la existencia de franquicia aumenta la probabilidad de aparición de ambos tipos de fraude mientras que el signo negativo para el segundo supondría una interpretación en sentido contrario. Cuando existe franquicia el asegurado sabe que ha de asumir una parte del coste del accidente y ésto le puede llevar, por ejemplo, a declarar en un mismo parte de siniestro daños anteriores del vehículo con el objetivo de hacer frente a una única franquicia. Cuando el asegurado posee cobertura de accesorios disminuye la necesidad de cometer fraudes para recuperar, por ejemplo, un radio-cassette robado o para intentar cambiar uno antiguo por otro más nuevo.

El historial de siniestros del asegurado tiene poder explicativo en la aparición de los dos tipos de fraude. El coeficiente observado, positivo y estadísticamente significativo (al 1%) en ambos vectores podría interpretarse de forma análoga a como lo hacíamos en el Modelo 1. El mayor conocimiento de la forma de actuar de la compañía, la no detección de fraudes anteriores o el hecho de que su clasificación en la escala de *bonus-malus* no pueda ser peor de lo que es, son algunas de las razones que ayudan a interpretar el valor obtenido para el coeficiente.

La edad del asegurado aparece multiplicada por un coeficiente con signo negativo y estadísticamente significativo (al 1%) en ambas categorías de fraude. El resultado obtenido permite señalar que la probabilidad de que el asegurado cometa cualquiera de los dos tipos de fraude disminuye con la edad.

La coincidencia de apellidos entre la parte asegurada y la contraria aumenta la probabilidad de aparición de fraude. El coeficiente que acompaña a esta variable es positivo y estadísticamente significativo en las dos alternativas. La consideración, dentro de la primera categoría de fraude, de aquellos casos en los que el asegurado declara en falso para eludir casos excluidos en la póliza (posibilidad de crear un siniestro con un familiar) permite interpretar el resultado

para el coeficiente que aparece en el primer vector de parámetros. Para el coeficiente que aparece en el segundo vector, bastaría pensar, por ejemplo, en un acuerdo entre familiares que se declaran mutuamente culpables en sus respectivas compañías.

La aceptación de culpa por parte del asegurado está directamente relacionada con un aumento de la probabilidad de aparición de ambos tipos de fraude siendo estadísticamente significativa en la explicación de ambos comportamientos. La interpretación de este resultado puede resultar complicada dada la gran variedad de situaciones que pueden plantearse y, en todo caso, estará sujeta a matizaciones. No obstante, el planteamiento de escenarios concretos puede ayudarnos en la explicación del valor obtenido para el coeficiente. Así, por ejemplo, si la fecha de efecto de la póliza ha sido retrocedida con el objetivo de cubrir un siniestro, parece más lógico pensar en la culpa del asegurado que en la del contrario. De la misma forma, la aceptación de culpa por parte de un asegurado que posee cobertura de daños propios permitirá cubrir los daños del contrario y a la vez reparar los propios, existiendo la posibilidad de que se beneficien ambos implicados.

El coeficiente para la variable que recoge la antigüedad en años del vehículo del asegurado no es estadísticamente significativo en ninguna de las dos alternativas de fraude, aunque para la categoría de fraude destinado a conseguir una indemnización prácticamente podríamos aceptar su significación al 10%. La aparición de un signo positivo para el parámetro estimado nos permite señalar que la probabilidad de existencia de fraude (sobre todo del primer tipo) aumenta cuanto más antiguo es el vehículo. La necesidad de reparar las cada vez más frecuentes averías puede ser una de las principales razones para que el asegurado prepare el siniestro.

Al igual que ocurría en el Modelo 1, las variables que recogen el hecho de que el siniestro sea comunicado a la compañía con posterioridad a la primera semana desde su ocurrencia o el que haya ocurrido a altas horas de la noche presentan coeficientes estadísticamente significativos y relacionados positivamente con la probabilidad de aparición de ambos tipos de fraudes. La interpretación que dábamos entonces para los mismos puede utilizarse ahora, en términos de que cuando el siniestro es legal cabe esperar brevedad en su comunicación a la compañía y de la mayor facilidad que puede existir para cometer fraudes en siniestros ocurridos a altas horas de la noche.

La ocurrencia del siniestro en carretera está directamente relacionado con la probabilidad de existencia de fraude dirigido a conseguir una indemnización. La obtención de un coeficiente positivo y significativo (al 5%) en el primer vector de parámetros permite realizar la

afirmación anterior, sin que podamos decir lo mismo para el coeficiente asociado al segundo tipo de fraude. En este último caso, el coeficiente carece de significación estadística y por tanto la variable analizada no puede considerarse relevante en la explicación de fraude para aumentar la indemnización.

Un comportamiento análogo se observa para la variable que recoge la ocurrencia del siniestro en fin de semana y para la variable que indica la ocurrencia del siniestro entre la fecha de efecto y de emisión de la póliza. Para la primera, la aparición de un parámetro positivo y estadísticamente significativo (al 10%) en el primer vector permite señalar que cuando el siniestro ocurre en sábado o domingo aumenta la probabilidad de que el asegurado defraude para conseguir una indemnización sin que la variable goce de significación en la explicación del segundo tipo de comportamiento fraudulento. Para la segunda, la conclusión es la misma, con la salvedad de que el coeficiente es significativo al 1%. Cabe destacar, en relación a esta variable, como los resultados obtenidos permiten validar los comentarios que para la misma hacíamos en el Modelo 1. De esta forma cabe esperar que el retroceso en la fecha de efecto esté directamente relacionado con la existencia de fraude para conseguir una indemnización e, implícitamente podríamos pensar en la existencia de fraude para beneficiar al asegurado.

Para las variables que recogen la presencia de testigos y la ocurrencia del siniestro en una zona de elevada siniestralidad, los coeficientes estimados muestran la existencia de efectos contrarios. Así, la existencia de testigos disminuye la probabilidad de que el asegurado cometa el primer tipo de fraude mientras que incrementa la probabilidad de existencia del segundo tipo, aunque en este segundo caso el coeficiente no resulta significativo. La interpretación para el primer resultado parece plausible. En relación al segundo podría pensarse en situaciones extremas en las que por ejemplo se diera la existencia de testigos falsos. Para la variable que recoge la ocurrencia del siniestro en una zona de elevada siniestralidad, el coeficiente es positivo y significativo en la primera categoría de fraude mientras que, para la segunda, presenta signo contrario y carece de significación estadística.

La acaecimiento del siniestro en una zona de baja siniestralidad presenta coeficientes positivos y estadísticamente significativos en ambos vectores de parámetros. Nótese de nuevo que tal vez la clasificación de zonas de siniestralidad utilizada no es la adecuada para discriminar entre tipos de fraude.

Por último, los coeficientes asociados a la existencia en la declaración relatada del asegurado de cierta terminología y a la intervención de policía en el siniestro muestran un comportamiento análogo en ambas categorías de fraude, siendo en todos los casos

estadísticamente significativos. Para la primera variable, la aparición de valores positivos permite señalar una relación directa con la probabilidad de aparición de ambos tipos de fraude. Para la segunda, la relación es inversa (disminución de la probabilidad de aparición de fraude cuando la policía interviene en el siniestro).

Al igual que veíamos para el Modelo 1, los coeficientes obtenidos sólo son interpretables en términos de dirección de variación de la probabilidad. Desde este punto de vista, su análisis únicamente nos ha permitido señalar si las variables a las que acompañan tiene una relación positiva o negativa con la probabilidad de elegir cada una de las alternativas de fraude (tomando como categoría de referencia la categoría de no fraude). Cuando la relación es positiva (coeficiente con signo positivo), un incremento en el valor de la variable (o la presencia de un valor igual a 1 si es dicotómica) va asociado a un aumento de la probabilidad correspondiente; cuando es negativa, la variación de la probabilidad se produce en sentido inverso. Los resultados obtenidos tras efectuar el cálculo de los efectos marginales de los regresores en las probabilidades, como solución a la no interpretación directa de los estimadores, reflejan valores muy bajos. Al igual que ocurría en el Modelo 1, el hecho de que numerosas de las variables consideradas sean dicotómicas dificulta su interpretación (recordemos que los efectos marginales son computados teniendo en cuenta el vector de medias de las variables explicativas y dicha medida estadística carece de sentido cuando la variable es categórica).

El cálculo de las probabilidades predichas por el modelo para cada una de las alternativas, teniendo en cuenta los vectores de coeficientes estimados, se ha realizado según las expresiones,

$$P(Y_{i0}=1)=\frac{1}{1+\Gamma_{i1}+\Gamma_{i2}},$$

$$P(Y_{i1}=1)=\frac{\Gamma_{i1}}{1+\Gamma_{i1}+\Gamma_{i2}},$$

$$P(Y_{i2}=1)=\frac{\Gamma_{i2}}{1+\Gamma_{i1}+\Gamma_{i2}},$$

donde,

$$\Gamma_{i1} = \exp(-3.119 + 0.426DRAMO - 0.658DUSO + 0.015SFRANQUI - 0.150ACCESORI + 0.188HISTSTR - 0.002EDAD + 1.379DESTAF + 0.537DCULPA + 0.032ANTIGUC + 1.240DCOMOCUR + 1.489ALTAHORA + 0.662CARRET + 0.344SABDOM + 2.734DEFOCEMI - 9.946STEST + 0.968DZONA1 + 1.195DZONA3 + 0.427DRELATO - 1.784SAUTDAD)$$

$$\Gamma_{i2} = \exp(-3.603 + 0.886DRAMO - 0.403DUSO + 0.100SFRANQUI - 0.429ACCESORI + 0.170HISTSTR - 0.021EDAD + 0.908DESTAF + 2.156DCULPA + 0.006ANTIGUC + 1.286DCOMOCUR + 1.146ALTAHORA + 0.244CARRET + 0.270SABDOM + 0.461DEFOCEMI + 1.646STEST - 0.422DZONA1 + 0.452DZONA3 + 0.444DRELATO - 1.211SAUTDAD).$$

siendo el cómputo para los log-odds entre las probabilidades,

$$\ln\left[\frac{P_{i(1)}}{P_{i(0)}}\right] = \ln(\Gamma_{i1}),$$

$$\ln\left[\frac{P_{i(2)}}{P_{i(0)}}\right] = \ln(\Gamma_{i2}).$$

6.3.2.2 Independencia de Alternativas Irrelevantes

La aplicación del test de Hausmann y McFadden (1984) permite demostrar la independencia entre las alternativas consideradas en el modelo. Al eliminar la alternativa de fraude para incrementar la indemnización, el test es igual a 0.004 y su comparación con el valor en tablas para una χ^2 con 20 grados de libertad supone no poder rechazar la hipótesis nula de independencia de alternativas irrelevantes, al 1% de significación. Una conclusión análoga ha sido obtenida al eliminar la alternativa de fraude para conseguir una indemnización, siendo el valor del test en este caso igual a 0.005.

De nuevo, el cumplimiento de la hipótesis nula puede verificarse comparando el vector de coeficientes estimados considerando las tres alternativas del modelo y el vector obtenido al eliminar la categoría de fraude para conseguir una indemnización y la de fraude para incrementar una indemnización, respectivamente. Cuando estimamos el modelo eliminando la segunda categoría de fraude (la muestra pasa a estar formada por 1300 observaciones) los coeficientes estimados para la categoría de fraude para conseguir una indemnización son los siguientes,

Tabla 11. Resultados de la estimación de un modelo logístico
(eliminando la alternativa de fraude para incrementar una indemnización)

variable	coeficiente	t-test	P-valor
CONSTANTE	-3.094	-4.491	0.000
DRAMO	0.397	0.796	0.426
DUSO	-0.628	-2.150	0.031
SFRANQUI	0.086	0.109	0.913
ACCESORI	-0.139	-0.354	0.723
HISTSTR	0.185	3.695	0.000
EDAD	-0.022	-2.709	0.007
DESTAF	1.362	3.992	0.000
DCULPA	0.521	2.425	0.015
ANTIGUO	0.035	1.556	0.120
DCOMOCUR	1.252	5.881	0.000
ALTAHORA	1.425	5.075	0.000
CARRET	0.579	1.723	0.085
SABDOM	0.332	1.494	0.135
DEFOCEMI	2.799	4.584	0.000
STEST	-9.827	-17.202	0.000
DZONA1	0.975	2.799	0.005
DZONA3	1.215	4.987	0.000
DRELATO	0.382	1.882	0.060
SAUTDAD	-1.827	-3.764	0.000
número de observaciones	1300	Chi-cuadrado	225.896
ln función verosimilitud	-375.676	grados de libertad	19
ln verosimilitud restringida	-488.624	nivel signif.	0.000

Se observa como los resultados prácticamente no difieren de los presentados para la misma alternativa en la Tabla 10, en base al uso de la muestra completa formada por 1676 observaciones.

Cabe señalar, que al igual que ocurría en la demostración de Independencia de Alternativas en el Modelo 1, las ponderaciones introducidas han sido debidamente corregidas para tener en cuenta los tamaños muestrales y el número de elecciones finales consideradas.

La similitud puede observarse también entre los parámetros estimados para la segunda categoría de fraude, una vez eliminada la categoría de fraude para conseguir una indemnización, y los obtenidos para la misma categoría considerando todas la alternativas. Trabajando con una muestra formada por 1374 casos, los resultados son los presentados en la Tabla 12.

Tabla 12. Resultados de la estimación de un modelo logístico
(eliminando la alternativa de fraude para conseguir una indemnización)

variable	coeficiente	t-test	P-valor
CONSTANTE	-3.572	-5.245	0.000
DRAMO	0.903	1.980	0.048
DUSO	-0.445	-1.471	0.141
SFRANQUI	0.027	0.036	0.971
ACCESORI	-0.316	-0.793	0.428
HISTSTR	0.151	2.428	0.015
EDAD	-0.021	-2.507	0.012
DESTAF	1.059	2.552	0.011
DCULPA	2.153	10.753	0.000
ANTIGUO	0.010	0.463	0.643
DCOMOCUR	1.330	6.326	0.000
ALTAHORA	1.207	3.995	0.000
CARRET	0.382	0.977	0.328
SABDOM	0.258	1.188	0.234
DEFOCEMI	0.385	0.414	0.678
STEST	1.645	1.420	0.156
DZONA1	-0.433	-1.053	0.292
DZONA3	0.386	1.809	0.070
DRELATO	0.434	2.094	0.036
SAUTDAD	-1.197	-2.781	0.005
número de observaciones	1374	Chi-cuadrado	309.207
ln función verosimilitud	-367.248	grados de libertad	19
ln verosimilitud restringida	-521.852	nivel signif.	0.000

No se observan diferencias notables en relación a los coeficientes presentados en el segundo vector de parámetros de la Tabla 10.

Al no poder rechazar la hipótesis nula de Independencia de Alternativas Irrelevantes, el cociente entre la probabilidad de que el asegurado elija la primera alternativa de fraude y la probabilidad de que no defraude ($P_{i(1)}/P_{i(0)}$) permanece constante cuando no consideramos el segundo tipo de fraude. Lo mismo podemos señalar para el cociente entre la probabilidad de que el asegurado elija el segundo tipo de comportamiento fraudulento y la probabilidad de que no cometa fraude ($P_{i(2)}/P_{i(0)}$), esta vez, al no considerar la alternativa de fraude para conseguir una indemnización.

6.3.2.3 Análisis de la calidad del ajuste

Para cada observación de la muestra, la comparación entre los valores observados y predichos para la variable dependiente queda recogida en la siguiente tabla de clasificación.

Tabla 13. Frecuencias de clasificación (utilizando el criterio de máxima probabilidad)

	Elección predicha			
	Legítimo	Conseguir una indemnización	Incrementar una indemnización	Total
Elección Observada				
Legítimo	964	12	22	998
Conseguir una indemnización	213	48	41	302
Incrementar una indemnización	243	15	118	376
Total	1420	75	181	1676

Los datos anteriores reflejan que un total de 1130 expedientes han sido clasificados por el modelo dentro la categoría correspondiente siendo el porcentaje de aciertos del 67.4%. Sin embargo, mientras que el porcentaje de siniestros sin fraude detectado clasificados de forma correcta es elevado, de un 96.6%, el error de clasificación observado para los siniestros fraudulentos es notable. Tan sólo un 15.9% de los siniestros con fraude destinado a conseguir una indemnización son clasificados correctamente. En relación a los siniestros con fraude destinado a incrementar la indemnización derivada de la ocurrencia del siniestro el porcentaje, si bien mayor, continua siendo preocupante, tomando un valor del 31.4%.

La situación es parecida a la que encontrábamos en el análisis de la Tabla 6, asociada al Modelo 1. De nuevo el problema no es sólo el escaso número de fraudes detectados, sino también el elevado porcentaje de siniestros fraudulentos clasificados por el modelo como legítimos. Así, el hecho de que 213 y 243 expedientes, asociados a la primera y la segunda categoría de fraude, respectivamente, sean clasificados como no fraudulentos induce a pensar en la escasa calidad predictiva del modelo desde la óptica de fraude (tras su aplicación la compañía dejaría de detectar el 70.5% de los siniestros que presentan el primer tipo de comportamiento fraudulento considerado y el 64.6% de los siniestros que presentan el segundo). La solución al problema, producido tal vez porque la muestra utilizada no sea suficientemente representativa, se encuentra en la optimización del punto de corte o criterio probabilístico utilizado en la delimitación de las zonas de aceptación de cada una de las alternativas.

El análisis gráfico de las probabilidades predichas por el modelo teniendo en cuenta los valores observados para la variable dependiente, confirma lo manifestado en el párrafo anterior.

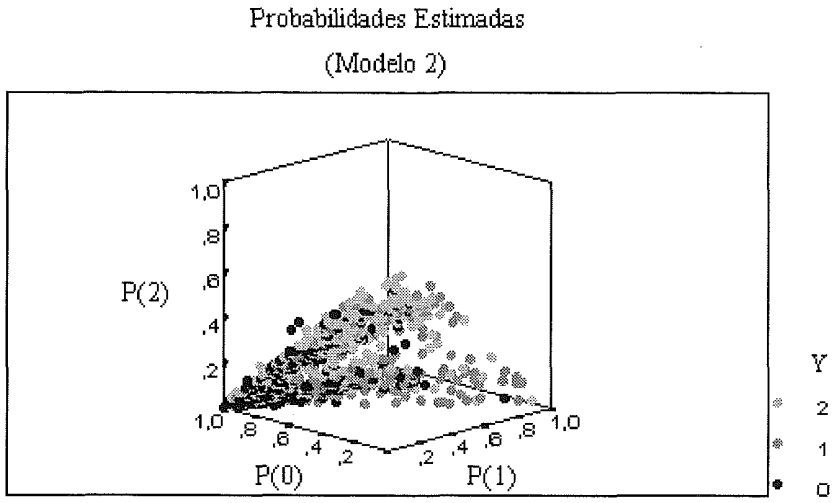

Figura 16

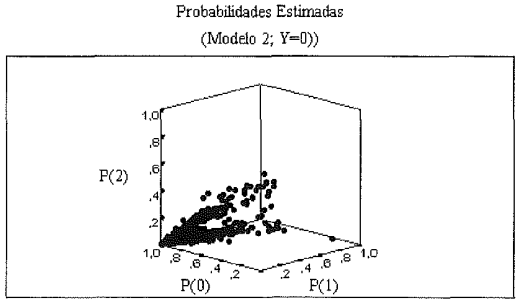

Figura 17

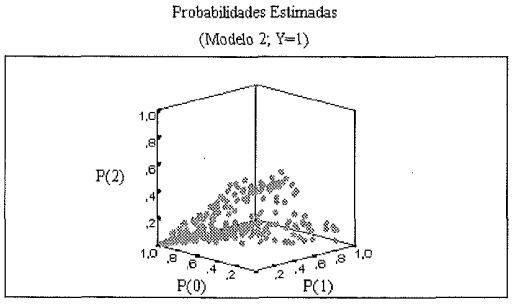

Figura 18

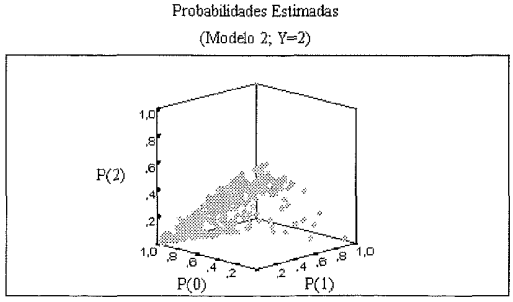

Figura 19