CUADERNOS DE LA FUNDACIÓN

Nº 88

* * * *

ANÁLISIS MULTIVARIANTE APLICADO A LA SELECCIÓN DE FACTORES DE RIESGO EN LA TARIFICACIÓN

Autores: Eva Boj del Val
Mª Mercè Claramunt Bielsa
Josep Fortiana Gregori

Diciembre, 2004

ISBN: 84-89429-78-2

Depósito Legal: M-53.029-2004

LISTA DE CUADERNOS DE LA FUNDACIÓN MAPFRE ESTUDIOS EDITADOS:

1. Filosofía Empresarial
2. Resultados de la Encuesta sobre "Altos Profesionales de Seguros" (A.P.S.)
3. Dirección y Gestión de la Seguridad
4. Los Seguros en una Europa cambiante: 1990-1995 (No disponible)
5. La Distribución Comercial del Seguro: Sus Estrategias y Riesgos
6. Elementos de Dirección Estratégica de la Empresa
7. Los Seguros de Responsabilidad Civil y su Obligatoriedad de Aseguramiento
8. La Implantación de un Sistema de Controlling Estratégico en la Empresa
9. Técnicas de Trabajo Intelectual
10. Desarrollo Directivo: Una Inversión Estratégica
11. El Concepto de Seguridad en la Ciencia y la Ciencia de la Seguridad
12. Los Seguros de Salud y la Sanidad Privada
13. Calidad Total y Seguridad
14. El Reaseguro de Exceso de Pérdidas
15. El Coste de los Riesgos en la Empresa Española 1991
16. La Legislación Española de Seguros y su Adaptación a la Normativa Comunitaria

Número Especial: Informe sobre el Mercado de Seguros 1993

17. Medio Ambiente Seguro: Desarrollo Futuro

18. El Seguro de Crédito a la Exportación en los países de la OCDE (Evaluación de los resultados de los aseguradores públicos)
19. Una Teoría de la Educación

20. El Reaseguro en los Procesos de Integración Económica

Número Especial: Informe sobre el Mercado de Seguros 1994

21. La Nueva Regulación de las Provisiones Técnicas en la Directiva de Cuentas de la C.E.E. Provisiones Técnicas de Seguros de Vida en las Directivas Comunitarias

22. Rentabilidad y Productividad de Entidades Aseguradoras

23. Análisis de la Demanda de Seguro Sanitario Privado

24. El Seguro: Expresión de Solidaridad desde la Perspectiva del Derecho

25. El Reaseguro Financiero

26. El Coste de los Riesgos en la Empresa Española 1993

27. La Calidad Total como Factor para elevar la Cuota de Mercado en Empresas de Seguros

28. La Naturaleza Jurídica del Seguro de Responsabilidad Civil

29. Ruina y Seguro de Responsabilidad Civil Decenal

Número Especial: Informe sobre el Mercado de Seguros 1995

30. El Tiempo del Directivo

31. Tipos Estratégicos, Orientación al Mercado y Resultados Económicos: Análisis Empírico del Sector Asegurador Español

32. Decisiones Racionales en Reaseguro

33. La función del Derecho en la Economía

34. El Coste de los Riesgos en la Empresa Española 1995

35. El Control de Riesgos en Fraudes Informáticos

36. Cláusulas Limitativas de los Derechos de los Asegurados y Cláusulas Delimitadoras del Riesgo Cubierto. Las Cláusulas de Limitación Temporal de la Cobertura en el Seguro de Responsabilidad Civil

Número Especial: Informe sobre el Mercado de Seguros 1996

37. La Responsabilidad Civil por Accidente de Circulación. Puntual Comparación de los Derechos Francés y Español

38. Legislación y Estadísticas del Mercado de Seguros en la Comunidad Iberoamericana

39. Perspectiva Histórica de los Documentos Estadístico-Contables del Órgano de Control: Aspectos Jurídicos, Formalización y Explotación

40. Resultados de la Encuesta sobre la Organización y Gestión de la Seguridad en la Empresa (1996)

41. De Maastricht a Amsterdam: Un paso más en la integración europea

Número Especial: Informe sobre el Mercado de Seguros 1997

42. La Responsabilidad Civil por contaminación del entorno y su aseguramiento

43. Resultados de la Encuesta sobre Disponibilidad de Instalaciones de Protección contra Incendios en la Empresa 1997"

44. Resultados de la Encuesta sobre Implantación en la Empresa de la Ley de Prevención de Riesgos Laborales

45. Los Impuestos en una Economía Global

46. Evolución y Predicción de las Tablas de Mortalidad Dinámicas para la Población Española

47. El Fraude en el Seguro del Automóvil: Cómo detectarlo

48. Matemática Actuarial no Vida con MapleV

49. Solvencia y Estabilidad Financiera en la Empresa de Seguros: Metodología y Evaluación Empírica mediante Análisis Multivariante

50. Mixturas de Distribuciones: Aplicación a las variables más relevantes que modelan la siniestralidad en la Empresa Aseguradora

51. Seguridades y Riesgos del joven en los grupos de edad

52. La Estructura Financiera de las Entidades de Seguros

53. Habilidades Directivas: Estudio de sesgo de género en instrumentos de evaluación

54. El Corredor de Reaseguro y su legislación específica en América y Europa

55. Resultados de la Encuesta: "La Seguridad contra Intrusión (Seguridad Privada) en la Empresa. 1999"

56. Análisis económico y estadístico de los factores determinantes de la demanda de los seguros privados en España

57. Informe final. Encuesta: "La Organización y Gestión de la Seguridad en la Empresa. 1999"

58. Problemática contable de las operaciones de reaseguro

59. Estudios sobre el Euro y el Seguro

60. Análisis Técnico y Económico del conjunto de las empresas aseguradoras de la Unión Europea

61. Sistemas Bonus-Malus generalizados con inclusión de los costes de los siniestros

62. Seguridad Social. Temas generales y régimen de clases pasivas del Estado

63. Análisis de la repercusión fiscal del seguro de vida y los planes de pensiones. Instrumentos de previsión social individual y empresarial

64. Fundamentos técnicos de la Regulación del Margen de Solvencia

65. Ética Empresarial y Globalización

66. Encuesta: "Seguridad contra Incendios en la empresa. 2000"

67. Gestión Directiva en la Internacionalización de la Empresa

68. Los seguros de crédito y de caución en Iberoamérica

69. Provisiones para prestaciones a la luz del Reglamento de Ordenación y Supervisión de los Seguros Privados: Métodos Estadísticos de Cálculo

70. El Cuadro de Mando Integral para las entidades aseguradoras

71. Gestión de activos y pasivos en la cartera de un fondo de pensiones

72. Análisis del proceso de exteriorización de los compromisos por pensiones

73. Financiación del capital-riesgo mediante el seguro

74. Estructuras de propiedad, organización y canales de distribución de las empresas aseguradoras en el mercado español

75. Incidencia de la Nueva Ley de Enjuiciamiento Civil en los procesos de responsabilidad civil derivada del uso de vehículos a motor

76. La incorporación de los sistemas privados de pensiones en las pequeñas y medianas empresas

77. Resultados de la Encuesta sobre *"El Coste de los Riesgos en la Empresa Española. 2001"*

78. Nuevas perspectivas de la educación universitaria a distancia

79. La actividad de las compañías aseguradoras de vida en el marco de la gestión integral de activos y pasivos

80. Los Planes y Fondos de Pensiones en el contexto europeo: la necesidad de una armonización

81. El Seguro de Dependencia. Una visión general

82. Informe Final. Encuesta: "La Organización y Gestión de la Seguridad en la Empresa 2002"

83. La teoría del valor extremo: fundamentos y aplicación al seguro, ramo de responsabilidad civil autos

84. Estudio de la estructura de una cartera de pólizas y de la eficiencia de un Sistema Bonus-Malus

85. La Matriz Valor-Fidelidad en el Análisis de los Asegurados en el Ramo del Automóvil

86. El Margen de Solvencia de las Entidades Aseguradoras en Iberoamérica

87. Dependencia en el modelo individual, aplicación al riesgo de crédito

88. Análisis Multivariante Aplicado a la Selección de Factores de Riesgo en la Tarificación

ANÁLISIS MULTIVARIANTE APLICADO A LA SELECCIÓN DE FACTORES DE RIESGO EN LA TARIFICACIÓN

Autores:

Dra. Eva Boj del Val
Departamento de Matemática Económica, Financiera y Actuarial de la Universidad de Barcelona

Dra. Mª Mercè Claramunt Bielsa
Departamento de Matemática Económica, Financiera y Actuarial de la Universidad de Barcelona

Dr. Josep Fortiana Gregori
Departamento de Estadística de la Universidad de Barcelona

Trabajo resultante de una beca RIESGO y SEGURO 2000/2001, concedida por la Fundación MAPFRE Estudios.

Presentación

Me resulta muy grato presentar este trabajo realizado por los profesores Eva Boj y Mª Mercè Claramunt del Departamento de Matemática Económica, Financiera y Actuarial y por el profesor Josep Fortiana del Departamento de Estadística, ambos Departamentos de la Universidad de Barcelona.

En primer lugar, quiero destacar la gran importancia que para el sector asegurador tiene la tarificación *a priori* y por ello la realización de estudios tanto teóricos como empíricos para la determinación de los principales factores de riesgo a incluir como posibles variables de tarifa en los distintos tipos de seguros generales o no vida.

El presente trabajo plantea un estudio teórico de la posible aplicación de la regresión basada en distancias para seleccionar las variables de tarifa como metodología complementaria o alternativa a los métodos actuariales tradicionalmente aplicados para dar solución a este problema.

También me gustaría destacar, como profesor universitario, que este trabajo hace realidad el ideal de interdisciplinariedad de la investigación universitaria, ya que en la realización han colaborado profesores de dos departamentos de la Universidad de Barcelona: Eva Boj y Mª Mercè Claramunt han aportado sus conocimientos más profundos en el análisis económico y actuarial, y el profesor Fortiana, de la misma manera, sus amplios conocimientos en estadística.

El trabajo va dirigido a los actuarios dedicados a seguros generales o no vida, con la esperanza de que les sea de utilidad en su trabajo diario y les permita aproximarse a una metodología de análisis de la siniestralidad que les aporte criterios para la toma de decisiones relacionada con la selección de variables de tarifa en la tarificación *a priori*.

Por último, quisiera resaltar la importancia que para la investigación actuarial tienen las becas Riesgo y Seguro que otorga la Fundación MAPFRE Estudios, ya que son las únicas que con carácter privado permiten co-financiar estudios actuariales con la universidad pública, apostando por una investigación actuarial de calidad.

Antonio Alegre Escolano
Catedrático de Matemática Actuarial
Universidad de Barcelona

Agradecimientos

El origen de este libro se remonta al año 1996 cuando el Dr. Antonio Alegre Escolano [Catedrático de Matemática Actuarial del Departamento de Matemática Económica, Financiera y Actuarial de la Universidad de Barcelona] propuso el tema de la selección de los factores de riesgo en la tarificación de los seguros como objetivo de investigación de gran importancia en el ámbito tanto práctico como teórico de los Seguros No Vida. Agradecemos al Dr. Antonio Alegre que nos animase a centrar nuesta investigación en este tema.

El trabajo realizado ha dado lugar a la tesis doctoral de Eva Boj del Val dirigida conjuntamente por la Dra. Mª Mercè Claramunt Bielsa y el Dr. Josep Fortiana Gregori, que obtuvo en junio de 2003 la calificación de Sobresaliente Cum Laude por unanimidad. Agradecemos a todos los miembros del tribunal, y en especial al Dr. Ángel Vegas Montaner, los comentarios realizados que nos han sido muy útiles en la redacción última del presente libro.

Tenemos que agradecer a dos compañías de seguros que operan en España la cesión gratuita y desinteresada de los datos sobre la experiencia de siniestralidad de su cartera de Responsabilidad Civil del Automóvil. Sin los mismos hubiera sido imposible la realización de las aplicaciones prácticas.

Por último, pero no por ello en menor grado, agradecemos a la Fundación MAPFRE Estudios la confianza depositada en nosotros al conceder una beca RIESGO y SEGURO 2000/2001 englobada en el área concreta "Utilización de modelos *data mining* para el análisis del riesgo y tarificación del seguro de automóviles en España". La obtención de la beca ha supuesto, por un lado, financiación adicional para reuniones de trabajo, material bibliográfico e informático y, especialmente, para la asistencia a congresos nacionales e internacionales del campo estadístico y actuarial. La asistencia y presentación de ponencias a los mismos nos ha permitido conocer de primera mano el estado actual del tema. Por otro lado, la beca nos ha servido para contar con el asesoramiento de una compañía real, MAPFRE, que opera en el Seguro del Automóvil. Agradecemos el constante apoyo recibido de Juan Antonio Rodrigo y Mónica Román del Área Técnica de MAPFRE Automóviles, y de José Luis Catalinas de la Fundación MAPFRE Estudios.

Los autores.

Índice

Capítulo 1

Introducción

El presente trabajo, **Análisis multivariante aplicado a la selección de factores de riesgo en la tarificación**, está dedicado al estudio de las metodologías estadísticas que dan solución al problema de la selección de variables de tarifa a partir del conjunto de factores potenciales de riesgo, como primer paso del proceso de tarificación *a priori* que se realiza en los seguros no vida.

El proceso de tarificación *a priori* tiene como finalidad asignar una prima, precio del seguro, a una póliza que entra en la cartera de seguros de un asegurador, a partir de unas ciertas características conocidas referentes a dicha póliza. Se denomina *a priori* puesto que nos permite asignar una prima a un riesgo que se incorpora a nuestra cartera sin tener necesariamente experiencia sobre la siniestralidad que conlleva.

El cálculo de las primas debe responder a los principios de equidad, solidaridad y suficiencia. El principio de equidad implica, precisamente, que la prima se ajuste al riesgo de siniestralidad de cada póliza, es decir, que el asegurado pague según el riesgo que incorpora. Desde el punto de vista actuarial, ello conlleva tener en cuenta, para la tarificación, los factores de riesgo que expliquen en mayor medida el comportamiento de la estructura de siniestralidad.

El primer paso del proceso de cálculo de dicha prima es la selección de las variables de tarifa, que consiste en la elección de los factores de riesgo o características que utilizaremos para distinguir a los asegurados/pólizas con diferentes riesgos asociados. La correcta selección de dichas variables es fundamental. Para ello, debemos hacer uso de técnicas de análisis estadístico multivariante, las cuales nos permiten organizar procesos de selección teniendo en cuenta simultáneamente el conjunto de factores.

Los objetivos del presente trabajo son:

- Realizar el estudio teórico de las metodologías estadísticas aplicadas en la literatura actuarial para la selección de variables de tarifa.
- Realizar el estudio teórico de la viabilidad de la regresión basada en distancias para constituir una herramienta alternativa en la selección de variables de tarifa.
- Ilustrar la aplicabilidad a la tarificación de las metodologías ya existentes y de la propuesta en el trabajo con datos reales de carteras de seguros no vida, analizando las distintas dificultades empíricas.

Estos tres objetivos básicos implican y se reflejan en la estructura del trabajo. Por un lado se estudian las metodologías de selección desde el punto de vista estadístico, y por otro, su aplicabilidad a la realidad actuarial de los seguros no vida.

En el capítulo 2, se estudia el proceso de tarificación *a priori* en los seguros no vida, detallando sus fases y justificando la importancia de una correcta selección de los factores de riesgo como base del proceso. Puesto que decidimos centrarnos en el seguro del automóvil, analizamos con detalle las peculiaridades de su funcionamiento en España.

El capítulo 3 se dedica a la revisión de las técnicas estadísticas aplicadas en el campo actuarial para llevar a cabo la selección de variables de tarifa, principalmente el análisis de segmentación y el modelo lineal generalizado.

En el capítulo 4, nos centramos en el estudio teórico de la regresión basada en distancias. Describimos su filosofía y funcionamiento, y además la desarrollamos en distintos aspectos teóricos, todos ellos necesarios para aplicarla a la selección de variables de tarifa. En primer lugar, necesitamos plantear un método de selección apropiado, con las correspondientes medidas y tests estadísticos. En el proceso nos encontramos con que no conocemos las distribuciones de los estadísticos de test para muestras finitas. Solventamos tal dificultad optando por simular las distribuciones mediante la metodología *bootstrap* que, como veremos, se adapta especialmente bien a las características peculiares de la regresión basada en distancias. En segundo lugar, proponemos la versión ponderada de la regresión basada en distancias para abordar el caso de la frecuencia de siniestralidad en el tratamiento del número de siniestros. En este capítulo se concentran las principales aportaciones teóricas del trabajo.

El capítulo 5 incluye las aplicaciones prácticas. En él se ilustra la aplicabilidad de las tres metodologías estudiadas con detalle en el trabajo: el análisis de segmentación, el modelo lineal generalizado y la regresión basada en distancias; lo que se corresponde con nuestro tercer objetivo.

Utilizamos cuatro conjuntos de datos distintos que se desarrollan en las cuatro aplicaciones en que se divide el capítulo.

La primera aplicación utiliza unos datos sobre las cuantías de los siniestros de los impagos de préstamos de una entidad financiera. La característica principal de estos datos es su simplicidad. La tercera aplicaeión utiliza unos datos sobre las cuantías de siniestros para la cobertura de daños propios del seguro del automóvil; también simples, pero agregados y, por lo tanto, ponderados. Estas dos aplicaciones se utilizan principalmente para comprobar el correcto funcionamiento en la práctica de la metodología de selección propuesta para la regresión basada en distancias.

Las aplicaciones segunda y cuarta utilizan datos reales de España, de las carteras del seguro del automóvil cedidas. La segunda se refiere a la cuantía por siniestro para daños personales y la cuarta al número de siniestros para daños materiales. Para analizar estos dos conjuntos de datos tenemos en cuenta las características peculiares que tienen en España los seguros de automóviles.

En las aplicaeiones se utiliza tanto el software disponible ya existente, como el implementado especialmente para las aplicaciones. Todo este conjunto de software se describe durante el trabajo.

Finalizamos con un capítulo de conclusiones, en el que se destacan las aportaciones del trabajo.

Capítulo 2

Tarificación en los seguros no vida

En este capítulo pretendemos, en primer lugar, centrar el escenario actuarial de los seguros no vida remarcando sus peculiaridades. En segundo lugar, definir el proceso de tarificación *a priori*, detallando sus fases y justificando la importancia de una correcta selección de los factores de riesgo como base del citado proceso.

El presente capítulo se estructura en tres apartados y tres anexos.

Dedicamos un primer apartado, 2.1, de introducción, en el que redactamos generalidades de los seguros hasta llegar al marco de los denominados no vida. Detallamos sus ramos y sus características principales en comparación con los de vida. Prestamos, durante todo el capítulo, especial atención al seguro del automóvil, ya que al mismo se refieren la mayor parte de aplicaciones presentadas en este trabajo.

Dedicamos un segundo apartado, 2.2, a la tarificación. Describimos el proceso de riesgo asociado al acaecimiento de los siniestros y a sus respectivas cuantías, dando lugar al coste total de los siniestros en un determinado período. Con las hipótesis de independencia y equidistribución del proceso de riesgo, obtenemos la esperanza matemática del coste total que nos lleva a la prima pura, como producto del número esperado de siniestros por la cuantía esperada de un siniestro.

Detallamos las distintas componentes de la prima hasta la construcción de la prima de recibo o precio final del seguro, y describimos los principios del cálculo de primas en lo que a la parte técnica actuarial se refiere: principios de equidad, solidaridad y suficiencia.

Llegados ya a los sistemas de tarificación, nos centramos en la tarificación *a priori*. Observamos que el principio de equidad en el cálculo de la tarifa necesita de los factores de riesgo como variables exógenas del modelo. A través de ellos seremos capaces de explicar la mayor parte de aleatoriedad del coste total esperado.

Puesto que la prima pura se obtiene como el producto de la esperanza del número de siniestros por la esperanza de la cuantía de un siniestro, los factores de riesgo pueden ser seleccionados por separado respecto a ambas variables.

Resaltamos la necesidad de una buena base de datos como paso previo a la explotación estadística. Describimos los datos de experiencia de siniestralidad necesarios para la realización de un proceso de tarificación *a priori*.

Exponemos el estado actual del seguro del automóvil en lo que a ficheros sectoriales se refiere, resaltando, por un lado, el gran avance que supone en la tarificación la constitución del Fichero Histórico de SINiestralidad de COnductores (SINCO); y por otro, la problemática asociada al Convenio entre Entidades Aseguradoras de Automóviles para la Indemnización Directa de Daños Materiales a Vehículos (CIDE), respecto a la selección de los factores de riesgo influyentes en la cuantía de un siniestro para daños materiales en el seguro de responsabilidad civil obligatoria.

Finalmente, dedicamos un tercer apartado, 2.3, a la mera descripción de la experiencia de siniestralidad y factores de riesgo asociados a los datos que serán utilizados en el capítulo 5 para ilustrar las metodologías estadísticas de selección presentadas en los capítulos 3 y 4.

2.1. Introducción

2.1.1. Generalidades de los seguros

Un contrato de seguro es un servicio de seguridad ofrecido por una entidad económica o ente asegurador (sociedad anónima, mutua, mutualidad de previsión social o cooperativa de seguros), por el cual el asegurador se obliga, mediante el cobro de una prima y para el caso de que se produzca el evento cuyo riesgo es objeto de cobertura, a indemnizar, dentro de los límites pactados, el daño producido al asegurado o a satisfacer un capital, una renta u otras prestaciones convenidas [Garrido y Comas (1987)].

Podemos señalar diversos hechos diferenciales de la identidad de un contrato de seguro:

- Es un contrato aleatorio: una de las partes, o ambas recíprocamente, se obligan a dar o hacer alguna cosa en equivalencia de lo que la otra parte ha de dar o hacer para el caso de un acontecimiento

incierto, o que ha de ocurrir en tiempo indeterminado. Por otro lado, en casi todos los seguros se desconoce, *a priori*, el importe de la cantidad a satisfacer por el asegurador en caso de siniestro. Y lo que es más importante, el hecho principal y sus consecuencias –el siniestro– han de ser producto de causas completamente aleatorias en su acaecimiento, es decir, deberse a circunstancias ajenas a la voluntad consciente y propósito deliberado del asegurado de provocar el siniestro. A este planteamiento hay que añadir que, para el asegurador, el riesgo de pérdidas o ganancias se ve considerablemente atenuado, porque para él, dedicado profesionalmente a la asunción de riesgos ajenos, los resultados del conjunto de riesgos asumidos han de ajustarse al modelo estadístico utilizado para el cálculo de las tarifas. Es decir, se trata de un coste global previamente conocido, dentro de una compensación recíproca de riesgos aceptados.

- Es un contrato consensual: se perfecciona por el mero consentimiento. Aunque existe una gran discusión en si debería ser consensual o formal, es decir, que requiera ciertos requisitos formales como la firma de la póliza para perfeccionarse.

- Es un contrato oneroso: ambas partes asumen la obligación de realizarse prestaciones recíprocas, que de ordinario revisten carácter económico. No hay ningún propósito de libertad en quienes intervienen en el contrato de seguro, todos ellos persiguen un beneficio que tiene que producirse a costa del oponente en el convenio. El asegurador promete la prestación, y la lleva a cabo en su momento, porque para ello percibió anticipadamente la prima. El asegurado, en cambio, realiza el pago del precio del seguro, porque desea liberarse del riesgo que transfiere al asegurador. En su caso, el beneficio consiste en hacer seguro lo que para el interesado es inseguro.

- Es un contrato de buena fe: es importante resaltar este punto dada la dependencia del asegurador respecto del tomador o asegurado, en cuanto afecta a su indispensable colaboración para conocer ciertos datos que pueden serle de importancia para tomar decisiones sobre la aceptación del seguro y sus condiciones, tramitar adecuadamente el siniestro y otras situaciones análogas.

- Es un contrato de adhesión: es un contrato a aceptar o dejar, el que quiera utilizar el convenio ofrecido no tiene otra alternativa que aceptarlo en bloque o rechazarlo, no cabe la discusión.

El contrato en que se materializa este convenio oneroso se denomina póliza del seguro y se define como el documento emitido por la entidad aseguradora, suscrito por ésta y el tomador, que tiene por objeto probar la existencia del contrato de seguro, concretando sus condiciones.

El precio del servicio de seguridad en un contrato de seguro es la prima de seguro, la cual es función del riesgo asegurable y de los restantes factores que integran el coste de la empresa.

El riesgo se define como el acontecimiento incierto, independiente de la voluntad exclusiva de las partes, cuya realización implica, normalmente, consecuencias desfavorables para el asegurado. Así, en un contrato de seguro hay dos componentes básicas e imprescindibles para su viabilidad: la existencia de un riesgo y el pago de un precio por su cobertura, la prima.

Los elementos personales que intervienen en el contrato de seguro son el asegurador, el tomador, el asegurado y el beneficiario:

- El asegurador es la persona jurídica que, constituida con arreglo a lo dispuesto por la legislación correspondiente, se dedica a asumir riesgos ajenos, cumpliendo a lo que este efecto establece aquella legislación, mediante la percepción de un cierto precio llamado prima.

- El tomador es la persona que contrata y suscribe la póliza de seguro, por cuenta propia o la de un tercero, asumiendo las obligaciones y derechos que se establecen en la Ley de Ordenación y Supervisión de los Seguros Privados en España.

- El asegurado es el titular del área de interés a que la cobertura del seguro concierne, y del derecho a la indemnización que en su día se satisfaga que, en ciertos casos, puede trasladarse al beneficiario. Es la persona natural o jurídica a quien el acontecimiento del siniestro va a afectar más directamente. Cuando el tomador y el asegurado sean personas distintas, las obligaciones y los deberes que derivan del contrato corresponden al tomador del seguro, excepto los que, por su naturaleza, deban ser cumplidos por el asegurado.

- El beneficiario es el que va a recibir la utilidad del seguro cuando se produzca el hecho contemplado en el mismo. Es aquél sobre quien recaen los beneficios de la póliza pactada, por voluntad expresa del tomador.

2.1.2. Clasificación de los seguros

La riqueza de modalidades aseguratívas que existe en el mercado correspondientes a la extensa gama de necesidades de previsión actuales, hacen que resulte posible señalar diversas clases de contratos de seguro atendiendo a las distintas circunstancias que en ellos se dan. Algunas clasificaciones posibles de los contratos son [Pérez (1986,2001)]:

a) Según la localización del riesgo: seguros marítimos, terrestres y aéreos

b) Según la clase de riesgos afectados: seguros de cosas o bienes, de personas y patrimoniales.

c) Según la condición jurídica: contrato de seguro puro y mutuo

d) Según la forma de pago de la prima: contrato de seguro a pagar mediante primas periódicas y a pagar de una sola vez con una única prima

e) Según la clase de obligación principal asumida por el asegurador: seguros de capital, de renta y de prestación de servicios

f) Según se refieren a la vida o no de las personas: seguros de vida y de no vida

Si nos regimos por el **marco legal español**[1] los seguros se clasifican en tres grandes grupos:

- Seguro directo de vida: Incluye un único ramo, el ramo de vida, junto con la posibilidad de cubrir riesgos complementarios
- Seguro directo distinto del seguro de vida: Los riesgos cubiertos se encuentran clasificados en una serie de ramos los cuales detallamos más abajo
- Reaseguro: Por el contrato de reaseguro el reasegurador se obliga a reparar, dentro de los límites establecidos en la Ley y en el contrato, la deuda que nace en el patrimonio del reasegurado como consecuencia de la obligación por éste asumida como asegurador en el contrato de seguro

[1] Ley 30/1995 de 8 de noviembre, sobre Ordenación y Supervisión de los Seguros Privados (LOSSP). BOE nº 268, de 9 de noviembre de 1995.
Real Decreto 2486/1998 de 20 de noviembre por el que se aprueba el Reglamento de Ordenación y Supervisión de los Seguros Privados. BOE nº 282, de 25 de noviembre de 1998.

Ramos en los seguros no vida

1. Accidentes
2. Enfermedad (comprendida la asistencia sanitaria)
3. Vehículos terrestres (no ferroviarios). Incluye todo daño sufrido por vehículos terrestres sean o no automóviles, salvo los ferroviarios.
4. Vehículos ferroviarios
5. Vehículos aéreos
6. Vehículos marítimos, lacustres y fluviales
7. Mercancías transportadas (comprendidos los equipajes y demás bienes transportados)
8. Incendio y elementos naturales. Incluye todo daño sufrido por los bienes (distinto de los comprendidos en los ramos 3, 4, 5, 6 y 7) causado por incendio, explosión, tormenta, elementos naturales distintos de la tempestad, energía nuclear y hundimiento de terreno.
9. Otros daños a los bienes. Incluye todo daño sufrido por los bienes (distinto de los comprendidos en los ramos 3, 4, 5, 6 y 7) causado por el granizo o la helada, así como por robo u otros sucesos distintos de los incluidos en el número 8.
10. Responsabilidad civil en vehículos terrestres automóviles (comprendida la responsabilidad del transportista)
11. Responsabilidad civil en vehículos aéreos (comprendida la responsabilidad del transportista)
12. Responsabilidad civil en vehículos marítimos, lacustres y fluviales (comprendida la responsabilidad del transportista)
13. Responsabilidad civil general. Comprende toda responsabilidad distinta de las mencionadas en los números 10, 11 y 12.
14. Crédito. Comprende insolvencia en general, venta a plazos, crédito a la exportación, crédito hipotecario y crédito agrícola.

15. Caución (directa e indirecta)

16. Pérdidas pecuniarias diversas. Incluye riesgos del empleo, insuficiencia de ingresos (en general), mal tiempo, pérdida de beneficios, subsidio por privación temporal del permiso de conducir, persistencia de gastos en general, gastos comerciales imprevistos, pérdida del valor venal, pérdidas de alquileres o rentas, pérdidas comerciales indirectas distintas a las anteriormente mencionadas, pérdidas pecuniarias no comerciales y otras pérdidas pecuniarias.

17. Defensa jurídica

18. Asistencia. Asistencia a las personas que se encuentren en dificultades durante desplazamientos o ausencias de su domicilio o de su lugar de residencia permanente. Comprenderá también la asistencia a las personas que se encuentren en dificultades en circunstancias distintas, determinadas reglamentariamente, siempre que no sean objeto de cobertura en otros ramos de seguro.

19. Decesos. Incluye operaciones de seguro que garanticen únicamente prestaciones en caso de muerte, cuando estas prestaciones se satisfagan en especie o cuando el importe de las mismas no exceda del valor medio de los gastos funerarios por fallecimiento.

Si a la entidad aseguradora se le concede la autorización simultánea para varios ramos dentro de los seguros directos distintos del de vida, la denominación será la que se indica en el cuadro siguiente:

RAMOS DE AUTORIZACIÓN SIMULTÁNEA	DENOMINACIÓN DE LA AUTORIZACIÓN
1 y 2	Accidentes y enfermedad
Cobertura de ocupantes de vehículos del ramo 1 junto con los ramos 3, 7 y 10	Seguro de automóviles
Cobertura de ocupantes de vehículos del ramo 1 junto con los ramos 4, 6, 7 y 12	Seguro marítimo y de transporte
Cobertura de ocupantes de vehículos del ramo 1 junto con los ramos 5, 7 y 11	Seguro de aviación
8 y 9	Incendio y otros daños a los bienes
10, 11, 12 y 13	Responsabilidad civil
14 y 15	Crédito y caución
Todos los ramos	Seguros generales

2.1.3. Características de los seguros no vida

Teniendo en cuenta la estructura de los seguros no vida, su estudio se aleja bastante del enfoque clásico de los seguros de vida. Algunas de sus características generales en comparación con los de vida, son [Nieto y Vegas (1993)]:

a) Son operaciones a corto plazo. En la mayoría de los casos la duración es anual, renovable tácitamente, por lo que el tipo de interés no juega, en general, un papel tan importante como en los seguros de vida.

b) El problema de la tarificación es complejo. El número de factores de riesgo que influyen en las probabilidades de los sucesos que dan lugar al pago de las indemnizaciones es mayor que en los seguros de vida, los cuales dependen de un número más limitado (usualmente edad y sexo del asegurado) [Almer (1957); Franckx (1974); Hey (1970)].

c) El entorno socio-económico influye en gran medida en la evolución temporal de la siniestralidad, especialmente en algunos ramos y modalidades (responsabilidad civil obligatoria del automóvil, robo, etc) a diferencia de lo que ocurre en los seguros de vida (excepto en la contingencia de invalidez);

d) Los problemas de estabilidad en la fluctuación respecto a los valores esperados (primas puras) son más complejos que en los seguros de vida, donde además la existencia de reservas matemáticas juega un papel estabilizador;

e) Las primas cubren, normalmente, el riesgo del período correspondiente y no llevan incorporada la componente de ahorro como en los seguros de vida.

f) Las prestaciones o indemnizaciones están en función de la cuantía del daño. Éste a su vez viene dado por un variable aleatoria (causal en el seguro de vida). Por lo que, a diferencia de los seguros de vida, se presentan problemas de infraseguro y sobreseguro cuando la suma asegurada no coincide con el valor del interés asegurado.

2.1.4. El seguro del automóvil

La sociedad actual ha impuesto un nivel de actividad en el que los desplazamientos son parte

fundamental. Por esto, la utilización de coches, camiones o motos se ha convertido en un elemento casi indispensable para el desarrollo de nuestra actividad diaria. Sin embargo, el gran número de vehículos que circulan por nuestras calles y carreteras generan unos riesgos en los que todos podemos vernos afectados. Aunque seamos unos magníficos conductores nada ni nadie puede garantizarnos la inexistencia de golpes o accidentes, ya que en éstos, además de la imprudencia, también juegan una baza importante factores difíciles de controlar. En unos casos podremos causar un daño físico o material a terceras personas, y en otros, seremos nosotros los directamente perjudicados por un siniestro. Este estado de riesgo permanente que genera la utilización de vehículos, unido a los elevados costes económicos y sociales que provocan los accidentes de tráfico, hizo necesario el establecimiento obligatorio de un sistema reparador de los daños causados por dicha utilización: el seguro del automóvil.

El seguro de responsabilidad civil de vehículos terrestres a motor (el seguro "a terceros") protege al asegurado de las consecuencias económicas adversas que pueda ocasionar en su patrimonio la reclamación, por parte de un tercero, de los daños provocados por su automóvil. Y, al tiempo, garantiza que toda víctima de un accidente va a ser reparada en los daños que se le causen.

El sistema de responsabilidad civil imperante en todos los países hasta finales del siglo XIX se basaba en la responsabilidad civil subjetiva o por culpa. La revolución industrial y la aparición de nuevas máquinas cada vez más rápidas y potentes, pero susceptibles de generar más y más graves accidentes, fueron factores que demandaban un cambio. Así, se dio paso a la paulatina adopción del denominado principio de responsabilidad objetiva o por riesgo, que es aquella que surge sin que exista necesariamente culpa de la persona causante del daño.

La responsabilidad civil objetiva

El fundamento de la responsabilidad objetiva está en el riesgo que suponen por sí mismas ciertas actividades (por ejemplo, la mera posesión y uso de una máquina llamada automóvil). Por lo tanto, se garantiza la reparación efectiva de los daños sufridos, aún cuando no exista culpa o negligencia de quien los causó. La transición de responsabilidad civil de tipo subjetiva a la de carácter objetivo, en materia de circulación de vehículos a motor, es trascendental. Pone de manifiesto la importancia que en el orden económico y social reviste el seguro como institución, puesto que sin la existencia del seguro no podría darse el paso que convierte unas obligaciones definidas por la existencia de culpa en un derecho a reparación que es independiente de dicha culpa.

La presencia del seguro ratifica el carácter social del mismo, ya que para producirse la reparación del daño no se precisa que el causante haya cometido delito alguno. Por el contrario, si fuesen los propios conductores los que resolviesen sus problemas, habría víctimas de siniestros causados por personas que no podrían reparar el daño causado. Con la existencia del seguro obligatorio todas las víctimas cobran, independientemente del patrimonio que tenga quien les provocó el daño. En definitiva, el seguro del automóvil es uno de los mayores logros sociales del siglo XX. Es un sistema que transfiere cada año, aproximadamente, cinco mil cuatrocientos nueve millones de euros a víctimas de accidentes de tráfico. No en vano, actualmente hay cerca de veinte millones de pólizas de autos emitidas. Hoy en día, el seguro del automóvil está adaptado para atender a todas las necesidades del conductor, que van más allá de la cobertura del riesgo, convirtiéndose de esta forma en un servicio integral.

Antecedentes históricos del seguro del automóvil

Se estima que el primer antecedente relacionado con el seguro de responsabilidad civil de automóviles es una Ordenanza de Policía dictada por el Prefecto de París en 1821. Cada cochero tenía que destinar veinte céntimos diarios de su salario a la creación de un fondo que pagase las multas y la reparación de daños causados a terceros. A partir de 1825 se forman algunas entidades para el seguro de responsabilidad civil de caballos y coches, como la francesa "L`Automedon", sociedad dedicada especialmente a este seguro.

El mecanismo del seguro fue aceptado por los Tribunales de Justicia. No obstante, entre los juristas se creó una corriente negativa, ya que pensaban que si los conductores de los carruajes tuviesen un seguro que cubriese sus posibles negligencias, no prestarían la suficiente atención para evitar accidentes. Este argumento fue aceptado por el Tribunal de Comercio del Sena, que decretó la nulidad de este seguro de responsabilidad civil en 1844. La compañía "L`Automedon" apeló la resolución ante el Tribunal de Casación. Éste revocó la sentencia anteriormente dictada y decretó la licitud del seguro de responsabilidad civil. Dicha resolución marcó un hito muy importante y fue la base del desarrollo posterior del seguro de responsabilidad civil.

Cuando aparecieron los vehículos de motor el seguro se adaptó a esta forma de locomoción. En algunos países, dada la proliferación de los automóviles y el riesgo evidente que éstos crean, se hizo obligatoria la contratación de un seguro de responsabilidad civil. En España, la primera normativa relativa al seguro del automóvil es la Ley sobre Uso y Circulación de Vehículos a Motor de 1962.

Legislación hasta la actualidad en referencia al seguro del automóvil en España

Recopilamos a continuación las referencias de la legislación española básica sobre el seguro del automóvil:

- **Decreto 632/1968, de 21 de marzo,** por el que aprueba el Texto Refundido de la Ley 122/1962, de 24 de diciembre, sobre responsabilidad civil y seguro en la circulación de vehículos a motor (BOE nº 85, de 8 de abril de 1968)
- **Real Decreto 2.641/1986, de 30 de diciembre,** por el que se aprueba el Reglamento del seguro de responsabilidad civil derivada del uso y circulación de vehículos a motor, de suscripción obligatoria (BOE nº 313, de 31 de diciembre de 1986; corrección de errores en BOE nº 18, de 21 de enero de 1987)
- **Real Decreto 447/1986, de 10 de enero,** por el que se adoptan las medidas provisionales necesarias para la adaptación del Seguro Obligatorio de Automóviles a las exigencias de la adhesión a la Comunidad Económica Europea (BOE nº 53, de 3 de marzo de 1986)
- **Orden de 23 de abril de 1987,** sobre requisitos del modelo para probar la existencia del seguro de responsabilidad civil derivada del uso y circulación de vehículos de motor, de suscripción obligatoria (BOE nº 104, de 1 de mayo de 1987; corrección de errores en BOE nº137, de 9 de junio)
- **Orden de 25 de septiembre de 1987,** por la que se dictan normas relativas al funcionamiento de la Oficina Española de Aseguradores Automóviles (OFESAUTO) (BOE nº 247, de 15 de octubre de 1987)
- **Convenio tipo Interbureaux para la ejecución de las estipulaciones de las "Recomendaciones de Ginebra",** adoptado por el Council of Boreaux en Asamblea General los días 19 y 20 de octubre de 1989
- **Convenio Multilateral de Garantía entre Oficinas Nacionales de Seguros de 15 de marzo de 1991:** Decisión de la Comisión 91/323/CEE, de 30 de mayo de 1991, sobre la aplicación de la Directiva 72/166/CEE, del Consejo, relativa a la aproximación de las legislaciones de los Estados miembros sobre el seguro de la responsabilidad civil que resulte de la circulación de vehículos automóviles y el control de la obligación de asegurar esta responsabilidad (DOCE nº L 177, de 5 de julio de 1991)
- **Convenio sobre la ley aplicable en materia de accidentes de circulación por carretera,** hecho en La Haya el 4 de mayo de 1971 (BOE nº 264, de 4 de noviembre de 1987; corrección de errores en BOE nº 307, de 24 de diciembre)
- **Real Decreto 7/2001, de 12 de enero**, por el que se aprueba el Reglamento sobre la responsabilidad civil y seguro en la circulación de vehículos a motor.
- **Resolución de 20 de enero de 2003**, de la Dirección General de Seguros y Fondos de Pensiones, por la que se da publicidad a las cuantías de las indemnizaciones por muerte, lesiones permanentes e incapacidad temporal, que resultarán de aplicar durante 2003 el sistema para la valoración de los daños y perjuicios causados a las personas en accidentes de circulación.

2.1.4.1. El seguro de suscripción obligatoria

Todo propietario de un vehículo a motor está obligado a contratar y mantener en vigor una póliza de seguro que cubra, hasta la cuantía que en cada momento se determine, la responsabilidad civil del

conductor que se derive de los daños, tanto personales como materiales, ocasionados a terceras personas como consecuencia de un hecho de la circulación.

Esta obligación impuesta al propietario del vehículo contrasta con la designación del conductor como sujeto responsable. Es decir, aunque el propietario del vehículo sea el tomador del seguro, el sujeto asegurado es el conductor del mismo, porque lo que se cubre no es la responsabilidad del propietario, sino la del conductor.

La vigencia del seguro se acredita exclusivamente con el recibo de prima correspondiente al periodo de seguro en curso. Dicho recibo debe llevarse siempre en el interior del vehículo ya que los agentes de tráfico pueden llegar a pedírnoslo. Si en el momento en que se nos solicita no disponemos de la documentación acreditativa del seguro seremos sancionados con 60.10 € (diez mil pesetas) de multa.

Coberturas del seguro obligatorio

El Seguro de Suscripción Obligatoria (SOA) garantiza la cobertura de la responsabilidad civil del conductor de cualquier vehículo terrestre a motor con estacionamiento habitual en España, mediante el pago de una sola prima, en todo el territorio del Espacio Económico Europeo y de los estados adheridos al Convenio Multilateral de Garantía.

Hoy en día, existen dos convenios internacionales entre oficinas nacionales: el tradicional Convenio de la Carta Verde y el Convenio Multilateral de Garantía (sistema de las directivas europeas).

En 1953, comenzó a operar el Sistema de Carta Verde, patrocinado por el Consejo Económico para Europa de la ONU. Mediante la presentación de la Certificación Internacional de Seguro, llamada carta verde por su color, se facilitaba el tránsito de vehículos entre países ya que garantizaba la existencia de un seguro de responsabilidad civil según la legislación obligatoria de cada país.

Para hacer efectivo este sistema, se crea en cada país una oficina nacional o bureau (en España OFESAUTO) que se responsabiliza frente a sus autoridades de las consecuencias de siniestros causados por vehículos extranjeros debidamente garantizados en su Carta Verde.

La Oficina Española de Aseguradores de Automóviles (OFESAUTO) agrupa obligatoriamente a todas las entidades aseguradoras autorizadas para operar en el ramo de responsabilidad civil de vehículos terrestres automóviles y al Consorcio de Compensación de Seguros. Sus funciones se desenvuelven en dos planos:

- Accidentes en el extranjero con vehículos matriculados en España garantizados mediante una carta verde emitida por una aseguradora asociada con la autorización de OFESAUTO.
- Accidentes ocurridos en España causados por vehículos matriculados en los países del Convenio Multilateral de Garantía o bien garantizados por carta verde emitida bajo autorización de una oficina nacional adherida al sistema de carta verde.

En los accidentes ocurridos en España causados por vehículos no nacionales, OFESAUTO, cuyos servicios son gratuitos, facilitará los medios para obtener la indemnización que le pudiera corresponder, bien informando sobre el representante del asegurador causante del daño, bien prestando esta atención directamente. Una vez confirmado el estacionamiento habitual del vehículo causante del accidente en España y la responsabilidad de su conductor, la oficina española procederá a valorar los daños y perjuicios que deben ser indemnizados por cuenta de la oficina nacional donde el vehículo está matriculado.

Por lo que respecta a los accidentes causados por vehículos españoles en el extranjero, OFESAUTO garantiza que tales daños están debidamente cubiertos por un seguro de responsabilidad civil de vehículos a motor, o, en su caso, por la cobertura del Consorcio de Compensación de Seguros como fondo nacional de garantía.

El Servicio de Riesgos Especiales de Automóviles (SEREA) es una agrupación voluntaria de entidades aseguradoras que practican el seguro del automóvil. Su misión principal es asumir riesgos que por su naturaleza se consideran como agravados. Algunos ejemplos de tales riesgos son las pruebas deportivas, los vehículos de bomberos, los autocares y las ambulancias.

Además, también se encarga de la contratación del seguro de frontera para aquellos vehículos que deseen entrar en España y, por su país de procedencia (no perteneciente a la Comunidad Europea, ni a un país adherido a los convenios complementarios) necesitan tener una carta verde. A diferencia de OFESAUTO, la pertenencia al SEREA es de carácter voluntario. Asimismo, su cobertura se extiende tanto a los seguros de responsabilidad civil de suscripción obligatoria como a los de suscripción voluntaria.

Siguiendo con el SOA, en lo que se refiere a los daños personales o corporales ocasionados a terceros, tiene una responsabilidad civil de carácter objetivo, es decir, sin culpa. Dicha responsabilidad sólo quedará exonerada cuando se pruebe que los daños fueron debidos únicamente a la conducta o

negligencia del perjudicado, o a fuerza mayor extraña a la conducción o al funcionamiento del vehículo. No se considerarán casos de fuerza mayor los defectos del vehículo, ni la rotura o fallo de alguna de sus piezas.

Este seguro cubre la responsabilidad civil del conductor frente a terceros, pero no los daños personales ni materiales que éste sufra. La cantidad máxima con que la aseguradora indemnizará a cada tercero víctima de un accidente es actualmente de trescientos cincuenta mil € (58 235 100 pesetas) para daños personales. A través del Sistema de Valoración del Daño Corporal en Accidentes de Tráfico, habitualmente conocido como el Baremo, se determina la cuantía de las indemnizaciones a pagar por las lesiones provocadas, así como los días de baja laboral derivados de las mismas. Además de las indemnizaciones fijadas con arreglo a las tablas del Baremo, el SOA cubre los gastos de asistencia médica y hospitalaria, así como, en las indemnizaciones por muerte, los gastos de entierro y funeral. La cobertura de los gastos sanitarios será ilimitada hasta la plena sanación y recuperación de los heridos, con independencia del centro sanitario en el que hayan sido tratados, y siempre que el gasto esté debidamente justificado atendiendo a la naturaleza de la asistencia prestada.

En cuanto a los daños materiales producidos a las cosas o animales, el conductor responderá frente a terceros si existe culpa o negligencia por su parte. En estos casos, la cantidad máxima con que cada asegurador indemnizará a los terceros perjudicados por un accidente será de cien mil € (16 638 600 pesetas) por siniestro, no por víctima. De esta forma, si en un mismo accidente amparado por un único asegurador resultan varios perjudicados por daños materiales, y éstos exceden del límite fijado, cada perjudicado verá reducida su indemnización en proporción a los daños sufridos.

El SOA no cubre los daños sufridos ni por el vehículo asegurado, ni por el conductor del mismo, pero sí lo hace con el resto de ocupantes, ya que éstos son "terceros". Además, si el vehículo hubiera sido robado, los daños personales y materiales producidos con motivo de su circulación serán indemnizados por el Consorcio de Compensación de Seguros. Este seguro obligatorio tampoco cubre los desperfectos ocasionados en las cosas transportadas en el vehículo asegurado, ni en los bienes de los que sean titulares el tomador, asegurado, propietario o conductor, así como los del cónyuge o los parientes hasta el tercer grado de consanguinidad de éstos.

El derecho de repetición

En el SOA el asegurador, una vez efectuado el pago de la indemnización, podrá repetir contra el conductor, el propietario del vehículo causante y el asegurado, si, entre otros casos, los daños

materiales y personales causados fuesen debidos a la conducción bajo la influencia de bebidas alcohólicas o de drogas tóxicas, estupefacientes o sustancias psicotrópicas.

Es decir, la compañía aseguradora pagará la indemnización correspondiente a la víctima de un accidente ocurrido con motivo de alguno de los supuestos anteriores, pero luego, podrá reclamar el pago de ese importe al conductor, al propietario y al asegurado.

Ámbito de cobertura del seguro obligatorio

- Todo el territorio del Espacio Económico Europeo.
- Estados adheridos al Convenio Multilateral de Garantía: Alemania, Austria, Bélgica, Croacia, República Checa, Dinamarca, República Eslovaca, Eslovenia, España, Finlandia, Francia, Gran Bretaña, Grecia, Holanda, Hungría, Irlanda, Islandia, Italia, Luxemburgo, Noruega, Portugal, Suecia y Suiza.

Y cualquier otro estado que determine el Ministerio de Economía y Hacienda.

Lista de países en los que se necesita Carta Verde

Albania, Andorra, Bosnia-Herzegovina, Bulgaria, Chipre, Estonia, Irak, Irán, Israel, Letonia, Macedonia, Malta, Marruecos, Polonia, Rumanía, Túnez, Turquía, Ucrania y Yugoslavia.

2.1.4.2. Responsabilidad civil voluntaria

Cuando la indemnización que corresponde al tercero o perjudicado excede a los capitales fijados para el seguro obligatorio de automóviles, el conductor responsable será quien, con su patrimonio, deba hacer frente al exceso de indemnización fijada. Las indemnizaciones dictadas por los tribunales son frecuentemente muy superiores a los límites contemplados por el seguro de responsabilidad civil obligatoria. Ello justifica la necesidad de dar cobertura a esos excesos de indemnización mediante un seguro complementario y de suscripción voluntaria.

Las entidades aseguradoras ofrecen la posibilidad de contratar un seguro de responsabilidad civil que amplía la cobertura indemnizatoria del seguro obligatorio. Esta garantía del seguro voluntario cubre, dentro de los límites pactados o ilimitadamente, las indemnizaciones que excedan del límite de la

responsabilidad civil de suscripción obligatoria.

La mayoría de las entidades aseguradoras ofrecen en sus pólizas, siempre que se haya contratado la cobertura de responsabilidad civil voluntaria, la garantía de defensa penal, fianzas y reclamaciones, también llamado seguro de protección jurídica del automovilista.

Este seguro cubre, en las causas penales dirigidas contra el conductor o propietario del vehículo, los honorarios del abogado y los derechos del procurador cuando su intervención sea preceptiva, incluyéndose el coste de los poderes procesalmente necesarios. Asimismo, también cubre la constitución de fianzas exigidas por la autoridad judicial para garantizar el pago de costas procesales y la libertad condicional.

Por lo que respecta a la reclamación de daños y perjuicios causados por un tercero al vehículo asegurado y a las personas que en él viajen, el seguro de protección jurídica del automovilista cubre los trámites y gestiones en vía amistosa, la asistencia jurídica para la reclamación vía arbitral o judicial, los gastos de peritación de los daños materiales ocasionados en el vehículo y la atención directa de esos daños reclamados cuando se haya obtenido la conformidad de pago por la entidad aseguradora responsable.

El asegurado tendrá derecho a elegir libremente el procurador que le represente y el abogado que le defienda desde el momento de la declaración del siniestro. Estos profesionales gozarán de una amplia libertad en la dirección técnica del asunto.

2.1.4.3. Coberturas complementarias del seguro del automóvil

Si además de la responsabilidad civil también aseguramos otros riesgos que afectan al vehículo, podemos decir que el seguro del automóvil es en realidad una póliza combinada o multi-riesgo. Eligiendo bien las distintas posibilidades que nos ofrecen las entidades aseguradoras lograremos un producto adaptado a nuestras necesidades reales. A continuación citaremos algunas de estas coberturas que podemos añadir a nuestra póliza del seguro del automóvil para lograr unas garantías completas, es decir, "un todo riesgo".

Es especialmente importante para el conductor, porque está excluido del seguro de responsabilidad civil, contratar la **cobertura de accidentes personales**, también llamada de ocupantes de vehículos.

A través de esa cobertura, se asegura una indemnización en caso de fallecimiento, lesiones o invalidez como consecuencia de un accidente de circulación. De la misma manera, también están garantizados los gastos de asistencia sanitaria que sean necesarios y el traslado en ambulancia.

Los **daños propios del vehículo**, la **rotura de lunas** y los riesgos de **robo** e **incendio** completan la lista de posibilidades que las entidades aseguradoras ponen a disposición de sus clientes.

La cobertura de daños propios en el vehículo asegurado cubre, hasta los límites fijados en las condiciones particulares de la póliza, los daños producidos en el vehículo como consecuencia de un accidente. Esta garantía puede abarcar la totalidad de los daños, o bien establecerse algún tipo de franquicia.

Algunas entidades aseguradoras ofrecen contratar esta garantía limitando su cobertura a los daños sufridos por el vehículo asegurado como consecuencia de la colisión con vehículos, personas o animales, siempre que las personas o los propietarios de los vehículos o animales resulten identificados. También puede limitarse exclusivamente a la pérdida total del vehículo.

Además, la garantía de daños propios se hace cargo de los gastos que se ocasionen por el transporte del vehículo al taller más cercano al accidente.

La cobertura de incendios suele contratarse junto con la anterior. Comprende tanto la combustión o abrasamiento con llama como la explosión y la caída del rayo.

Por lo que respecta a la garantía de robo del vehículo, el asegurador se obliga, dentro de los límites de la póliza, a indemnizar al asegurado en caso de sustracción ilegítima del vehículo por terceras personas. Los riesgos excluidos en esta modalidad son los siguientes:

- La sustracción que tenga su origen en negligencia grave del asegurado, del tomador, o de las personas que de ellos dependan o convivan con ellos. Un ejemplo de esta negligencia puede ser el caso en el que se dejan las llaves en el interior del vehículo, estando éste sin cerrar.

- Las sustracciones de que fueran autores, cómplices o encubridores los familiares del asegurado o del tomador del seguro, hasta el tercer grado de consanguinidad o afinidad, o los dependientes o asalariados de cualquiera de ellos.

La indemnización de la cobertura de robo del vehículo y sus accesorios se determina en el contrato. La más frecuente es:

SITUACIÓN	INDEMNIZACIÓN
Sustracción vehículo completo	80% de su valor venal*
Sustracción piezas fijas del vehículo	80% de su valor de nuevo
Batería/neumáticos	80% de su valor venal
Daños: por aparición del vehículo con éstos o por intento de robo	80% del importe de la reparación
Accesorios	De tenerlos expresamente incluidos, 80% de su valor a nuevo

* El valor venal es el valor de venta del vehículo asegurado inmediatamente antes de la ocurrencia del siniestro.

En caso de robo del vehículo, el asegurado deberá denunciarlo ante las autoridades competentes. Si el vehículo aparece en el plazo fijado por la póliza, el asegurado está obligado a admitir su devolución. Si, por el contrario, el vehículo aparece con posterioridad al plazo fijado en la póliza, éste quedará en manos del asegurador. Si el asegurado quisiera recuperar su coche, tendría que devolver la indemnización percibida por parte de la entidad aseguradora.

Por último, a través de la cobertura de rotura de lunas se garantizan los gastos de colocación y reposición de las lunas originales del vehículo asegurado, en caso de rotura de las mismas.

Coberturas del Consorcio de Compensación de Seguros referentes al seguro de automóviles

- Indemnizar los siniestros ocurridos en España, cuando el vehículo causante sea desconocido.
- Indemnizar los siniestros producidos por vehículos con estacionamiento habitual en España, que estando asegurados, hayan sido robados.
- Indemnizar los siniestros de los casos anteriores, cuando exista controversia entre el Consorcio y una entidad, sin perjuicio de que una vez resuelta en contra de la entidad, ésta abone al Consorcio el importe de la indemnización, con el recargo correspondiente.
- Indemnizar los siniestros, cuando la entidad aseguradora hubiera sido declarada en quiebra, suspensión de pagos o en situación de liquidación intervenida o asumida por la Comisión Liquidadora de Entidades Aseguradoras.

2.2. Tarificación

2.2.1. Proceso de riesgo

Consideremos una cartera de riesgos, C, en un intervalo temporal determinado τ, generalmente $\tau = [0,1)$ correspondiente a la duración de un año. Nos interesa estudiar la siniestralidad de C respecto de un determinado riesgo que será el que cubrirá la operación de seguro contemplada.

Se denomina *proceso de riesgo* al proceso estocástico que está asociado al acaecimiento de los siniestros y a sus respectivas cuantías.

Sea $\{S_t\}_{t\in\tau}$ el proceso estocástico cuantía total, siendo S_t, para $t \in \tau$, la variable aleatoria cuantía total de los siniestros ocurridos en $[0,t]$. Llamemos:

- N: variable aleatoria número de siniestros del intervalo τ
- X_i: variable aleatoria cuantía del siniestro i-ésimo, para $i = 1,2,\ldots,N$
- t_i: variable aleatoria instante de ocurrencia del siniestro i-ésimo, con $t_i \in \tau$, para $i = 1,2,\ldots,N$

El valor en 0 de las indemnizaciones a realizar por la compañía o valor actual del coste total por los siniestros ocurridos en τ, es:

$$\begin{cases} X_1 v^{t_1} + X_2 v^{t_2} + \cdots + X_N v^{t_N} & \text{si } N > 0 \\ 0 & \text{si } N = 0 \end{cases} \qquad (2.1)$$

siendo $v^{t_i} = (1+I_1)^{-t_i}$ los factores de actualización financiera de t_i a 0, para $i = 1,2,\ldots,N$.

Observamos que además de todas las componentes aleatorias (cuantías, X_i, tiempos, t_i, y número de siniestros, N), mirado desde 0, necesitamos fijar *a priori* el tipo de interés, I_1.

Como generalmente el plazo para la tarificación es inferior o igual al año, se desprecia el efecto de la actualización financiera, pues los resultados no se alterarían significativamente. Además el modelo se simplifica y se hace más aplicable.

Resultando el *coste total* por indemnizaciones de siniestros acaecidos, S, en el período τ, la suma

aritmética:

$$S = \begin{cases} X_1 + X_2 + \cdots + X_N & \text{si } N > 0 \\ 0 & \text{si } N = 0 \end{cases} \tag{2.2}$$

que depende de las cuantías de los sumandos y del número de sumandos.

Hipótesis del proceso de riesgo

Las hipótesis que se realizan usualmente para el proceso de riesgo en la teoría colectiva son:

- **Equidistribución**: la distribución de probabilidad de las cuantías X_i, $i = 1,2,\ldots,N$, no cambia con el número que corresponde al orden de ocurrencia de siniestro, $F_X = F_{X_i} = F_{X_j}$ $\forall i, j \in \{1,2,\ldots,N\}$, y adicionalmente se supone $X > 0$, no se contempla la posibilidad de cuantías negativas.

- **Independencia entre cuantías**: las distribuciones de probabilidad de las cuantías $X_i, X_j \quad \forall i, j \in \{1,2,\ldots,N\}$ se suponen independientes, $F_{(X_i,X_j)}(x_i,x_j) = F_{X_i}(x_i)F_{X_j}(x_j)$.

- **Independencia entre el coste por siniestro y el número de siniestros**: la distribución de probabilidad del coste por siniestro y la del número de siniestros se suponen independientes, $F_{(X_i,N)}(x_i,n) = F_{X_i}(x_i)F_N(n)$.

Esperanza y varianza del coste total

A partir de estas hipótesis podemos calcular fácilmente la esperanza y la varianza del coste total en un período τ, a partir de las esperanzas y varianzas del número de siniestros y del importe de cada siniestro.

- **Esperanza**:

  El coste total es la suma de un número aleatorio de sumandos aleatorios, $S = \sum_{i=1}^{N} X_i$. Para calcular su esperanza hacemos uso de la formulación de la esperanza condicionada (véase Ross (1989) pp. 93-102):

$$E[S]=E\left[\sum_{i=1}^{N}X_i\right]=E\left[E\left[\sum_{i=1}^{N}X_i \mid N\right]\right] \tag{2.3}$$

Si se supone $N = n$, se tiene que $E\left[\sum_{i=1}^{N}X_i \mid N=n\right]=E\left[\sum_{i=1}^{n}X_i \mid N=n\right]=E\left[\sum_{i=1}^{n}X_i\right]$. Teniendo en cuenta las hipótesis de independencia y equidistribución del proceso de riesgo resulta que $E\left[\sum_{i=1}^{n}X_i\right]=nE[X]$.

De ello, se deduce que $E\left[\sum_{i=1}^{N}X_i\right]=NE[X]$. Por lo que tomando esperanzas,

$$E\left[E\left[\sum_{i=1}^{N}X_i \mid N\right]\right]=E[NE[X]]=E[N]E[X],$$

i.e.,

$$E[S]=E[N]E[X]. \tag{2.4}$$

- **Varianza**:

$$Var(S)=Var\left[\sum_{i=1}^{N}X_i\right]=E\left[\left(\sum_{i=1}^{N}X_i\right)^2\right]-\left(E\left[\sum_{i=1}^{N}X_i\right]\right)^2 \tag{2.5}$$

La esperanza condicionada es también útil en el cálculo de la varianza del coste total. Utilizaremos la formulación condicionada para rescribir el primero de los sumandos:

$$E\left[\left(\sum_{i=1}^{N}X_i\right)^2\right]=E\left[E\left[\left(\sum_{i=1}^{N}X_i\right)^2 \mid N\right]\right],$$

$$E\left[\left(\sum_{i=1}^{N}X_i\right)^2 \mid N=n\right]=Var\left(\sum_{i=1}^{n}X_i\right)+\left(E\left[\sum_{i=1}^{N}X_i\right]\right)^2=nVar(X)+\left(nE[X]\right)^2,$$

por lo que $E\left[\left(\sum_{i=1}^{N} X_i\right)^2 \mid N\right] = NVar(X) + N^2\left(E[X]\right)^2$ y tomando esperanzas en ambos lados:

$$E\left[\left(\sum_{i=1}^{N} X_i\right)^2\right] = E[N]Var(X) + E\left[N^2\right]\left(E[X]\right)^2.$$

Así, sustituyendo, obtenemos,

$$\begin{aligned} Var\left(\sum_{i=1}^{N} X_i\right) &= E[N]Var(X) + E\left[N^2\right]\left(E[X]\right)^2 - \left(E\left[\sum_{i=1}^{N} X_i\right]\right)^2 \\ &= E[N]Var(X) + E\left[N^2\right]\left(E[X]\right)^2 - \left(E[N]E[X]\right)^2 \\ &= E[N]Var(X) + \left(E[X]\right)^2\left(E\left[N^2\right] - \left(E[N]\right)^2\right) \\ &= E[N]Var(X) + \left(E[X]\right)^2 Var(N), \end{aligned}$$

i.e.,

$$Var[S] = E[N]Var(X) + \left(E_X[X]\right)^2 Var(N). \tag{2.6}$$

2.2.2. Estructura de la prima

Supongamos que disponemos de una cartera infinita cuyos elementos son idénticos, es decir, que observamos el mismo riesgo, de un ramo de no vida. Si estudiamos el coste total por póliza, su valor esperado viene dado por (2.4), siendo N el número de siniestros por póliza en un período (de un año). En tal caso, tenemos que

$$P = E[N]E[X] \tag{2.7}$$

es la **prima pura con bases de segundo orden** (véase esquema p. 29). Cambiamos el riesgo aleatorio de la póliza, que puede tomar cualquier valor positivo o nulo, por un valor cierto fijo, P, que coincide con la esperanza matemática del coste total, $E[S]$.

La prima pura, P, es la componente base del precio del seguro, pues con esta parte de la prima, la entidad aseguradora tendrá que acumular la cantidad suficiente de recaudación para hacer frente a los

siniestros previstos y esperados[2]. Si dispusiéramos sólo de un riesgo sería difícil cambiarlo por la esperanza, pero al suponer la cartera infinita, unos riesgos se compensarán con otros. Su cálculo estará basado en información estadística propia de cada ramo, y en su caso, de cada producto o modalidad de seguro.

Pero, en la práctica, no se dispone ni de una cartera infinita, ni de información estadística suficiente y fiable, por lo que la **prima de riesgo** se corresponde con la esperanza matemática más un recargo de seguridad. El recargo de seguridad tiene como finalidad constituir una reserva técnica que permita absorber con cargo a la misma los excesos de siniestralidad que se hayan podido producir en un determinado período.

Se distingue [Nieto y Vegas (1993)]:

⇒ Recargo *implícito*, o tarificación con bases de primer orden

⇒ Recargo *explícito*, o tarificación con bases de segundo orden

En los casos en que el recargo de seguridad está incluido en la prima pura, recargo de seguridad implícito, se dice que se trata de una **prima pura con bases de primer orden**. El recargo implícito surge generalmente cuando los aseguradores aplican tarifas uniformes. En tal caso se supone que la primas puras no responden a principios técnicos de acuerdo con la estructura de cada empresa, por lo que la solvencia se consigue a un precio elevado del servicio de seguridad. Es decir, el precio es común para todas las entidades con independencia de las restantes magnitudes de estabilidad o del tamaño y composición de la cartera, por lo que no se verifican los principios de equidad y suficiencia, que en inmediato detallamos, y entonces la prima figura como una variable exógena en el modelo y no en función de los costes técnicos del asegurador.

Por el contrario, el recargo explícito, λ, supone una evolución hacia criterios de eficacia económica, en donde dentro de un marco de competencia se produce el fenómeno de que los costes técnicos (la siniestralidad esperada y la media de sus fluctuaciones respecto a su media), junto con las magnitudes de estabilidad de la propia empresa determinan los precios o **primas de riesgo**, P_1. Este recargo se destina a cubrir las desviaciones aleatorias de la siniestralidad con respecto a su valor medio, por lo

[2] Al suponer una cartera infinita, bajo la Ley de los grandes números, la esperanza matemática deberá ser suficiente para cubrir los riesgos esperados en el periodo, pues estaremos exigiendo que exista un número elevado de sujetos expuestos al mismo riesgo que hará que el sistema sea viable.

que su cálculo dependerá de las restantes magnitudes de estabilidad de la entidad (reaseguro, reservas de solvencia), es decir tiene por objeto financiar las fluctuaciones negativas de la siniestralidad y contribuye a garantizar la solvencia del asegurador.

Es posible recargar explícitamente la prima pura para obtener la prima de riesgo, P_1, con diferentes criterios [Heilmann (1988), capítulo 4]. Usualmente el recargo se fija proporcional a la prima pura:

$$P_1 = P(1+\lambda) = P + \lambda P . \tag{2.8}$$

A este criterio se le denomina criterio de la esperanza, pues el recargo está en función de ésta. Otros criterios de recargo podrían ser: criterio de la varianza; de la desviación estándar; de la semi-varianza; del percentil; de la máxima pérdida; de la esperanza de la utilidad nula; etc.

El precio que un asegurado paga finalmente por su riesgo se compone de diferentes elementos:

- Recargo para gastos de gestión interna, g_1 : porción destinada a gastos de gestión interna para cubrir gastos propios de la gestión de la póliza dentro de la compañía aseguradora, es decir, para financiar los gastos de administración como sueldos, cargas sociales, amortizaciones, gastos generales, etc. Cabe notar que los gastos de gestión de los siniestros se incluirán en todo caso en la prima pura. Los gastos de gestión interna pueden ser fijos sin depender de la cuantía de la prima, o proporcionales, en tanto por uno, al importe de la prima de tarifa. A efectos de la periodificación contable de los ingresos de primas, la prima resultante al añadir esta componente se define como **prima de inventario**, P',

  $$P' = P_1 + g_1 , \tag{2.9}$$

  o bien si el recargo es proporcional,

  $$P' = P_1 + g_1 P'' . \tag{2.10}$$

- Recargo para gastos de gestión externa, g_2 : proporción destinada a gastos de gestión externa calculada habitualmente proporcional, en tanto por uno, a la prima de tarifa para cubrir gastos externos de producción o adquisición como comisiones de la red comercial destinadas a remunerar a los mediadores entre asegurados y aseguradores, gastos de adquisición, y gastos de cobro y mantenimiento de la cartera.

- Recargo para beneficios o excedente, β : beneficio directamente para la compañía o excedente que se destinará a remunerar los recursos financieros e incrementar la solvencia dinámica de la empresa. Representarán habitualmente un tanto por uno respecto de la prima de tarifa, en función del margen de beneficio que se quiera conseguir. La prima resultante al añadir las dos componentes se denomina **prima de comercial o de tarifa**, P'',

$$P'' = P' + g_2 P'' + \beta P''. \tag{2.11}$$

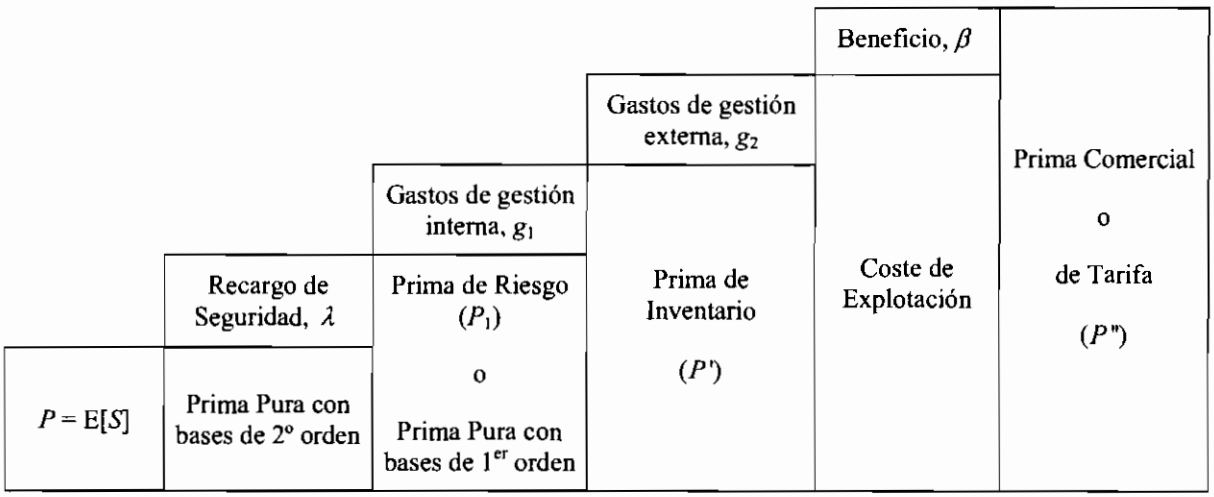
Si los gastos de gestión interna son proporcionales, tenemos que la prima de tarifa se obtendría como:

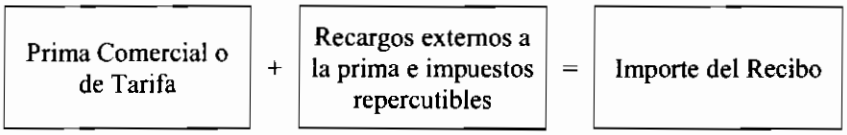
$$P'' = \frac{P + \lambda P}{1 - g_1 - g_2 - \beta}. \tag{2.12}$$

Además, en España, sobre las primas comerciales giran, en aquellos casos en que así está establecido, los recargos para la Comisión Liquidadora de Entidades Aseguradoras (CLEA) y el Consorcio de Compensación de Seguros, así como los impuestos y tasas repercutibles. También la legislación española prevé la posibilidad de establecer un recargo externo a la prima comercial, destinado a compensar las modificaciones que puedan ocurrir en los gastos de administración y producción. Por lo que, si denominamos δ a todo este montante, la **prima total** o **prima de recibo** por la contratación del seguro se corresponde con la suma:

$$\text{Prima Total} = P'' + \delta \tag{2.13}$$

Esquemáticamente:





Principios para el cálculo de primas

Las primas resultantes cuando agregamos todas las componentes deben cumplir los principios de equidad, solidaridad y suficiencia de acuerdo con la naturaleza de los riesgos asumidos por el asegurador:

- El **principio de equidad** se refiere a que la prima o precio del seguro se ajuste al riesgo de siniestralidad de cada póliza, es decir, que el asegurado pague según el riesgo que incorpora. Desde el punto de vista actuarial, este criterio implica tener en cuenta, para la tarificación, los factores de riesgo que expliquen en mayor medida el comportamiento de la estructura de siniestralidad.

- El **principio de solidaridad** se refiere a la repartición del riesgo y de la prima total de manera que las pólizas que incorporan un menor riesgo paguen más de lo que estrictamente les correspondería y compensen una prima menor a pagar por las que incorporan más riesgo, de manera que en términos esperados se tienda a la media de siniestralidad. De hecho este criterio se deriva del anterior, puesto que no se puede pretender que cada individuo pague estrictamente por lo que ha "gastado".

- El **principio de suficiencia** se refiere a que en términos esperados las primas sean suficientes para cubrir todos los riesgos de la cartera considerada y que permitan hacer rentable, en condiciones de estabilidad a largo plazo, a la empresa aseguradora.

Notamos que el precio forma parte del servicio concebido como producto puesto a disposición del cliente y preparado para su introducción en el mercado. Los Actuarios, deben tener presente la técnica actuarial que hace referencia a los verdaderos costes de siniestralidad, gastos y márgenes que en todo momento han de dominar la estructura del precio, pero se comprende que en su política de precios existirá también una componente de adecuación al mercado y al poder adquisitivo de los segmentos de población a los que irá dirigido cada producto o cobertura. Así, el equilibrio entre las tendencias que

defienden la asequibilidad del precio, por una parte, y su equidad y suficiencia, por otra, se hace primordial a la hora de diseñar la política del precio [Coutts (1984b)].

Clasificación

Podemos distinguir las primas en función de si se pagan de una sola vez o de forma periódica, y si son constantes o no a lo largo de la vida del contrato. Las clasificaciones más usuales son:

a) *Prima única* y *prima periódica.*

Prima única es aquella mediante cuyo pago, el tomador se libera totalmente de satisfacer nuevas cantidades, por este concepto, durante la duración del seguro.

Primas periódicas son las que satisface periódicamente dentro de los plazos previstos para la duración del seguro. Normalmente es por anualidades.

b) *Prima fraccionada* y *prima fraccionaria.*

Ambos conceptos responden a un fraccionamiento de la prima, que se realiza para mayor comodidad en el pago. Así, pueden establecerse períodos de pago semestrales, trimestrales o mensuales. Las consecuencias del régimen establecido son, sin embargo, distintas en los dos supuestos.

La prima fraccionada es aquélla que, aunque calculada en períodos anuales, es liquidada mediante pagos más reducidos; por tanto, si la prima señalada lo ha sido en concepto de prima fraccionada, y el siniestro se produce, la entidad aseguradora puede exigir al tomador el abono de las restantes fracciones de prima no abonadas, o lo que es igual, descontárselas de la cantidad que en virtud del siniestro deba pagar.

La prima fraccionaria está calculada estrictamente para un período de tiempo inferior al año, durante el cual tiene vigencia el seguro. En el caso de la prima fraccionaria, si el siniestro se produce, el asegurador deberá satisfacer la indemnización pactada sin poder reclamar el abono de las restantes fracciones de la prima que faltarán por vencer hasta el final de la anualidad en curso.

c) *Prima fija* y *prima variable.*

Ésta es una diferenciación que responde a las dos distintas maneras de operar en la industria del

seguro que pueden tener las mutuas y las cooperativas de seguros.

Prima fija es la que corresponde a la cobertura de riesgos asegurados en entidades que adopten la forma jurídica de sociedad anónima, mutua o cooperativa a prima fija. En ellas, la prima se establece por adelantado para el período de cobertura pactado en el seguro, con independencia de la posible participación del asegurado en los resultados desfavorables de cada ejercicio, como sucede en el caso de mutuas y cooperativas.

Prima variable es la que se corresponde a la cobertura de riesgos asegurados en entidades que adopten la forma jurídica de mutua o cooperativa a prima variable, y se materializa mediante el pago de derramas con posterioridad a los siniestros. Previamente, y como fondo de maniobra, se exige a los asegurados la aportación de una cuota de entrada para hacer frente al pago de siniestros y gastos.

2.2.3. Sistemas de tarificación

El objetivo de todo sistema de tarificación es la obtención de las primas o precios del seguro. Las primas resultantes deben responder a los principios de libertad de competencia y deben estar fundadas en la equidad y suficiencia.

Los principios técnicos en que se basa la elaboración de una tarifa constituyen el sistema de tarificación.

Distinguimos, en el campo actuarial, dos sistemas de tarificación:

- **Tarificación *a priori* o *class-rating***

  Este sistema se denomina *a priori* puesto que nos permite asignar una prima a un riesgo que se incorpora a nuestra cartera sin tener necesariamente experiencia sobre la siniestralidad que conlleva. Únicamente es necesario conocer determinadas características para asignar una siniestralidad esperada y con ella una prima.

  En este sistema la agrupación de riesgos en clases homogéneas se hace teniendo en cuenta los llamados factores de riesgo, es decir, aquellas características exógenas significativas cuya presencia explica una parte importante de la siniestralidad. La necesidad de elaborar tarifas equitativas obliga

a considerar la presencia de estos factores de riesgo y el nivel con que lo hacen.

La elección de estos factores de riesgo que han de incorporarse a una tarifa es objeto de estudio por la *Estadística Actuarial*. Esta selección debe hacerse con el criterio estadístico de que la media de daños resultante en cada clase sea distinta y que la dispersión dentro de cada clase sea mínima.

- **Tarificación *a posteriori* o *experience-rating***

La tarificación *a posteriori*, en oposición a la tarificación *a priori*, parte de una prima inicial para cada unidad de riesgo, individuo o grupo, que se va modificando en períodos sucesivos de acuerdo con la experiencia individual o colectiva para dar lugar a las primas de los períodos sucesivos. En un sentido amplio, la expresión *experience-rating* se aplica a todo problema de actualización de tarifas mediante la incorporación de nueva información. Podríamos distinguir:

- La basada en la eficacia de la tarificación, que considera el riesgo individual: Bonus-Malus, Merit-rating, Retrospective-rating
- Y la basada en la eficacia y estabilidad, que considera el riesgo de un colectivo o grupo: Distribución de dividendos (Premium refund) y participación en beneficios

Centrados en el primer caso, la justificación de estos sistemas está en que dentro de cada clase de riesgo existe heterogeneidad, debida a la influencia de ciertos factores de riesgo no considerados (conocidos o desconocidos) o a la incorrecta agrupación de las clases de los si considerados, que pondrá de manifiesto la siniestralidad con el transcurso del tiempo. Al considerar la experiencia obtenemos un mayor grado de equidad en las primas de los ejercicios posteriores, que en la inicialmente cobrada. Una manera de conseguir este mayor grado de equidad, es incorporar la información evolutiva de los riesgos mediante un sistema de bonificaciones y penalizaciones (sistema *bonus-malus*) de acuerdo con los resultados obtenidos [Nieto y Vegas (1993); Lemaire (1995); Vegas (1992a,b,1993)].

Cabe notar, que en la tarificación *a posteriori*, también es interesante realizar un estudio de cuales son los factores de riesgo influyentes en la siniestralidad de la prima inicial. De esta forma, si los factores y la agrupación de sus niveles es la adecuada, la heterogeneidad que se intenta corregir con las bonificaciones y penalizaciones sería menor y éstas a su vez más leves [Dione and Vanesse (1989); Lemaire (1988)].

En el seguro del automóvil, el sistema *bonus-malus* se aplica generalmente sobre las coberturas de responsabilidad civil obligatoria y voluntaria, daños propios, incendio, robo y rotura de lunas, quedando excluidas las demás garantías del seguro. Su aplicación suele tomar como referencia el número de siniestros declarados en un período de 12 meses; si no se han declarado siniestros a las garantías computables se asciende por la escala de descuentos uno o más escalones. Por cada siniestro declarado a las garantías computables se desciende por la escala uno o más tramos, bien reduciéndose los descuentos, bien aplicando recargos. Históricamente, lo usual ha sido basar la escala en el número de siniestros, aunque existen estudios basados en las cuantías de los siniestros declarados [Morillo (2000)].

2.2.4. Tarificación *a priori* o *class-rating*

Partimos de la experiencia de una cartera para una determinada cobertura en un período fijado, en general un año, en la que se ha observado para cada póliza de la cartera la siniestralidad y una serie de características o factores potenciales del riesgo observado.

Se trata de realizar un proceso que pasará por las siguientes fases [de Wit (1986); van Eeghen, Greup y Nijssen (1983)]:

a) **Determinación de la estructura de tarifa**, resolviendo:

- **selección de las variables tarificadoras**: es la elección de los factores de riesgo o características que utilizaremos para distinguir a los asegurados con diferentes riesgos asociados, puesto que influirán en la siniestralidad. Los factores seleccionados pasarán a ser variables tarificadoras o variables de tarifa;
- **determinación de las clases de tarifa**: es la elección de las clases o agrupaciones de clases de las variables tarificadoras anteriormente seleccionadas, que acabarán discriminando a los diferentes grupos de riesgo en la tarifa final;
- **obtención de los grupos de tarifa**: es la obtención de grupos homogéneos de riesgo, exclusivos y exhaustivos, formados a partir de las clases de tarifa anteriores;
- **inclusión de los gastos en la tarifa**;

- **tratamiento adecuado de los grandes riesgos**.

b) **Cálculo de un nivel adecuado de prima para cada grupo de tarifa**: es la estimación de las primas de riesgo (equitativas y suficientes), que ajusten la siniestralidad para cada grupo de tarifa en términos esperados a partir de la esperanza del número de siniestros y del coste medio por siniestro, para así obtener la prima pura de la clasificación.

c) Y por último, la **implementación de la tarifa en un mercado competitivo**: es la adecuación de la tarifa a la práctica. A parte de la justificación teórica de la selección de los factores de riesgo es necesario tener presente la competencia de mercado y los segmentos de población a los cuales va dirigida la cobertura. Hay factores de riesgo que por su propia naturaleza supondrían una discriminación indeseada y el producto no sería aceptado, y sin embargo, hay otros que invitan a ser elegidos por la relación intuitiva que merecen con el riesgo y que serían fácilmente aceptados.

Dejando a un lado el aspecto de mercado y centrando la atención en la parte técnica actuarial, dado el objetivo de equidad y suficiencia en las primas, buscamos la formación de grupos de riesgo homogéneos determinados por combinaciones de clases de tarifa, que tendrán internamente una siniestralidad esperada similar y por lo tanto poca dispersión entorno a su valor esperado.

Formaremos pocos grupos si buscamos una tarifa resultante sencilla y aplicable, que diferencie sólo mínimamente los riesgos de calidades diferentes, o formaremos una agrupación más fina, es decir, con más grupos, si el objetivo es mayor ajuste en la prima individual y más detalle en la tarifa final.

Cabe notar, que los pasos de selección de variables tarificadoras, de determinación de las clases de tarifa y de obtención de los grupos de tarifa, dentro de la fase de determinación de la estructura de tarifa, están entre ellos estrechamente vinculados, y su resolución no es independiente en función de las metodologías utilizadas.

Es conveniente que la experiencia en que se base la tarifa pertenezca a un intervalo temporal lo más cercano posible al momento de actualización, y serán necesarias revisiones periódicas con datos actualizados que repetirán el proceso con todas sus fases, comenzando por la selección de variables de

tarifa. Además debemos verificar si las hipótesis iniciales realizadas sobre los datos han variado, en especial si la base de datos es pequeña [Jonson y Hey (1972); Lanteli (1962)].

Es importante realizar correctamente los pasos de la fase inicial de determinación de la estructura de tarifa para llegar a una correcta realización de la tarifa final en el cumplimiento del principio de equidad. Esta fase inicial forma parte del informe técnico actuarial en la justificación de las primas resultantes para la Dirección General de Seguros (DGS). La DGS puede requerir la presentación, siempre que lo entienda pertinente, de los modelos de pólizas, tarifas de primas y las bases técnicas[3] al objeto de controlar si respetan las disposiciones técnicas y sobre contrato de seguro.

2.2.4.1. Factores de riesgo

Las variables consideradas en un estudio de análisis estadístico multivariante pueden ser clasificadas de diferentes formas [Andenberg (1973) pp. 26-27]:

a) Según el objetivo y la interpretación del análisis:

- Variables respuesta o dependientes.
- Variables intermedias: son las que son tratadas como respuesta para algunas variables y como explicativas para otras.
- Variables explicativas o independientes.

b) Según la escala de medida de los posibles valores:

- Nominal: En este tipo, las diferentes categorías de la variable no tienen ningún tipo de ordenación.

[3] Las bases técnicas han de ser suscritas por un actuario de seguros, y deberán comprender:
- Información genérica: explicación del riesgo asegurable conforme a la póliza respectiva, los factores de riesgo considerados en la tarifa y los sistemas de tarificación utilizados
- Información estadística sobre el riesgo: se aportará información sobre la estadística que se haya utilizado, indicando el tamaño de la muestra, las fuentes y método de obtención de la misma y el período a que se refiera
- Información sobre la imputación en la prima pura del recargo de seguridad
- Información sobre la imputación de los recargos de gestión y beneficios
- Descripción de la equivalencia financiero-actuarial utilizada para el cálculo de la prima de tarifa o comercial
- Y si es el caso el detalle sobre el cálculo de las provisiones técnicas

- Ordinal: En este tipo, las diferentes categorías de la variable tienen implícita alguna ordenación natural.

- Intervalo: En este tipo, dada una diferencia entre dos valores, ésta tiene el mismo significado que la misma diferencia entre otros dos valores de la escala.

- Razón: Son variables intervalo que además tienen un cero natural.

Estas escalas están ordenadas jerárquicamente, desde la nominal hasta la razón. Cada tipo incorpora alguna cosa más al siguiente, por ejemplo, una variable intervalo es también ordinal pero no al revés, es decir, una ordinal en general no es intervalo. Así, dada una escala podemos reducir la información proporcionada por la variable a una escala de orden menor. En la práctica el cambio más utilizado es el cambio de intervalo a ordinal. Usualmente las variables de la escala nominal y ordinal son denominadas como categóricas o cualitativas. Y las de la escala intervalo o razón como cuantitativas.

c) Según el tamaño del rango de los posibles valores:

- Continuas: En este tipo, el conjunto de valores tiene un rango infinito no numerable. Típicamente estas variables suponen valores pertenecientes a intervalos reales o bien a una colección de ellos.

- Discretas: En este tipo, el conjunto de valores tiene un rango finito, a lo sumo infinito numerable. Estas variables pueden ser numéricas o no numéricas.

- Binarias o dicotómicas: Éstas son variables discretas que toman sólo dos valores.

d) Además, de cara a las aplicaciones informáticas, distinguiremos:

- Variable Frecuencia: Nos es útil para disminuir el volumen de los datos en el aspecto informático, en lugar de poner caso a caso, ésta nos sirve para especificar los contajes de la misma combinación.

- Variable Peso: Nos sirve para ponderar el peso que tiene una observación en la muestra. Si se utiliza conjuntamente con una variable frecuencia hay que ir con cuidado en la interpretación pues no será la ponderación de un solo individuo sino la del conjunto de individuos de la misma combinación. Aunque es usual que la variable frecuencia juegue el papel de peso

directamente.

Para la realización del proceso de tarificación *a priori*, dispondremos de la experiencia de una cartera de riesgos compuesta por un conjunto de pólizas, para un período determinado. De cada póliza tendremos observación de la siniestralidad y de una serie de características o factores potenciales del riesgo.

Hemos visto que la prima pura (2.7) es calculada como el producto de la esperanza del número de siniestros, E[N], por la esperanza de la cuantía de un siniestro, E[X]. Dada tal relación, estaremos interesados en estudiar el comportamiento aleatorio de ambas variables por separado. Así, para la realización del proceso de tarificación *a priori*, estaremos interesados, como primer paso, en la selección de las variables de tarifa que influyan en el número de siniestros, N, y en la selección de las que influyan en la cuantía de un siniestro, X, también por separado.

Respecto al objetivo e interpretación del análisis, se trata de clasificar las pólizas según el riesgo que incorporan, por lo que la variable aleatoria que recoja la siniestralidad (coste total, S, número de siniestros, N, o cuantía de un siniestro, X) jugará el papel de variable dependiente. La variable N es una variable discreta numérica que toma valores en los naturales. Dependiendo de la metodología estadística utilizada, a veces será tratada como cuantitativa, y a veces como categórica ordinal. Las variables X y S son variables continuas, y como tales deben ser tratadas.

Los *factores potenciales de riesgo*, $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_P$, son las variables independientes o explicativas, pues a través de algunos de ellos (variables de tarifa) seremos capaces de explicar la estructura de riesgo. A las variables explicativas también se les denomina variables predictoras o predictores pues a modo de variables independientes constituirán modelos que servirán para la predicción de la variable dependiente, a la que por el mismo motivo se la denomina variable respuesta.

Los factores de riesgo, son características "medibles" que habremos observado y que tendrán una posible relación de causa con la siniestralidad objeto de estudio. Es necesario que puedan tener alguna definición y que puedan ser representados como variables, además de ser tenidos en cuenta de modo expreso como datos codificados para posibles estudios.

Una fase previa a todo el proceso de tarificación, incluido el paso de selección de variables tarificadoras, es "la selección o recopilación de posibles factores potenciales de riesgo". Es imprescindible conocer y procesar la máxima información en torno al riesgo asegurado [Ingenbleek y

Lemaire (1988); Gogol (1993)]. La siniestralidad evoluciona en el tiempo, por lo que es posible que a partir de un momento sea explicada por factores no tenidos en cuenta anteriormente, o bien porque no explicaban suficientemente el riesgo en el momento de la tarificación, o bien porque nunca habían sido considerados como posibles factores.

En general el conjunto de factores de riesgo será de tipo mixto (mezcla de variables cuantitativas y cualitativas). Y será importante disponer para el estudio de los datos en sus escalas originales.

En ocasiones encontramos predictores continuos discretizados[4] de antemano (por ejemplo la edad del conductor en intervalos de edad). Esto implica una pérdida de información al pasar a una escala de medida menor. Resulta imposible obtener los datos originales, cuantitativos, de los ya codificados como discretizaciones, pues no sabemos qué valor tomó la variable dentro de cada grupo, sólo sabemos entre qué valores osciló. En tal caso, para la formación o agrupación de clases de tarifa, si no disponemos de los datos originales, no seremos capaces de deshacer la agrupación original para realizar otras que quizá proporcionarían mejores resultados. En el proceso de tarificación posiblemente acabaremos discretizando las variables continuas para obtener los grupos de tarifa finales, pero no debemos discretizar cuando aún no hemos confirmado su relación con la siniestralidad.

Notamos en mayúscula que el pequeño esfuerzo que representa la correcta gestión inicial de datos (caso de no ser estadísticas comunes realizadas por entidades de interés, por tener dificultades añadidas), nos llevará a una mejora significativa al largo y costoso proceso de tarificación que ha de servir, a largo plazo a la empresa aseguradora, a la obtención de mayores beneficios y mejor gestión de los riesgos de la cartera. Un mercado tan competitivo como es el español implica una gran amenaza para las compañías que todavía confían en métodos simples de tarificación y en el análisis superficial de datos estadísticos. Es imprescindible invertir de forma importante en conocer y analizar la propia experiencia.

Los factores de riesgo podrán hacer referencia tanto a características del objeto asegurado como a otros condicionamientos de éste: características del asegurado, del tomador, condiciones socio-económicas que lo rodean, etc [Booth, Chadburn, Cooper, Haberman y James (1999)]. En el seguro del automóvil, y dependiendo de la cobertura, los factores generalmente tenidos en cuenta son:

[4] Discretizar una variable continua: realizar una cantidad numerable de intervalos en los que clasificar los valores continuos, y sustituir el valor continuo de la variable original por la pertenencia al intervalo o por una puntuación ordinal.

- *Factores relativos al vehículo asegurado*: valor, antigüedad, categoría, clase, tipo, marca, modelo, número de plazas, potencia, peso, o relación potencia / peso, color, etc
- *Factores relativos al conductor*: edad, sexo, antigüedad del carnet, estado civil, profesión, número de hijos, posibilidad de conductores ocasionales, resultado de la experiencia en el pasado, etc
- *Factores relativos a la circulación*: zona de circulación[5], uso del vehículo[6], kilómetros anuales, etc

En ocasiones, puede interesar tener en cuenta información excesivamente privada o comprometedora para el asegurado (o tomador en su caso), que éste no esté dispuesto a responder. En tal caso, se debe intentar buscar una alternativa que resuma aproximadamente la misma información y que sí esté dispuesto a responder sin reticencias [Harrington y Doerpinghaus (1993)].

2.2.4.2. Datos de experiencia de siniestralidad

La información disponible puede estar agregada o desagregada. Veamos en cada caso el tipo de información a manejar:

- **Información desagregada**:

  Los datos tendrán el siguiente aspecto:

$\mathbf{Y}$	$\mathbf{F}_1$	$\mathbf{F}_2$	...	$\mathbf{F}_P$
y_1	f_{11}	f_{12}	...	f_{1P}
y_2	f_{21}	f_{22}	...	f_{2P}
...	...	...	...	...
y_n	f_{n1}	f_{n2}	...	f_{nP}

[5] En la zona geográfica las diferencias de siniestralidad son achacables a múltiples factores tales como la densidad de tráfico, el clima, la calidad de la red viaria, la edad media de los conductores...etc.
[6] En el uso que se da al vehículo, se tiene en cuenta que los profesionales que trabajan con su coche lo usan más y por lo tanto están más expuestos al riego.

donde:

- n es el número de pólizas si $\mathbf{Y}$ representa el número de siniestros por póliza o si representa la cuantía total de los siniestros por póliza;
- n es el número de siniestros acaecidos en la cartera si $\mathbf{Y}$ representa la cuantía por siniestro;
- $(\mathbf{F}_1,\mathbf{F}_2,\ldots,\mathbf{F}_P)$ es, en cualquier caso, la matriz de tipo mixto que recogerá las características o factores de riesgo asociados a las pólizas correspondientes

Por ejemplo, para el número de siniestros y la cuantía total:

Número de póliza	**Número de siniestros**	**Cuantía total**	**Período de exposición al riesgo**	**Sexo**	**Edad**	**Antigüedad del carnet**	**Potencia**	**...**	**Zona**
892356	0	0	1	M	25	6	95	...	A
892357	0	0	1	M	36	15	120	...	B
892358	2	660	1	H	60	40	85	...	A
892359	1	560	1	H	54	20	150	...	C
892360	3	6600	0.6	H	19	0.5	260	...	D
...	...	...	...	...	...	...	...	...	...
892389	0	0	1	M	38	17	120	...	B
892390	1	300	1	H	22	4	150	...	D

Y para las correspondientes cuantías de los siniestros:

Número de póliza	Cuantía por siniestro	Período de exposición al riesgo	Sexo	Edad	Antigüedad del carnet	Potencia	...	Zona
892358	200	1	H	60	40	85	...	A
892358	460	1	H	60	40	85	...	A
892359	560	0.6	H	54	20	150	...	C
892360	900	0.6	H	19	0.5	260	...	D
892360	2200	0.6	H	19	0.5	260	...	D
892360	3500	1	H	19	0.5	260	...	D
892390	300	1	H	22	4	150	...	D
...	...	...	...	...	...	...	...	...

Respecto al seguro del automóvil, cabe notar que, el ficheros generado de la base de datos en referencia al número de siniestros y a la cuantía total, es sustancialmente de mayor volumen que el de las cuantías individuales, ya que contiene mayor número de filas al contemplar todas las pólizas con o sin siniestro. Y el fichero con las cuantías tiene tantas filas como siniestros acaecidos. Si no ocurre así, es mala señal para la compañía!

- **Información agregada**:

En este caso los factores de riesgo son todos categóricos (o si eran continuos discretizados en clases), pues la información viene dada por una tabla cruzada de todos los factores. Supongamos que disponemos de dos factores, a modo de ilustración, A y B con 2 y 3 clases respectivamente (fácilmente extrapolable al caso de P factores con $n_1, n_2, \ldots, n_P$ clases respectivamente). El total de celdas de la tabla cruzada (en nuestro ejemplo 2x3 = 6 celdas) es el total de observaciones de qué dispondremos.

Así tendremos,

	B1	B2	B3
A1	$y_{11} \mid w_{11}$	$y_{12} \mid w_{12}$	$y_{13} \mid w_{13}$
A2	$y_{21} \mid w_{21}$	$y_{22} \mid w_{22}$	$y_{23} \mid w_{23}$

donde:

- si estamos analizando el número de siniestros por póliza: y_{ij} es el número medio de siniestros en la celda ij, y w_{ij} es el número de pólizas que pertenecen a la combinación ij, es decir, el número de pólizas con las que hemos calculado el número medio correspondiente;

- si estamos analizando la cuantía por siniestro: y_{ij} es la cuantía media por siniestro en la celda ij, y w_{ij} es el número de siniestros de la combinación ij;

- si estamos analizando las cuantías totales por póliza: y_{ij} es la cuantía total media de los siniestros pertenecientes a la celda ij, y w_{ij} es el número de pólizas que pertenecen a la combinación ij.

En todos los casos es usual tener celdas vacías, pues hay combinaciones de la tabla cruzada en las que no hay pólizas con tales características, incluso combinaciones imposibles (como edad entre 18 y 20 años combinada con antigüedad del carnet de más de 10 años). En el caso de las cuantías por siniestro es aún más usual, pues tendremos muchas combinaciones en las que sí hay pólizas, pero que no han sufrido siniestro en el período de observación.

Notamos que aunque agrupar datos origina pérdida de información, a menudo conlleva un sorprendente incremento en el ajuste de regresiones. El estudio clásico de Cramer (1964) sobre este tema constituye un buen ejemplo. Además supone una reducción espectacular del tamaño de los datos a manejar en carteras de gran volumen.

A modo de ejemplo visual, los datos de Baxter, tabla 2.2, descritos en el apartado 2.3.3, cruzan 3 factores y se refieren a cuantías medias por siniestro. Si los colocamos en un fichero tendrán el siguiente aspecto:

Coste medio por siniestro	Número de siniestros	Edad del conductor	Grupo de vehículo	Antigüedad del vehículo
289	8	17-20	A	0-3
372	10	17-20	B	0-3
189	9	17-20	C	0-3
763	3	17-20	D	0-3
282	8	17-20	A	4-7
249	28	17-20	B	4-7
288	13	17-20	C	4-7
850	2	17-20	D	4-7
133	4	17-20	A	8-9
288	1	17-20	B	8-9
179	1	17-20	C	8-9
160	1	17-20	A	10 o más
11	1	17-20	B	10 o más
...	...	...	...	...
123	6	60 o más	D	10 o más

En la práctica, tanto si la información se codifica agregada como desagregada, dispondremos de pólizas que no han estado vivas durante todo el período de observación (bien porque se han incorporado a mitad del período, o bien porque han vencido a mitad del período y no han renovado), este hecho debemos tenerlo en cuenta. Una posibilidad es extrapolar el resultado de siniestralidad a todo el período de observación, y otra es ponderar la siniestralidad según el tanto por uno de período en que la póliza ha estado viva, teniendo las vivas un peso de uno. En el caso de datos agregados éste hecho debe reflejarse en la variable que podemos denominar *expuestos al riesgo*, que vendrá representada por el número de pólizas para el número medio de siniestros y el coste total medio, y por el número de siniestros para el coste medio por siniestro.

También es posible que nos encontremos con la dificultad de datos faltantes cuando los siniestros sean siniestros en curso o pendientes de reclamación, pues el período de observación debe ser el año más reciente que representará la estructura más actual de la cartera, y hay que tener en cuenta que ese año es el que tiene los siniestros más inmaduros, por lo que es imprescindible realizar previamente una revisión de reservas de siniestros pendientes con el fin de obtener los niveles últimos de siniestralidad.

Por ejemplo, en el seguro del automóvil, respecto a daños personales, si estamos estudiando la cuantía por siniestro, puesto que el período de maduración de los siniestros es muy largo, será imprescindible realizar un análisis de reservas en curso que nos ofrezca el coste último de los siniestros.

Respecto a los factores potenciales de riesgo, también es posible que de pólizas antiguas no tengamos nota de ciertos factores considerados recientemente.

2.2.5. Notas referentes al seguro del automóvil

A la hora de recoger los datos de siniestralidad en el SOA debemos separar la información en lo que se refiere a daños personales y a daños materiales, ya que tienen frecuencia y coste medio muy diferentes. Adicionalmente debemos tener en cuenta el resto de coberturas hasta llegar al todo riesgo, en el que se tratará a parte la modalidad de daños propios por el mismo motivo. En este sentido, las *primas puras totales por póliza* se calcularán como la suma aritmética de las primas puras correspondientes a cada cobertura [Haberman y Renshaw (1996)]. Si denotamos por λ^c al número esperado de siniestros respecto a la cobertura c, y por m^c a la cuantía esperada de un siniestro respecto a la cobertura c (siendo c: daños propios, daños materiales, daños personales, etc), la prima pura total la calcularemos como:

$$PT = \sum_c \lambda^c \times m^c \tag{2.14}$$

Por supuesto, para poder realizar un estudio completo la información de partida debería ser desagregada.

Veamos el panorama actual por lo que respecta a información sectorial en ayuda de una mejor tarificación y gestión de riesgos (para mayor detalle nos referimos al anexo 2.1 del capítulo):

Ficheros sectoriales

En los últimos años, desde la Comisión Técnica de Seguros de Automóviles de UNESPA, se han creado varios ficheros sectoriales. A continuación los presentamos y extraemos las principales conclusiones en lo que respecta a la tarificación *a priori*.

- Base SIETE:

El Sistema Informativo de Especificaciones Técnicas, Base SIETE, es una base de datos elaborada por CENTRO ZARAGOZA en la que se incluyen las principales características técnicas de todos los vehículos automóviles susceptibles de ser asegurados en España.

- El Fichero de Vehículos Sustraídos e Indemnizados:

Está destinado a la localización de vehículos sustraídos mediante el intercambio de información a través de un fichero accesible a través de Internet. La información contenida se comparte con la Dirección General de la Policía de forma que cuando ésta recupera cualquier vehículo, las entidades están puntualmente informadas del lugar de localización y pueden proceder a su recogida.

- Estadística de automóviles:

TIREA pone a disposición del sector asegurador el servicio Estadística del Seguro del Automóvil (ESA), para la consulta de forma *on-line* de la información resultante de la explotación de los datos aportados por las entidades participantes. El servicio responde a la necesidad de obtener información sobre el riesgo elemental, que permita el conocimiento completo del mercado en el que se opera. Pretende recuperar la Estadística del Seguro del Automóvil de UNESPA para así coordinar los esfuerzos del Sector.

Encontramos el siguiente resumen de su funcionamiento en la Memoria de Actividades del año 2001 de UNESPA respecto al seguro de automóviles[7]:

"La Comisión Técnica de Seguros de Automóviles ha puesto en funcionamiento la nueva Estadística de Automóviles, que tiene como objetivo fundamental llenar el vacío de información técnica producido en el sector desde la última elaboración de la Estadística Común de Automóviles en el año 1997. Se trata de proveer a las entidades de la información básica que les permita el conocimiento completo del mercado en el que operan. Se pretende aumentar el número de entidades participantes y de esta forma enriquecer la información resultante.

[7] http://www.unespa.es/memorias/memoria2001/35-39_Automoviles.pdf

Aunque se basa en una estadística anterior, incorpora importantes novedades que la actualizan y ha contado con la información de 23 entidades, que representan el 57.52% del total de pólizas del mercado del seguro del automóvil y el 53.94% del total de primas del sector de autos (según datos de ICEA correspondientes al año 2000).

Sólo las entidades que colaboren de forma activa tendrán acceso a los resultados agregados de las explotaciones estadísticas, lo que implica un mayor grado de compromiso por parte de las entidades y una mayor validez en los resultados. El resto de las entidades del sector tan sólo recibirán un documento-resumen con información breve y genérica.

Esta estadística proporcionará nuevas posibilidades de explotación gracias a la incorporación de la información del fichero Base 7 (gestionado por Centro Zaragoza).

La primera edición se va a elaborar con los datos de 1999 y 2000, lo que permitirá verificar la evolución del sector de seguros de esos años.

Algunas variables se han modificado para adaptarse a las nuevas costumbres del sector y ser un instrumento de mejor uso. Por ejemplo, la situación del riesgo se mide por la provincia de la póliza y no por el lugar de ocurrencia del siniestro. De la misma manera, se han actualizado las categorías y usos de vehículos.

No menos importante es la naturaleza dinámica de la estadística, ya que se irá enriqueciendo año tras año con la experiencia y las aportaciones de todas las entidades participantes y por lo tanto se constituye como un instrumento activo y vivo, al servicio de la toma de decisiones de las entidades."

- El Fichero Histórico de Seguros de Automóviles:

Se trata de un fichero de datos de carácter personal constituido por las compañías de seguros del automóvil con el fin de permitir la tarificación, la selección de riesgos y la elaboración de estudios de técnica aseguradora.

Empezó a funcionar en noviembre del 2000 y se denomina fichero histórico de SINiestralidad de COnductores (SINCO). Proporciona información objetiva sobre la siniestralidad del tomador del seguro referente a los últimos 5 años, por lo que permite ajustar la prima que realmente le corresponde.

El contenido del mismo está elaborado de acuerdo con la Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal, y únicamente tienen acceso a los datos las entidades adheridas, las cuales lo pueden utilizar para realizar consultas en el momento de la solicitud de nuevas pólizas. De este modo, pueden realizar una valoración técnica y objetiva del riesgo para aplicar correctamente las tarifas de prima que tengan recogidas en sus bases técnicas.

En la citada Memoria de Actividades del año 2001 encontramos:

"*Transcurrido ya más de un año desde que comenzó a funcionar el Fichero Histórico de Seguros de Automóviles, el primer balance puede resumirse en los siguientes puntos:*

- *Han firmado su adhesión al fichero 44 entidades que suponen una cuota de mercado del 85%.*
- *De las anteriores, vienen desarrollando normalmente las cargas de datos 38 entidades, con lo que la cuota de mercado que está en disposición de ser consultada es de un 76%.*
- *Una de las dificultades con que se han encontrado la práctica totalidad de las entidades ha sido la adaptación de sus procedimientos internos a fin de estar en disposición de efectuar consultas al fichero. El ritmo de en la resolución de este problema ha sido muy desigual y así, a finales del 2001, 32 entidades formulaban consultas y de ellas sólo 18 lo hacen de manera habitual.*
- *Es de resaltar el esfuerzo por conseguir que la calidad de los datos cargados y sus sucesivas actualizaciones alcance el nivel necesario para que el fichero rinda su enorme potencial de utilidad al sector. En este sentido, hay que destacar los importantísimos avances conseguidos en los últimos tiempos, aunque ésta es una cuestión capital sobre la que no cabe relajar en absoluto la atención ni el nivel de exigencia...."*

De lo que deducimos que a medida que se vayan adhiriendo nuevas entidades al fichero histórico, los datos serán más completos y a medida que pase el tiempo y las compañías vayan renovando las pólizas, los costes podrán ser ajustados mucho mejor. Cabe notar que dadas unas características de seguro, un tomador con un riesgo determinado podrá encontrar diferentes tarifas en el mercado porque los precios de partida de cada entidad dependerán de su política, pero el trato como buen o mal conductor será el mismo. No será de sorprender que un asegurado en una compañía con antigüedad de la póliza no superior a 5 años, se encuentre con una prima bonificada o penalizada en la renovación sin

haber tenido siniestro en el último período. Si va a otra compañía obtendrá el mismo resultado.[8]

Hasta el momento, podían existir casos de asegurados que estaban expuestos a riesgos superiores a la prima que pagaban. En consecuencia, había personas que pagaban primas superiores a la valía de su riesgo. El fichero, en el medio plazo, tenderá a equilibrar este efecto, ya que cada uno pagará lo que en realidad le corresponda.

Con una correcta tarificación de riesgos, ayudada de todos estos ficheros y plataformas de comunicaciones, se logrará un claro control y disminución de costes y gastos de las entidades, objetivo primordial en un mercado de gran competitividad como es el del automóvil.

Para finalizar, en Vegas (1992a) pp. 32 encontramos el siguiente comentario respecto a los sistemas de tarificación *a posteriori*:

"*...Se demuestra matemáticamente que el sistema Bonus-Malus que acabamos de presentar está equilibrado financieramente, al verificarse que, transcurrido un año, la media de las frecuencias individuales de los siniestros coincide con la media ... Es decir, si la cartera no se modificara, si todos los asegurados con "malus" satisfacieran su sobreprima correspondiente y no se fueran a otra Entidad, el sistema estaría en equilibrio financiero. El problema es el de* ***la fuga de malus***."

En referencia a este problema, el fichero histórico supone también un gran avance.

Convenio CIDE

El Convenio entre Entidades Aseguradoras de Automóviles para la Indemnización Directa de Daños Materiales a Vehículos (CIDE) se implantó en España en enero de 1988. El objeto de este convenio, establecido para las entidades aseguradoras adheridas al mismo, es acelerar la liquidación y pago a sus respectivos asegurados de los daños causados exclusivamente a los vehículos, en aquellos accidentes de circulación que se produzcan por colisión directa entre dos de ellos, cualquiera que sea la clase y uso de los mismos, de acuerdo a los principios de responsabilidad que se determinen en el convenio.

[8] "... Si està pensant en comprar un segon cotxe per vostè o la seva parella, no es faci enrera pensant que l'assegurança del vehicle li suposarà un cost adicional molt elevat, perquè ara, si té el seu cotxe assegurat a MAPFRE, li aplicarem les mateixes condicions en la nova contractació. **Però si vostè encara no ens ha consultat el preu de l'assegurança del seu cotxe, ara és el moment perquè li respectem les mateixes bonificacions que a la seva companyia actual**, i gaudirà d'altres importants avantatges per ser client de MAPFRE, com: poder asegurar ciclomotors del dels 14 anys a fills de clients, asegurar motocicletes, tractors, remolcs, quads i carretilles..." (AGENCIA MAPFRE, Plaça Pau Casals 10 baixos, 08770 Sant Sadurní d'Anoia).

Es indispensable que los dos vehículos estén amparados por el seguro de responsabilidad civil de suscripción obligatoria. La aplicación del convenio sólo será posible cuando exista la declaración amistosa debidamente cumplimentada y firmada por los dos conductores.

Liquidación de siniestros vía CIDE: El hecho de que sea la aseguradora del perjudicado (acreedora) la que haga efectivo el coste de la reparación, en aras a la mayor celeridad, exige, lógicamente, que sea reembolsada por la aseguradora del causante (deudora). Sin embargo, en lugar de tomar como base del recobro el importe real de la reparación, se establece como fórmula de compensación un módulo determinado por el coste medio de los siniestros amparados por el convenio.

A grandes rasgos, cuando una entidad sea acreedora, anticipará la reparación en la cuantía peritada y recibirá a cambio el coste medio establecido por el convenio para ese período, y cuando sea deudora pagará el coste medio independientemente de la cuantía real del siniestro, la cual no llegará a conocer.

El sistema CICOS es el entorno informático por el que se tramitan los convenios de indemnización de daños materiales entre todas las entidades de automóviles.

El Centro Compensador residente en TIREA, realiza un cálculo mensual de los saldos que cada entidad debe o le deben con respecto a las restantes entidades y liquida los saldos mediante recibos y transferencias bancarias entre la cuenta del Centro de Compensación y las cuentas de cada entidad.

Respecto a la tarificación a priori:

En RC de daños materiales debemos entender como coste del siniestro el pago total último que ha de realizar la entidad. Este punto es remarcable porque un porcentaje elevado de los costes, aproximadamente el 80%, es de cuantía la que se establece en el coste medio sectorial del convenio (que encontramos detallado en la tabla 2.1 del anexo 2.I), que al fin y al cabo es el coste real para la entidad. Por ejemplo en los datos de las carteras que pasaremos a detallar en inmediato, la mayoría de siniestros son de 75 000 pesetas para la cartera C1 y de 90 000 pesetas para la cartera C2.

Debido a que no sabremos el coste real de los siniestros en los que la entidad es deudora, la selección de variables de tarifa respecto al coste de un siniestro no tiene sentido. Adicionalmente, a la entidad no le es de ningún interés saber cuales son las variables que explican el riesgo, ya que para ella, independientemente de las características de la póliza, el coste siempre será el que establezca el convenio.

Al igual que en el proceso de riesgo, se debe suponer una cartera infinita en la que los costes por siniestro tiendan a la media sectorial. Una cuestión importante para la estabilidad económica de una entidad es que la media de los siniestros en los que sea acreedora vía CIDE, también se corresponda con la media sectorial. Pero este hecho es no predecible ya que dependerá de las características de los causantes y no de sus propias pólizas.

De todo ello, deducimos que en el caso de daños materiales, respecto de una compañía, sólo será de interés el estudio de los factores de riesgo que influyen en el número de siniestros por póliza y no en la cuantía por siniestro.

Para finalizar este apartado, en la Memoria de Completa de Actividades del año 2001 de UNESPA[9], página 4, encontramos el siguiente párrafo textual en referencia conjunta al fichero SINCO y al convenio CIDE, el cual resume la buena marcha del sector:

"*En el importante capítulo del seguro del automóvil hemos vivido un ejercicio durante el que parecen haberse moderado las tensiones en los costes. La política de las entidades, mediante la adecuada selección y tarificación de riesgos, ha sido la razón fundamental para recuperar niveles razonables de equilibrio. Paralelamente, hemos continuado desarrollando proyectos de interés general altamente beneficiosos para el conjunto del ramo. En este sentido hay que citar la mejora del sistema CICOS y la creación de un nuevo procedimiento de gestión para los siniestros sin daños personales. También se ha producido, durante el pasado año 2001, la auténtica puesta en marcha del Fichero Histórico del Seguro de Automóviles (SINCO)...*"

2.3. Descripción de los datos relativos a las aplicaciones

Ahora pasamos a detallar los datos que serán utilizados durante el trabajo para la ilustración de los métodos estadísticos de selección de variables de tarifa.

Las dos primeras carteras, C1 y C2, nos han sido cedidas por dos compañías diferentes, y aunque su generosidad nos ha permitido realizar aplicaciones con un volumen significativo y con datos reales, queremos resaltar que no ha supuesto un estudio privado de su cartera y por tanto no hemos tenido la posibilidad de investigar el resto de información por ellos recopilada referente a tales siniestros. En

[9] http://www.unespa.es/memorias/memoria2001/Memoria_2001_completa.pdf

ambos casos, los datos nos han sido cedidos en ficheros de SPSS, por lo tanto han sido datos inicialmente codificados y depurados. En ningún caso hemos podido investigar sobre los datos de años posteriores referentes a los siniestros que en nuestros ficheros estaban pendientes. Tampoco disponemos de información particular del nombre del titular de la póliza, sólo disponemos de un registro en el que aparece el número de póliza.

2.3.1. Datos de la cartera C1 de responsabilidad civil de automóviles

Los datos de esta cartera hacen referencia al seguro de Responsabilidad Civil (RC) de automóviles, tanto a RC materiales como a RC personales en España. Supone un total de 169 618 pólizas. El período de observación de un año es el del 1 de enero de 1996 al 1 de enero de 1997. Son datos desagregados.

2.3.1.1. Experiencia de siniestralidad

Disponemos sólo de un fichero, por lo que las cuantías son totales, no se detalla siniestro a siniestro. Disponemos de información sobre el número de siniestros y sobre la cuantía total para RC total (materiales + personales), para RC de daños materiales y para RC de daños personales, por separado.

RC de daños materiales:

Respecto a los costes totales, se han contabilizado cuantías totales negativas. La compañía ha contabilizado como siniestros los acometidos por otros conductores y que han provocado un parte de otra compañía vía CIDE. Ha contabilizado cuantías positivas y negativas con los excedentes y pagos que ha tenido que realizar por sobra o falta en las reparaciones siendo acreedora. El hecho de anotar los siniestros que le han sido realizados a sus conductores por esta vía, está bien desde el punto de vista contable, pero no a la hora de utilizar la información de siniestralidad para la selección de variables de tarifa. Suponemos que la información cedida, representa un fichero parcial de su trabajo.

Imaginemos que en esta situación estudiamos los factores de riesgo que influyen en el número de siniestros por póliza para RC materiales, en concreto si la edad del conductor principal es significativa: dentro del número de siniestro habrán siniestros que se correspondan al conductor habitual de la póliza

y habrán otros que pertenecerán a conductores de otras edades que ni sabremos, por lo que no tendrá ningún sentido utilizar la información mezclada. Si a caso deberíamos tener por separado los siniestros de nuestro asegurado y los de los otros conductores, de los cuales sí conocemos las cuantías, con los correspondientes factores de riesgo.

Así, para estos datos, tampoco será correcto analizar el número de siniestros en lo que respecta a daños materiales. Por el mismo motivo, no tendrá sentido analizar la RC total, al incluir ésta RC de daños materiales. En conclusión, descartamos cualquier selección utilizando la información relativa a tanto a RC materialcs, como a RC total.

RC de daños personales:

El número de siniestros no es de gran interés utilizado como experiencia de siniestralidad. Del total de pólizas disponibles, 169 618 pólizas: 1 699 han tenido un siniestro y sólo 21 han tenido dos.

El interés de estudio en personales son las cuantías. En estos datos sólo disponemos de la cuantía total, pero como ya hemos visto, la mayoría de siniestros están anotados uno a uno por lo que no tendremos problema en utilizar la información individual. En las pocas pólizas con dos siniestros podemos proceder a realizar la media y a asignar dos siniestros de cuantía media. De hecho, cuando estudiamos los factores influyentes en las cuantías individuales, de algún modo el número de siniestros afecta, ya que las pólizas con dos siniestros aparecerán dos veces, y sus características, que serán las mismas, también.

No disponemos de la variable "expuesto" que nos indique el porcentaje de año en que la póliza ha estado viva, pero disponemos de la fecha de inicio de la póliza y de la fecha de anulación, lo cual nos permite generarla fácilmente.

2.3.1.2. Factores de riesgo

Los factores potenciales de riesgo están muy bien codificados, es decir, tenemos que las variables de naturaleza continua como la edad del conductor están anotadas como tales a fecha 1 de enero de 1997. Estos son:

- Factores cuantitativos continuos:

- Edad del conductor habitual en años
- Antigüedad del carnet del conductor habitual en años
- Valor monetario en miles de pesetas del vehículo a nuevo

➢ Factores cuantitativos discretos:

- Antigüedad del vehículo asegurado en años (está codificada con valores discretos, en intervalos de 1, aunque es de naturaleza continua)
- Potencia del vehículo en caballos de potencia
- Porcentaje de *bonus* referente a RC total (con valores de 0 a 0.7 con intervalos de 0.05: 0, 0.05, 0.1, 0.15, 0.2,..., 0.65, 0.7)
- Porcentaje de *malus* referente a RC total (con valores en positivo y la misma escala que el *bonus* aunque con sentido contrario)

➢ Factores cualitativos:

- Tipo de producto (3 categorías: estándar, senior, dual)
- Modalidad de producto (3 categorías: terceros, combinado sin daños, todo riesgo)
- Sexo del primer conductor (2 categorías: hombre, mujer)
- Zona de tarificación (10 zonas)
- Provincia de circulación habitual (61 categorías codificadas con los códigos postales)
- Municipio (con los códigos postales más detallados)
- Grupo de vehículo (5 categorías: ficticia, 1ª categoría[10], 2ª categoría[11] camiones, 2ª categoría resto, 3ª categoría[12])

[10] Vehículos de 4 y más ruedas (turismos y vehículos comerciales de menos de 3 500 Kg)
[11] Camiones y vehículos industriales de más de 3 500 Kg
[12] Motocicletas y vehículos de 2 y 3 ruedas para cuya conducción sea necesario un permiso o licencia

- Tipo de vehículo (5 categorías: todo terreno, monovolumen, balilla duro, balilla blando, resto de vehículos)
- Uso del vehículo (6 categorías: particular, empresa, otros usos, transportes a terceros, transportes propios, todos los usos)
- Clase de vehículo (9 categorías: turismo, furgoneta, camión, vehículo industrial, autocar, tractor, motocicleta, triciclo, remolque)
- Respecto al vehículo de los cuales no somos capaces de decodificar tenemos la marca, el modelo y el submodelo.
- Tipo de combustible (2 categorías: diesel, gasolina)
- Forma de pago de la prima (3 categorías: anual, semestral, trimestral)

Estos datos son analizados en la aplicación 2 del capítulo 5.

2.3.2. Datos de la cartera C2 de responsabilidad civil de automóviles

Los datos de esta cartera hacen referencia al seguro de RC de automóviles, en concreto a RC materiales sólo para turismos en España. Han sido extraídos del fichero global mediante un muestreo aleatorio del 10 %, suponiendo éste 43 560 pólizas. El período de observación de un año es el del 30 de junio de 1998 al 30 de junio de 1999. Son datos desagregados.

2.3.2.1. Experiencia de siniestralidad

Disponemos de dos ficheros, uno que contiene el número de siniestros y otro con las cuantías una a una correspondientes a los siniestros de esas pólizas.

Casi todas las cuantías son de 90 000 pesetas, lo ideal sería saber el coste real del siniestro algo imposible con el convenio CIDE, por lo que descartamos las cuantías como experiencia de siniestralidad en la selección de factores. Exactamente, de las 3 496 cuantías individuales de que disponemos, 2 768 son de 90 000 pesetas, es decir, el 79.18%.

Disponemos de una variable "expuesto" que nos indica en tanto por ciento el porcentaje de año en que la póliza ha estado viva hasta el 30 de junio de 1999. Si nos fijamos sólo en las pólizas expuestas al riesgo durante todo el período, éstas son 31 551; así, del total de la muestra (43 560), el 72.4% ha durado todo el año.

2.3.2.2. Factores de riesgo

Los factores potenciales de riesgo son todos discretos:

- Factores cuantitativos:
  - Antigüedad de la póliza en años (8 intervalos: [0,1), [1,2), [2,3), [3,4), [4,5), [5,6), [6,7), [7,...))
  - Antigüedad del carnet de conducir del primer conductor en años (11 intervalos: [0,1), [1,2), [2,3), [3,4), [4,5), [5,6), [6,7), [7,8), [8,9), [9,10), [10,...))
  - Antigüedad del vehículo asegurado en años (11 intervalos: [0,1), [1,2), [2,3), [3,4), [4,5), [5,6), [6,7), [7,8), [8,9), [9,10), [10,...) + 1 clase de missings)
  - Valor monetario en millones de pesetas del vehículo a nuevo (15 intervalos: (0,1], (1,1.2], (1.2,1.5], (1.5,1.7], (1.7,2], (2,2.2], (2.2,2.5], (2.5,2.7], (2.7,3], (3,3.5], (3.5,4], (4,5], (5,6], (6,10], (10,...] + 1 clase de missings)
  - Potencia del vehículo en caballos de potencia (9 intervalos: hasta 28, 29-33, 34-42, 43-53, 54-75, 76-94, 95-118, 119-215, más de 216)
  - Número de plazas (8 valores: 2, 3, 4, 5, 6, 7, 8, 9)
  - Edad del primer conductor en años (14 intervalos: [18,20), [20,25), [25,30), [30,35), [35,40), [40,45), [45,50), [50,55), [55,60), [60,65), [65,70), [70,75), [75,80), [80,...))
  - Edad del segundo conductor en años (14 intervalos: [18,20), [20,25), [25,30), [30,35), [35,40), [40,45), [45,50), [50,55), [55,60), [60,65), [65,70), [70,75), [75,80), [80,...) + 1 clase de missings)

- Edad de máximo riesgo en años, calculada como el mínimo entre la edad del primer conductor y del segundo, y si hay missing en la edad del segundo conductor se elige la del primer conductor (14 intervalos: [18,20), [20,25), [25,30), [30,35), [35,40), [40,45), [45,50), [50,55), [55,60), [60,65), [65,70), [70,75), [75,80), [80,...))
- Nivel de *bonus* (12 valores de escala: -50, -40, -20, -10, 0, 10, 20, 30, 35, 40, 45, 50)

Algunos de estos factores se hubieran podido codificar como datos continuos sin ninguna dificultad, por ejemplo la antigüedad de la póliza a partir de la fecha de inicio, las edades a partir de las fechas de nacimiento y la antigüedad del carnet a partir de la fecha de expedición.

➢ Factores cualitativos:

- Sexo del primer conductor (2 categorías: hombre, mujer)
- Forma de pago de la prima (3 categorías: anual, semestral, trimestral)
- Modalidad (6 categorías: terceros, terceros más rotura de lunas, terceros más rotura de lunas más incendios, todo riesgo, todo riesgo con franquicia de algún tipo, otra)
- Tipo de combustible (3 categorías: diesel, eléctrico, gasolina)
- Microzona de circulación habitual (52 categorías)
- Zona de circulación habitual (30 categorías)
- Provincia de circulación habitual (61 categorías codificadas con los códigos postales)

Estos datos son analizados en la aplicación 4 del capítulo 5.

2.3.3. Datos de Baxter

Estos datos [Baxter, Coutts y Ross (1980), tabla 1, pp. 21], que encontramos reproducidos el anexo 2.2 del capítulo (tabla 2.2), pertenecen a una cartera de seguros privados relativa al seguro del automóvil. Se dispone de información agregada, y los datos consisten en 128 cuantías medias de siniestros ocurridos durante el año 1975 referentes a daños propios junto con el correspondiente número de siniestros. Éstos han sido agregados de acuerdo a tres factores de riesgo:

- Antigüedad de la póliza: con 8 clases (*Policy holder Age*, PA)
- Antigüedad del vehículo: con 4 clases (*Vehicle Age*, VA)
- Grupo de vehículo: con 4 clases (*Car Group*, CG)

Por ello se dispone de 8x4x4 = 128 observaciones. Si analizamos un poco los datos, observamos lo siguiente:

Inicialmente se partía de:

- 8 902 siniestros con sus correspondientes cuantías individuales y no sabemos el número total de pólizas
- dos factores de riesgo continuos y uno categórico nominal

Finalmente se dispone de:

- 8x4x4 = 128 celdas que conforman la tabla cruzada de los tres factores, menos 5 celdas vacías, 128 – 5 = 123 cuantías medias con el correspondiente número de siniestros
- dos factores categóricos ordinales (discretizados) y uno categórico nominal

Aunque con la tabla cruzada se ha pretendido reducir el volumen de información, no se ha contemplado el número de pólizas implicadas en cada perfil, por lo que el estudio sólo puede realizarse sobre la cuantía por siniestro (con datos agregados y ponderados), y no sobre el número de siniestros o la cuantía total por póliza.

Estos datos son analizados en la aplicación 3 del capítulo 5.

2.3.4. Datos de impagos de préstamos de una entidad financiera

Estos datos [Bermúdez y Pons (1997)], que encontramos el anexo 2.3 del capítulo (tabla 2.3), son datos relativos a las pérdidas monetarias ocasionadas a una entidad financiera por aquellos clientes que no pudieron hacer frente al pago de la deuda del préstamo contraído, en un determinado momento. Los datos consisten en 401 cuantías, que encontramos clasificadas de acuerdo a dos factores de riesgo, uno de naturaleza categórica nominal y otro continuo pero discretizado de antemano:

- E: estado civil con tres categorías
  - ⇒ E1: aparejado
  - ⇒ E2: divorciado o separado
  - ⇒ E3: soltero

- A: antigüedad en el puesto laboral con tres clases:
  - ⇒ A1: menos de 2 años (antigüedad<2)
  - ⇒ A2: entre 2 y 10 años ($2 \leq$ antigüedad<10)
  - ⇒ A3: más de 10 años (antigüedad ≥ 10)

Los datos, aunque no son agregados están presentados en una tabla de dos factores puesto que ambos son discretos. Se anexa también la forma que tendría la información si se hubiera tratado de datos agregados (tabla 2.4).

Al igual que en el apartado anterior, aunque aquí disponemos de las cuantías por siniestro una a una y de ellas podemos realizar un estudio completo respecto a la influencia de los factores, con el número de siniestros y con la cuantía total por póliza no podemos hacer nada, ya que no sabemos el número total de clientes con el mismo riesgo.

Estos datos son analizados en la aplicación 1 del capítulo 5.

ANEXO 2.1. Seguro del Automóvil

1. Los convenios CIDE y ASCIDE, y el sistema CICOS

La información que a continuación detallamos ha sido extraída de http://www.unespa.es/.

1.1. Convenios CIDE y ASCIDE

El Convenio entre Entidades Aseguradoras de Automóviles para la Indemnización Directa de Daños Materiales a Vehículos (**CIDE**) se implantó en España en enero de 1988. El objeto de este convenio, establecido para las entidades aseguradoras adheridas al mismo, es acelerar la liquidación y pago a sus respectivos asegurados de los daños causados exclusivamente a los vehículos, en aquellos accidentes de circulación que se produzcan por colisión directa entre dos de ellos, cualquiera que sea la clase y uso de los mismos, de acuerdo a los principios de responsabilidad que se determinen en el convenio.

Por tanto, quedan excluidos de la aplicación del CIDE:

- Los daños a los vehículos cuando no exista colisión directa.
- Los daños cuando en el accidente intervengan más de dos vehículos.
- Cualesquiera otros daños materiales ajenos a los propios de los vehículos o perjuicios originados en el accidente.
- Los daños corporales.

Cuando se den estos supuestos, la tramitación del siniestro deberá efectuarse por el sistema tradicional. Sin embargo, el hecho de que existan lesionados no impide que pueda resolverse de acuerdo con el convenio la parte de daños a los vehículos.

Es indispensable que los dos vehículos estén amparados por el seguro de responsabilidad civil de suscripción obligatoria. La aplicación del convenio sólo será posible cuando exista la **declaración amistosa** debidamente cumplimentada y firmada por los dos conductores.

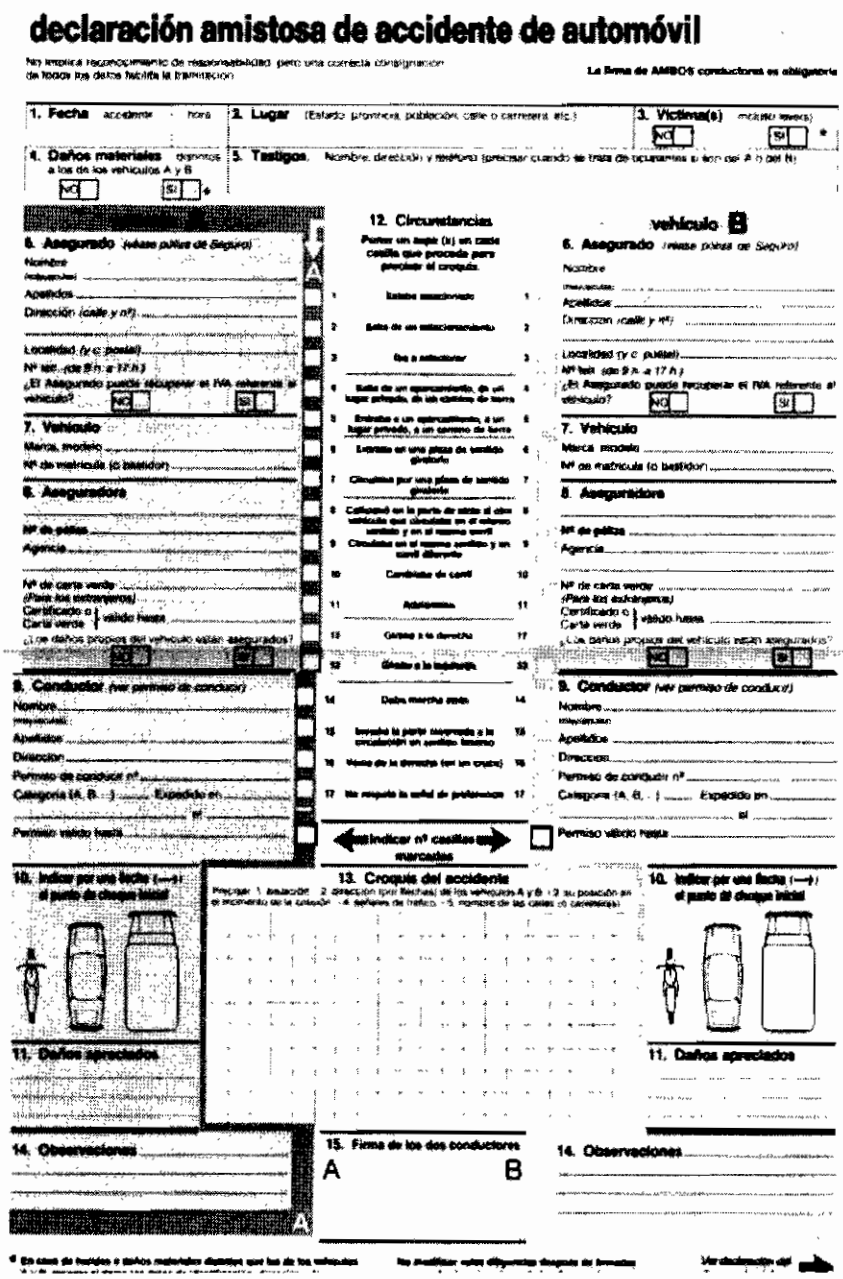
Declaración amistosa:

> El 1 de junio de 1987 se implantó en España un nuevo modelo de impreso para la declaración de siniestros de automóviles, la denominada Declaración

Amistosa de Accidente de Automóvil. Se caracteriza por su uniformidad, común para todas las entidades aseguradoras y para todos los países de la Comunidad Europea. Su aplicación más efectiva se refiere a los accidentes ocurridos con intervención de dos vehículos y por lo que concierne a los daños de los mismos.

Dicha declaración solicita los datos relativos al suceso, tales como fecha, lugar, si ha habido víctimas y daños materiales distintos de los dos vehículos y si ha sido presenciado por testigos. A continuación, en dos columnas laterales, se preguntan los datos relativos a cada uno de los vehículos intervinientes, a sus propietarios, a sus conductores y a sus respectivas compañías aseguradoras. La columna central es la más importante. Se refiere a las circunstancias del hecho y detalla hasta un total de 17 supuestos para cada vehículo. Únicamente se ha de señalar con un aspa aquella descripción que más se aproxime a la realidad del suceso ocurrido. Seguidamente, unos gráficos permiten señalar el punto de choque inicial de cada automóvil y después los daños apreciados y las observaciones que cada conductor considere oportuno declarar. Un croquis del lugar del accidente completa los elementos para poder enjuiciar el siniestro.

El factor que da verdadera validez al parte es la firma de los conductores. Ésta constituye una aportación muy importante al trámite del siniestro, puesto que la firma conjunta en una sola declaración da lugar a una versión única. De esta forma, se descarta el grave problema de los partes de accidente clásicos en que, en muchos casos, las versiones no coinciden. La declaración con una versión única y la firma de los dos conductores abre el camino para que la liquidación del siniestro pueda ser practicada en un plazo muy corto de tiempo. El impreso consta de dos hojas idénticas, que se cumplimentan al mismo tiempo, y cada conductor se queda con una de ellas. En el reverso de la hoja figura la petición de una serie de datos complementarios, que cada conductor o tomador debe facilitar a su aseguradora. Estos datos complementarios también sirven para suministrar información sobre lesionados o sobre daños a otros vehículos, animales o cosas.



declaración amistosa de accidente de automóvil

No implica reconocimiento de responsabilidad, pero una correcta consignación de todos los datos facilita la tramitación

La firma de AMBOS conductores es obligatoria

1. Fecha accidente hora
2. Lugar (Estado, provincia, población, calle o carretera, etc.)
3. Víctima(s) (indicar leves)
4. Daños materiales (ajenos a los de los vehículos A y B)
5. Testigos. Nombre, dirección y teléfono (precisar cuando se trate de ocupantes u otro del A o del B)

vehículo A

6. Asegurado (véase póliza de Seguro)
Nombre
Apellidos
Dirección (calle y nº)
Localidad (y c. postal)
Nº tel. (de 9 h. a 17 h.)
¿El Asegurado puede recuperar el IVA referente al vehículo?
7. Vehículo
Marca, modelo
Nº de matrícula (o bastidor)
8. Aseguradora
Nº de póliza
Agencia
Nº de carta verde (Para los extranjeros) Certificado o Carta verde válido hasta
¿Los daños propios del vehículo están asegurados?
9. Conductor (ver permiso de conducir)
Nombre
Apellidos
Dirección
Permiso de conducir nº
Categoría (A, B, ...) Expedido en
Permiso válido hasta
10. Indicar por una flecha (→) el punto de choque inicial
11. Daños apreciados
14. Observaciones

12. Circunstancias
Poner un aspa (X) en casa casilla que proceda para precisar el croquis.

1	Estaba estacionado	1	
2	Salía de un estacionamiento	2	
3	Iba a estacionar	3	
4	Salía de un aparcamiento, de un lugar privado, de un camino de tierra	4	
5	Entraba a un aparcamiento, a un lugar privado, a un camino de tierra	5	
6	Entraba en una plaza de sentido giratorio	6	
7	Circulaba por una plaza de sentido giratorio	7	
8	Cambiaba de carril	8	
9	Circulaba en el mismo sentido y en carril diferente	9	
10	Cambiaba de carril	10	
11	Adelantaba	11	
12	Giraba a la derecha	12	
13	Giraba a la izquierda	13	
14	Daba marcha atrás	14	
15	Invadía la parte reservada a la circulación en sentido contrario	15	
16	Venía de la derecha (en un cruce)	16	
17	No respetó la señal de prioridad	17	

Indicar nº casillas marcadas

13. Croquis del accidente
Precisar: 1. situación 2. dirección (con flechas) de los vehículos A y B 3. su posición en el momento de la colisión 4. señales de tráfico 5. nombre de las calles o carreteras

15. Firma de los dos conductores
A B

vehículo B

6. Asegurado (véase póliza de Seguro)
Nombre
Apellidos
Dirección (calle y nº)
Localidad (y c. postal)
Nº tel. (de 9 h. a 17 h.)
¿El Asegurado puede recuperar el IVA referente al vehículo?
7. Vehículo
Marca, modelo
Nº de matrícula (o bastidor)
8. Aseguradora
Nº de póliza
Agencia
Nº de carta verde (Para los extranjeros) Certificado o Carta verde válido hasta
¿Los daños propios del vehículo están asegurados?
9. Conductor (ver permiso de conducir)
Nombre
Apellidos
Dirección
Permiso de conducir nº
Categoría (A, B, ...) Expedido en
Permiso válido hasta
10. Indicar por una flecha (→) el punto de choque inicial
11. Daños apreciados
14. Observaciones

No obstante, desde el 1 de mayo de 1990 está en vigor un Acuerdo Suplementario del Convenio de Indemnización Directa Español (**ASCIDE**), que es de aplicación a aquellos siniestros que escapan del ámbito del CIDE, fundamentalmente por el hecho de no haberse cumplimentado la Declaración Amistosa de Accidentes de Automóvil o no ser ésta válida por carecer de alguno de los requisitos exigidos. Las entidades que se integran en el ASCIDE tienen que, previa o simultáneamente, estar adheridas al CIDE. El convenio CIDE es aplicable en los accidentes ocurridos en todo el territorio del

Estado Español y Andorra. También se aplicará cuando los accidentes se produzcan en los países integrados en el Sistema Internacional de Seguro y los vehículos intervinientes tengan contratado el seguro mediante pólizas españolas emitidas por entidades adheridas al convenio. En cuanto a la tramitación del siniestro vía CIDE, la culpabilidad será imputada al vehículo que resulte culpable según las Tablas de Culpabilidad que contiene el convenio. Dichas tablas describen distintas maniobras y situaciones posibles.

En el supuesto de que el accidente se hubiera producido en una situación no comprometida en las Tablas de Culpabilidad, la determinación del responsable se efectuará en función de las disposiciones del Código de Circulación. A los efectos del convenio, debe entenderse por entidad acreedora a la aseguradora del perjudicado, y por entidad deudora a la aseguradora del responsable. La acreedora formulará reclamación a la deudora por el conducto más rápido posible y ésta deberá contestar en el plazo de 72 horas sobre la aceptación o no del caso.

La respuesta negativa sólo podrá apoyarse en la inexistencia en su cartera de póliza de seguro de responsabilidad civil obligatoria del vehículo que se indique. También es motivo de negativa el rehúse total del siniestro a su asegurado, conforme a las disposiciones que configuran el seguro de responsabilidad civil de suscripción obligatoria en su aspecto de daños materiales.
El hecho de que una aseguradora no haya recibido la declaración amistosa, no la exime de cumplir las obligaciones del convenio, si tal declaración le ha sido facilitada por la acreedora.

Liquidación de siniestros vía CIDE

El convenio establece un límite cuantitativo, incluido el IVA y los gastos de traslado del vehículo siniestrado. Si la cuantía de los daños o el valor venal de vehículo excede del límite, el siniestro queda al margen del convenio y, por tanto, su tramitación será la utilizada para los demás siniestros.

La valoración de los daños del vehículo asegurado no culpable la efectuará el perito tasador designado por la acreedora, siendo a cargo de ésta los honorarios y gastos que se causen. Cuando la acreedora prevea que los daños o el valor venal pueden superar el límite establecido, tiene la obligación de ofrecer a la entidad deudora la posibilidad de peritación. Si la compañía acreedora no recibe objeción ninguna por parte de la deudora, tiene que abonar rápidamente el importe de la reparación de los daños sufridos por el vehículo de su asegurado. El hecho de que sea la aseguradora del perjudicado (acreedora) la que haga efectivo el coste de la reparación, en aras a la mayor celeridad, exige, lógicamente, que sea reembolsada por la aseguradora del causante (deudora). Sin embargo, en lugar de

tomar como base del recobro el importe real de la reparación, se ha establecido como fórmula de compensación un módulo determinado por el coste medio de los siniestros amparados por el convenio.

De todo lo anterior se deduce que la entidad acreedora abonará los daños sufridos por el vehículo que asegura, en la cuantía que estipule el dictamen pericial, hasta el límite cuantitativo establecido, y recobrará siempre el módulo vigente. Puede darse el caso de que la valoración supere el valor venal del vehículo y que el interesado no acepte el cobre de dicho valor. También, puede ocurrir que el perjudicado no admita el precio de la reparación fijado por el perito. En ambas situaciones, si el perjudicado reclama judicialmente y consigue su pretensión, la que deberá hacer efectivo el importe definitivo será la entidad deudora, en su condición de aseguradora del culpable. Pero en aplicación estricta del convenio, esta deberá ser reembolsada por la acreedora, de manera que sólo quede a cargo de la entidad deudora el módulo establecido. Aún cuando la sentencia establezca una valoración superior, el siniestro quedará dentro del convenio cuando el valor venal o el de los daños, fijados por los peritos designados, no excedan del límite cuantitativo fijado.

La Comisión de Vigilancia y Arbitraje tiene como funciones interpretar el convenio, velar por su cumplimiento, y dar solución a las cuestiones conflictivas que puedan suscitarse entre las entidades adheridas. Las decisiones de esta comisión son inapelables y obligan a las entidades aseguradoras adheridas, que quedan comprometidas a acatarlas.

En http://www.tirea.es/manuales/marco.htm encontramos el Manual de Criterios de las Comisiones CIDE / ASCIDE / CICOS de la Comisión Técnica de Seguros de Automóviles, manual editado el 10-5-2001. De él hemos extraído el límite, la evolución coste medio sectorial-límite cuantitativo aplicación Convenio y el valor venal que detallamos a continuación:

Límite

Cuando no existe pérdida total lo que determina el límite de los Convenios es el valor de los daños incluido remolcaje/rescate e IVA, aún cuando no se haga efectivo el pago de alguno de éstos. No se tendrán en cuenta depreciaciones por uso ni descuentos de talleres reparadores. Cuando existe pérdida total lo que determina el límite de los Convenios es el valor venal (sin deducción de restos) más el remolcaje/rescate si lo hubiese y no el valor de la reparación u otros. De existir daños sin reparar previos al accidente, no se tendrán en cuenta para determinar el valor venal. Si no hay acuerdo entre las partes sobre el importe de los daños o valor venal, en cuanto a si sobrepasan o no el límite de los

Convenios, deberán de común acuerdo nombrar un tercer perito que finalmente dictamine de forma vinculante para ambas partes. Tendrá también carácter de tercería la intervención de un perito judicial con ocasión de la existencia de actuaciones judiciales. En caso de tercería, el total de los gastos de la intervención del 3º perito, deben asumirlo al 50% entre las dos Entidades. Cuando la Acreedora prevea que el valor de los daños o el valor venal pueden exceder del límite de los Convenios, debe ofrecer la peritación a la presunta Deudora. Si finalmente el importe de los daños supera el límite y la Entidad Acreedora no ha ofrecido en tiempo la peritación a la Deudora, son de aplicación los Convenios, debiendo la Acreedora indemnizar a su asegurado el total de sus daños, incluso por encima del límite, y ello como medida de protección de quién pudiendo no tuvo posibilidad de intervenir en la peritación. Sin embargo, con el ánimo de flexibilizar estas situaciones, cuando resulte evidente el exceso del límite no se aplicarán los Convenios. A estos efectos, se considera evidente el exceso, cuando los daños superen el límite más el 50% del mismo. Como esta ampliación de criterio supone un cambio sustancial respecto a la situación anterior, esta modificación entró en vigor para los accidentes ocurridos a partir del 1.6.96 incluido.

Cuando la Entidad Acreedora solicita el abono del Módulo significa que ha indemnizado los daños del vehículo de su asegurado, por lo que no cabe alegar posteriormente exceso de límite. Informáticamente (en el sistema CICOS que en inmediato detallamos) no se puede comunicar el posible exceso de límite hasta que no se recibe el mensaje "aceptamos reclamación". Sin embargo, ello no exime de la obligación de ofrecer la peritación por fax o correo electrónico en el momento que se conoce el posible exceso de límite. En los casos en que la Entidad Acreedora recibe aceptación, no solicita Módulo, no indemniza a su asegurado y acude a la vía judicial, la Deudora sólo debe pagar el exceso del límite, siendo a cargo de la Acreedora el límite menos un Módulo más las costas e intereses.

Evolución coste medio sectorial-límite cuantitativo aplicación Convenio

Fecha	C.M.S (Ptas)	C.M.S (Euros)	Límite (Ptas)	Límite (Euros)
01.01.1988	40 000	240	500 000	3 006
01.01.1989	40 000	240	500 000	3 006
01.01.1990	50 000	301	1 000 000	6 010
01.01.1991	60 000	361	1 000 000	6 010
01.01.1992	70 000	421	1 000 000	6 010
01.01.1993	70 000	421	1 000.000	6 010
01.01.1994	70 000	421	1 000.000	6 010

01.01.1995	75 000	451	1 000 000	6 010
01.01.1996	75 000	451	2 500 000	15 025
01.01.1997	90 000	541	5 000 000	30 051
01.01.1998	90 000	541	16 000 000	96 162
01.01.1999	90 000	541	16 000 000	96 162
01.01.2000	95 000	571	16 000 000	96 162
01.01.2001	95 006	571	16 000 000	96 162
13.02.2001	95 006	571	16 638 000	100 000

Tabla 2.1. Coste medio sectorial.

Valor Venal

Cuando existe pérdida total, lo que determina el límite de los Convenios es el valor venal (sin deducción de restos) más el remolcaje/rescate si lo hubiese, y no el valor de reparación u otros.

Si existieran en el vehículo daños previos al accidente, correspondientes a otros accidentes, no se tendrán presentes en la estipulación del valor venal.

Cuando el valor venal es inferior al límite, y el perjudicado no acepta la indemnización con arreglo a éste, si el valor de reparación es superior al límite y existe sentencia concediendo el valor de reparación, la Acreedora asumirá hasta el límite del Convenio y la Deudora el exceso de este más el correspondiente Módulo de compensación.

Cuando se prevea que el valor venal pueda exceder del límite de los Convenios, debe ofrecerse la peritación a la otra Entidad. Si no hay acuerdo entre las partes sobre valor venal, en cuanto a si sobrepasa o no el límite de los Convenios, deberán de común acuerdo nombrar un tercer perito que finalmente dictamine de forma vinculante para ambas partes. Tendrá también carácter de tercería la intervención de un perito judicial con ocasión de la existencia de actuaciones judiciales.

En caso de tercería, el total de los gastos de la intervención del tercer perito, deben asumirlo al 50% entre las dos Entidades.

Si finalmente el valor venal supera el límite y la Entidad Acreedora no ha ofrecido en tiempo la peritación a la Deudora, son de aplicación los Convenios, debiendo la Acreedora indemnizar a su asegurado el valor venal, incluso por encima del límite, y ello como medida de protección de quién pudiendo no tuvo posibilidad de intervenir en la peritación. Sin embargo, con el ánimo de flexibilizar estas situaciones, cuando resulte evidente el exceso del límite no se aplicarán los Convenios. A estos

efectos, se considera evidente el exceso, cuando el valor venal supere el límite más el 50% del mismo. Como esta ampliación de criterio supone un cambio sustancial respecto a la situación anterior, esta modificación entrará en vigor para los accidentes ocurridos a partir del 1.6.96 incluido.

1.2. Sistema CICOS

El sistema CICOS es el entorno informático por el que se tramitan los convenios de indemnización de daños materiales entre todas las entidades de automóviles (CIDE y ASCIDE), en el que se ha introducido la transmisión de documentos escaneados y distribuidos mediante Infovía, habiendo logrado además crear un sistema multilateral bancario como cámara de compensación de todos los siniestros tramitados al año (más de 1 500 000 siniestros).

En lo que a la parte técnica se refiere Tecnologías de la Información y Redes para las Entidades Aseguradoras TIREA) es la responsable a través del sistema Tire@Cicos,

http://www.tirea.es/servicios/cicos/servicio_cicos.htm.

Las comunicaciones entre Entidades se realizan mediante los códigos de mensajes tipificados en el Reglamento CICOS y a través de un Sistema de Intercambio Electrónico de Documentos (EDI). El Centro Compensador residente en TIREA, realiza un cálculo mensual de los saldos que cada Entidad debe o le deben con respecto a las restantes Entidades. La liquidación de saldos se realiza mediante recibos y transferencias bancarias entre la cuenta del Centro de Compensación y las cuentas de cada Entidad.

Adicionalmente TIREA pone a disposición de las Entidades Aseguradoras el Servicio Tire@SDM: Gestión informatizada de los Siniestros de Daños Materiales que quedan excluidos de los convenios CIDE / ASCIDE,

http://www.tirea.es/servicios/sdm/servicio_sdm.htm.

La buena experiencia obtenida del sistema CICOS constituido en 1994 para mecanizar y regular los intercambios de información y saldos resultantes de la aplicación de los convenios CIDE/ASCIDE, planteó la necesidad de crear un sistema que permitiera la gestión informatizada de los siniestros excluidos en dichos convenios. Aprovechando la plataforma tecnológica que soporta TIREA, se dota de esta manera de la tecnología CICOS a aquellos siniestros/conceptos de daños materiales no CICOS.

El sistema Tire@SDM, basa sus funcionalidades en las reuniones mantenidas entre TIREA y el grupo de expertos creado al efecto por representantes del Ramo del Automóvil. Un diálogo automatizado a través de mensajes, permite a las Entidades Aseguradoras alcanzar acuerdos en la resolución de dichos siniestros y compensar los importes relativos a los mismos mensualmente mediante transferencia bancaria. Las Entidades cuentan además con la posibilidad de intercambiar documentación a lo largo del proceso de tramitación. A través de este sistema se gestionan: Siniestros en los que intervienen más de dos vehículos; Siniestros sin colisión directa; Daños causados por carga desprendida; Daños materiales ajenos a los vehículos o perjuicios (paralización, cascos, gafas, otros daños); y Siniestros tramitados CIDE/ASCIDE.

Los beneficios son que las Entidades Aseguradoras podrán alcanzar ventajas operativas y estratégicas al aprovechar la sinergia que ofrece la explotación conjunta de servicios dentro del Ramo del Automóvil. TIREA como proveedor de servicios para el Sector Asegurador pretende identificar los aspectos que puedan aportar mayor valor a la explotación de estos servicios y potenciarlos.

2. Ficheros sectoriales

En los últimos años, desde la Comisión Técnica de Seguros de Automóviles de UNESPA, se han creado varios ficheros sectoriales [Font (1999)]:

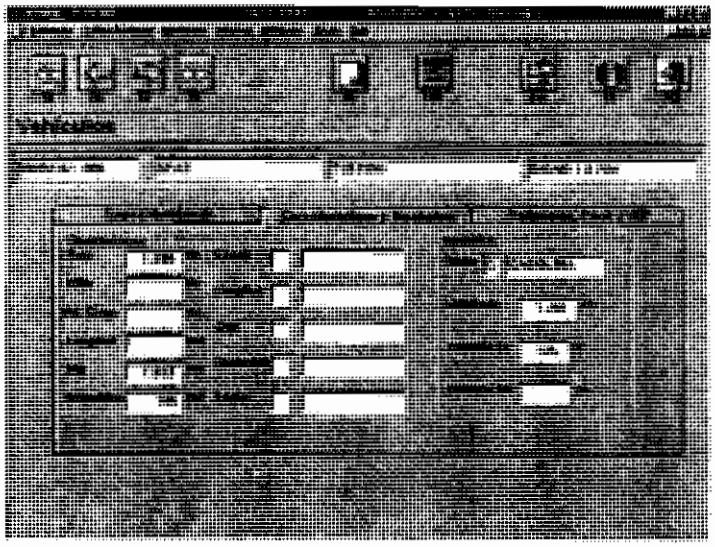
2.1. Base SIETE

El Sistema Informativo de Especificaciones Técnicas, Base SIETE, es una base de datos elaborada por CENTRO ZARAGOZA (http://www.centro-zaragoza.com/) en la que se incluyen las principales características técnicas de todos los vehículos automóviles susceptibles de ser asegurados en España.

Este producto se gesta como respuesta a la demanda del sector asegurador, y por tanto dirigido a él, de una base de datos fiable y constantemente actualizada en la que figuren los principales datos que cada una de las entidades aseguradoras precisan para el cálculo de sus primas.

Todos los vehículos automóviles contemplados en la base de datos disponen de un código de identificación, el Código Base SIETE, que es único para cada versión específica de los vehículos, el cual permite un ágil tratamiento de todo tipo de información, soslayando los inconvenientes que suponen las distintas formas en que pueden figurar los literales de marca, modelo y versión.

La Tabla de Vehículos, núcleo principal de Base SIETE, está estructurada en un total de 37 campos, en los que figuran cada una de las principales características de los vehículos automóviles:



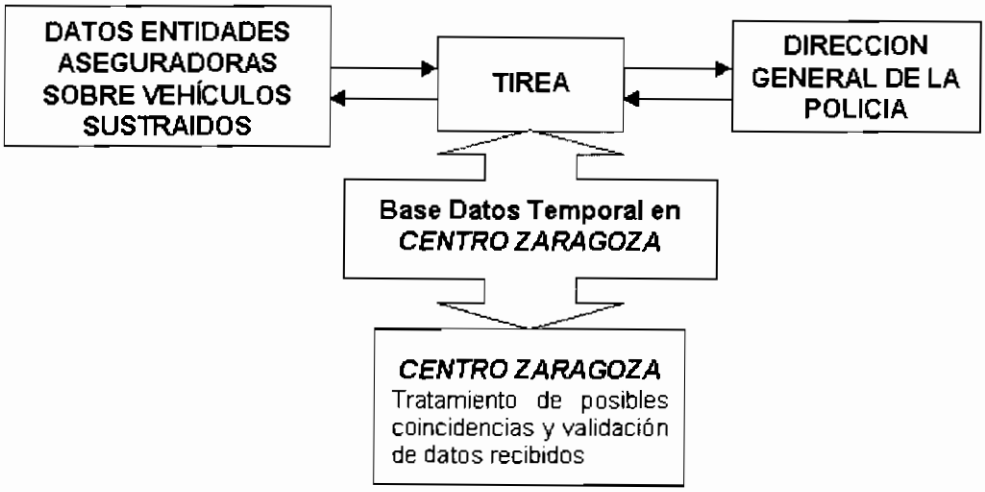
Es de destacar que la base de datos de vehículos de Base SIETE incluye todos los tipos de vehículos automóviles de las tres categorías aseguradoras. Actualmente cuenta con más de 30.000 registros, correspondientes a otras tantas versiones específicas. La clasificación de los distintos tipos de vehículos se realiza con los campos: CATEGORÍA, TIPO y CLASE. Cada vehículo pertenece a una de las tres CATEGORÍAS aseguradoras, y dentro de ésta a un determinado TIPO, del que a su vez pueden distinguirse distintas CLASES.

2.2. El Fichero de Vehículos Sustraídos e Indemnizados

Servicio destinado a la localización de vehículos sustraídos, mediante el intercambio de información a través de un Fichero accesible a través de Internet.

La información contenida en dicho Fichero se comparte con la Dirección General de la Policía de forma que cuando ésta recupere cualquier vehículo, las Entidades estén puntualmente informadas del lugar de localización y puedan proceder a su recogida. De esta manera, se produce una colaboración más estrecha entre las Aseguradoras y las Fuerzas de Seguridad del Estado.

Actualmente, el Fichero radica en Centro Zaragoza y únicamente se registrarán los datos de vehículos sustraídos que hayan sido indemnizados por la Entidad Aseguradora:



La labor de TIREA es la de asegurar la transmisión de la información y gestionar las comunicaciones y accesos de los usuarios, manteniendo un registro de los ficheros enviados y recibidos, a través del servicio Tire@FVSI,

http://www.tirea.es/servicios/fvsi/servicio_fvsi.htm.

2.3. La Estadística del Seguro del Automóvil

TIREA pone a disposición del Sector Asegurador el servicio Estadística del Seguro del Automóvil (ESA), para la consulta de forma on-line de la información resultante de la explotación de los datos aportados por las Entidades participantes. El Servicio responde a las necesidades de las Entidades de obtener información sobre el riesgo elemental, que permita el conocimiento completo del mercado en el que operan.

Pretende recuperar la Estadística del Seguro del Automóvil de UNESPA para así coordinar los esfuerzos del Sector.

Se realiza a través del servicio Tire@ESA,

http://www.tirea.es/servicios/esa/servicio_esa.htm,

que está compuesto por una base de datos con el total de expuestos y siniestros aportados para cada una de las garantías contempladas. Las Entidades Aseguradoras facilitan una carga de expuestos y siniestros por cada uno de los años contemplados para la Estadística. Y las consultas de los datos resultantes de la explotación se realizan de forma on-line. Existe la posibilidad de descargar de las consultas en ficheros Excel.

Como proveedor de Servicios Básicos de Telecomunicación para el Sector Asegurador, TIREA proporciona la infraestructura necesaria para soportar la interconexión de las Entidades a través de la red privada del Seguro, Tire@Net.

2.4. El Fichero Histórico de SINiestralidad de COnductores (SINCO)

El sector asegurador se basa fundamentalmente en una correcta valoración del riesgo, pero para tener una buena tarificación es necesario poseer una excelente base estadística. Es decir, la aplicación de criterios tarifarios equitativos, tal y como se viene haciendo en todos los países de nuestro entorno, requiere, dentro del ramo del seguro del automóvil, el conocimiento del comportamiento de la siniestralidad de quienes contratan la póliza. Y esto únicamente puede llevarse a cabo mediante la creación de un "Fichero Histórico".

Así, en noviembre del 2000, empezó a funcionar el Fichero Histórico de Seguros de Automóviles (SINCO), después de un dilatado período de diseño y desarrollo en el que ha tenido un gran protagonismo la Comisión Técnica de Seguros de Automóviles. Se trata de un fichero de datos de carácter personal constituido por las compañías de seguros del automóvil con el fin de permitir la tarificación, la selección de riesgos y la elaboración de estudios de técnica aseguradora.

El fichero se crea según lo dispuesto en la ley 30/1995 de Ordenación y Supervisión de los Seguros Privados, que permite a las entidades aseguradoras constituir ficheros comunes que contengan datos de carácter personal para la liquidación de siniestros. El contenido del mismo está elaborado de acuerdo con la Ley Orgánica 15/1999 de Protección de Datos de Carácter Personal, y únicamente tienen acceso a los datos las entidades adheridas a dicho fichero, quienes lo podrán utilizar para realizar consultas en el momento de la solicitud de nuevas pólizas. De este modo, pueden realizar una valoración técnica y objetiva del riesgo para aplicar correctamente las tarifas de prima que tengan recogidas en sus bases técnicas.

El SINCO, al proporcionar información objetiva sobre la siniestralidad del tomador del seguro referente a los últimos 5 años, permite ajustar la prima que realmente le corresponde a cada asegurado. Hasta el momento, podían existir casos de asegurados que estaban expuestos a riesgos superiores a la prima que pagaban. En consecuencia, había personas que pagaban primas superiores a la valía de su riesgo. El fichero, en el medio plazo, tenderá a equilibrar este efecto, ya que cada uno pagará lo que en realidad le corresponda, ni más ni menos.

El Fichero contiene los siguientes datos para cada asegurado:

- Vehículo asegurado
- Datos del tomador del seguro
- Datos del contrato: coberturas contratadas y período de vigencia, por lo que no incluye información sobre tarifas
- Datos del siniestro: cobertura afectada, fecha del siniestro y existencia de daños corporales o materiales

Actualmente, aproximadamente el 85% del mercado asegurador está adherido al fichero, que puede contener información de hasta los últimos 5 años del asegurado. En cuanto a los siniestros, únicamente tienen cabida los de responsabilidad civil a terceros. El fichero no contiene datos de otras coberturas.

Respecto a la adscripción de las entidades al Fichero, ésta no se ha producido en un determinado momento, sino que van adhiriéndose de forma escalonada

Para que una compañía adherida al fichero pueda realizar una consulta sobre el historial como conductor de un determinado asegurado, éste le tiene que facilitar una serie de datos. Así, para que la consulta se lleve a cabo, la entidad tiene que saber:

- Cuáles son las últimas cinco cifras de la póliza
- y uno de los siguientes datos: la matrícula de vehículo, el DNI del tomador o su nombre y apellidos.

Resulta completamente imposible reunir dos de los datos necesarios (uno de ellos la terminación de la póliza) si el asegurado no lo comunica. Es decir, aunque figure en el fichero, el asegurado es el único dueño de la información sobre él. Todas las compañías adheridas al SINCO están conectadas por una red informática que les permite leer la información pero no copiar o reescribir sobre los datos que aparecen en la pantalla del ordenador. La seguridad del fichero es absoluta.

En lo que a la parte técnica se refiere TIREA es la responsable del diseño y mantenimiento de la arquitectura de programación y de comunicaciones del Fichero a través del servicio Tire@Sinco,

http://www.tirea.es/servicios/sinco/servicio_sinco.htm.

TIREA proporciona la infraestructura necesaria para la interconexión de las Entidades a través de la red privada del Seguro, Tire@Net.

ANEXO 2.2. Datos de Baxter

POLICY HOLDER AGE (PA)	CAR GROUP (CG)	VEHICLE AGE (VA)							
		0 - 3		4 - 7		8 - 9		10 & Over	
		Mean Claim Amount	Number of Claims	Mean Claim Amount	Number of Claims	Mean Claim Amount	Number of Claims	Mean Claim Amount	Number of Claims
17-20	A	289	8	282	8	133	4	160	1
	B	372	10	249	28	288	1	11	1
	C	189	9	288	13	179	1	0	0
	D	763	3	850	2	0	0	0	0
21-24	A	302	18	194	31	135	10	166	4
	B	420	59	243	96	196	13	135	3
	C	268	44	343	39	293	7	104	2
	D	407	24	320	18	205	2	0	0
25-29	A	268	56	285	55	181	17	110	12
	B	275	125	234	172	179	36	264	10
	C	334	163	274	129	208	18	150	8
	D	383	72	305	50	116	6	636	1
30-34	A	236	43	270	53	160	15	110	12
	B	259	179	226	211	161	39	107	19
	C	340	197	260	125	189	30	104	9
	D	400	104	345	55	147	8	65	2
35-39	A	207	43	129	73	157	21	113	14
	B	208	191	214	219	149	46	137	23
	C	251	210	232	131	204	32	141	8
	D	233	119	325	43	207	4	0	0
40-49	A	254	90	213	98	149	35	98	22
	B	218	380	209	434	172	97	110	59
	C	239	401	250	253	174	50	129	15
	D	387	199	299	88	325	8	137	9
50-59	A	251	69	227	120	172	42	98	35
	B	196	366	229	353	164	95	132	45
	C	268	310	250	148	175	33	152	13
	D	391	105	228	46	346	10	167	1
60 & Over	A	264	64	198	100	167	43	114	53
	B	224	228	193	233	178	73	101	4
	C	269	183	259	103	227	20	119	6
	D	385	62	324	22	192	6	123	6

Tabla 2.2. Datos de Baxter.

ANEXO 2.3. Datos de impagos

		A: Antigüedad laboral											
		A1: Menos de 2 años				A2:Entre 2 y 10 años				A3:Más de 10 años			
E: Estado civil	E1: Aparejado	363.10	540.88	15.10	24.92	173.69	12.82	403.24	350.82	48.56	386.37	394.34	685.96
		267.58	523.66	371.92	94.36	63.47	142.28	221.20	497.90	57.06	103.11	145.99	162.40
		66.82	103.35	166.47	261.02	414.06	608.41	196.30	12.11	518.60	331.60	95.04	602.69
		93.05	140.82	312.18	109.34	205.19	311.53	400.06	122.62	287.28	105.88	110.22	203.00
		295.78	196.93	103.37	81.67	295.89	496.28	405.70	457.48	480.38	643.38	463.98	591.94
		59.36	298.89	452.58	150.11	536.37	268.45	203.34	168.02	423.87	724.22	129.01	239.66
		313.61	96.28	293.79	87.44	48.29	111.55	398.01	309.32	711.86	648.69	469.67	221.54
		395.56	187.40	286.09	137.64	74.51	168.30	90.92	381.98	259.00	783.54	516.97	182.72
		118.71	136.81	96.27	113.57	100.01	33.02	609.78	748.68	505.18	814.95	203.34	297.93
		51.60	332.67	403.13		20.97	84.40	366.05		177.51	251.56	509.60	556.78
										469.48	133.16	274.23	208.54
	E2: Separado Divorciado	54.92	109.34	268.77	47.18	513.14	97.04	316.37	483.78	215.06	414.72	238.92	157.63
		164.22	154.87	209.69	87.88	40.35	150.13	133.37	134.27	440.24	529.82	313.01	89.62
		396.50	184.61	80.80	321.16	529.67	332.02	532.93	147.06	48.16	471.13	388.53	36.85
		554.12	55.57	166.02	86.70	260.59	184.14	145.99	192.16	10.29	272.92	139.42	99.73
		163.77	173.10	148.62	425.24	425.27	102.29	42.67	149.56	382.19	520.18	77.37	290.97
		52.67	7.06	71.84	300.52	163.77	398.88	157.20	309.78	374.19	258.75	314.01	78.43
		77.99	486.38	84.18	69.13	640.47	77.34	202.68	98.96	563.36	232.75	60.60	
		65.40	176.24	254.47	141.89	53.00	47.94	331.85	111.04	271.07	94.52	134.11	
		102.35	62.66	97.79	134.91	272.63	319.02	27.32	166.00	37.10	421.51	163.87	
		107.72	67.74	80.97	118.57	81.67	97.72	563.61	56.23	36.43	487.45	290.97	
		79.17	26.78	665.58	180.37	378.23	416.72	273.65	134.27	324.44	380.65	209.32	
		208.78	305.27	73.22	106.38	494.40	456.57	228.23		133.16	416.76	209.51	
		106.38	213.05	321.15		264.02	107.96	194.81		392.61	277.07	143.69	
		93.98	5.05	491.71		95.01	150.90	46.66		135.26	214.37	361.60	
	E3: Soltero	75.30	156.06	242.69	265.84	158.72	101.68	165.78	47.99	293.78	155.79	295.07	157.63
		7.72	43.80	113.68	231.10	616.01	358.21	477.92	638.10	185.62	34.23	447.74	145.27
		133.02	484.22	196.30	10.29	405.37	56.18	107.48	97.31	488.89	170.73	242.96	314.85
		202.63	90.07	380.65	325.23	53.58	513.20	253.04	133.48	314.31	252.84	298.26	400.06
		223.68	27.06	388.53	68.63	123.04	82.79	406.20	113.57	46.47	84.31	222.98	56.56
		199.95	122.02	326.63	87.61	34.98	151.52	58.31	177.84	49.86	358.32	219.79	210.21
		49.62	80.97	91.41	84.79	316.44	342.59	77.93	406.86	483.03	287.12	539.43	251.92
		63.97	61.24	326.63	27.48	323.07	402.08	103.37	225.10	73.09	411.58	22.68	561.44
		293.02	133.68	106.38	314.31	364.56	413.70	400.06	234.76	527.52	434.72	420.87	
		218.24	445.67	266.48	248.60	221.55	257.69	550.82	406.86	214.25	502.44	145.92	
						147.45	56.09	25.05		37.12	247.09	117.82	

Tabla 2.3. Datos de impagos mostrando la información desagregada.

	A1	A2	A3
E1	208.816 | 39	269.565 | 39	366.609 | 44
E2	172.045 | 54	232.667 | 53	253.215 | 48
E3	180.380 | 40	246.705 | 43	261.575 | 41

Tabla 2.4. Datos de impagos mostrando la información agregada.

Capítulo 3

Selección de variables de tarifa

El primer paso dentro del proceso de tarificación *a priori* es el de selección de las variables de tarifa y sus clases, a partir de un conjunto de factores potenciales de riesgo. En este capítulo revisamos las técnicas estadísticas aplicadas en el campo actuarial para llevar a cabo dicha selección.

Nos centramos en las dos metodologías principalmente utilizadas en los seguros no vida: el análisis de segmentación y el modelo lineal generalizado. Respecto de cada una de estas metodologías se ha efectuado un resumen de sus fundamentos teóricos, prestando luego una especial atención a su aplicabilidad en la tarificación de los seguros. Para ello, se han analizado también las distintas dificultades empíricas y el software disponible con sus limitaciones.

El presente capítulo se estructura en seis apartados y tres anexos.

Dedicamos inicialmente un apartado, 3.1, al estudio de la relación de dependencia entre la variable univariante cuantitativa *experiencia de siniestralidad* (número de siniestros por póliza, o cuantía de un siniestro, o bien cuantía total de los siniestros) y cada factor de riesgo univariante, cuantitativo o cualitativo de forma individualizada. Para ello utilizamos medidas de asociación entre pares de variables, que también nos permiten estudiar las relaciones entre factores.

Puesto que el objetivo es la selección del conjunto de variables de tarifa que mejor explique la estructura de riesgo, debemos introducir técnicas de análisis estadístico multivariante, las cuales nos permiten organizar procesos de selección teniendo en cuenta simultáneamente el conjunto de factores. En el apartado 3.2 clasificamos dichas técnicas según su filosofía: técnicas de regresión, de análisis discriminante y de análisis cluster, y según las fases de la tarificación que nos permiten cubrir: algunas nos sirven para realizar una tarificación completa, y otras para cubrir tan sólo algunas de las fases de la obtención de la estructura de tarifa.

En el apartado 3.3, citamos las diferentes metodologías que se encuentran en la bibliografía actuarial. Cada una incorpora sus hipótesis y, con ellas ventajas e inconvenientes; así como un coste computacional mayor o menor. Generalmente se recomienda la utilización de varios métodos para decidir finalmente un "buen" subconjunto de variables tarificadoras. Los métodos utilizados deberían coincidir aproximadamente en los resultados obtenidos.

En los apartados 3.4 y 3.5, dedicamos especial atención al funcionamiento técnico de dos de las metodologías por ser las mayormente utilizadas: el análisis de segmentación y los modelos lineales generalizados. De ambas hacemos uso en las aplicaciones del capítulo 5.

Finalmente dedicamos un apartado, 3.6, al estudio de criterios de discretización de variables continuas, distinguiendo si se trata de respuesta o predictores, puesto que, dependiendo de la metodología de selección utilizada, y del tipo de información con el que queramos trabajar, necesitaremos, adicionalmente, de estos criterios.

3.1. Medidas de asociación entre pares de variables

Una medida de asociación, A, entre dos variables, **X** e **Y**, idealmente debe cumplir:

- $A(\mathbf{Y},\mathbf{X}) = A(\mathbf{X},\mathbf{Y})$
- $0 <= A(\mathbf{Y}, \mathbf{X}) <= 1$
- $A(\mathbf{Y}, \mathbf{X}) = 0$ si hay independencia estocástica entre **Y** y **X**
- $A(\mathbf{Y}, \mathbf{X}) = 1$ si hay alguna relación funcional entre **Y** y **X**

Como medidas de asociación, distinguiendo el tipo de variables, tenemos [Cuadras (2003)]:

- **Cuantitativa-Cuantitativa**: partimos de dos variables, **X** e **Y**, cuantitativas $(n \times 1)$, en este caso tenemos el coseno del ángulo entre los dos vectores:

$$\cos\alpha = \frac{\mathbf{X}^T\mathbf{Y}}{|\mathbf{X}||\mathbf{Y}|} = \frac{\sum_{i=1}^{n} y_i x_i}{\sqrt{\sum_{i=1}^{n} y_i^2 \sum_{i=1}^{n} x_i^2}}, \tag{3.1}$$

donde α es el ángulo entre **X** e **Y**. El coseno del ángulo es una medida de similitud entre **X** e **Y**. Ésta es independiente de la longitud de los vectores. En aplicaciones estadísticas es más frecuente utilizar el coseno del ángulo entre los vectores centrados, de lo que resulta el coeficiente de correlación, al cual trataremos en valor absoluto, $|\rho|$:

$$\rho = \frac{Cov(\mathbf{Y},\mathbf{X})}{\sqrt{Var(\mathbf{Y})Var(\mathbf{X})}} = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{\sqrt{\left(\sum_{i=1}^{n}(y_i - \overline{y})^2\right)\left(\sum_{i=1}^{n}(x_i - \overline{x})^2\right)}}, \tag{3.2}$$

siendo:

$$Var(\mathbf{Y}) = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n}; \quad Var(\mathbf{X}) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n};$$

$$Cov(\mathbf{Y},\mathbf{X}) = \frac{\sum_{i=1}^{n}(y_i - \overline{y})(x_i - \overline{x})}{n}; \quad \overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}; \quad \overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}.$$

La diferencia esencial entre el coseno y el coeficiente de correlación es que el coseno se basa en las puntuaciones originales (desviaciones del origen), mientras que el coeficiente de correlación se basa en puntuaciones centradas (desviaciones de la media). Por lo tanto, el coseno hace uso de información de escala razón y el coeficiente de correlación de escala intervalo. En general, las variables cuantitativas que analizaremos en las aplicaciones del capítulo 5 serán de la escala intervalo (por ejemplo, la potencia del vehículo, la antigüedad del carnet, la edad del conductor, etc), por lo que nos fijaremos en el coseno de los vectores centrados, ρ.

- **Cuantitativa-Cualitativa**: partimos de una variable, **Y**, cuantitativa $(n \times 1)$ y de una variable, **X**, cualitativa con k clases, de manera que para cada valor x_i de **X** tenemos los valores

$y_{i1}, y_{i2}, \ldots, y_{in_i}$ de **Y** para i = 1,2,...,k, con $\sum_{i=1}^{k} n_i = n$. En este caso, a partir de la descomposición del análisis de la varianza[13],

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_{..}\right)^2 = \sum_{i=1}^{k} n_i \cdot \left(\overline{y}_{i.}-\overline{y}_{..}\right)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_{i.}\right)^2 , \tag{3.3}$$

tenemos la medida:

$$\eta = \left\{ 1 - \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_{i.}\right)^2}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_{..}\right)^2} \right\}^{1/2} \tag{3.4}$$

donde $\overline{y}_{i.} = \dfrac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$ para $i = 1,2,\ldots,k$.

Su interpretación es la siguiente:

$$\eta^2 = 1 - \frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_{i.}\right)^2}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_{..}\right)^2} = 1 - \frac{SCD}{SCT} = \frac{SCE}{SCT} = \frac{\sum_{i=1}^{K} n_i\left(\overline{y}_{i.}-\overline{y}_{..}\right)^2}{\sum_{i=1}^{K}\sum_{j=1}^{n_i}\left(y_{ij}-\overline{y}_{..}\right)^2}$$

$\Rightarrow$ si $\eta = 0$, las medias de cada clase coinciden con la media global y por tanto no hay diferencias entre las clases: $\overline{y}_{i.} = \overline{y}_{..}$

$\Rightarrow$ si $\eta = 1$, no hay diferencia dentro de las clases: $y_{ij} = \overline{y}_{i.}$ para $j = 1,2,\ldots,n_i$

Por otro lado, podemos organizar unas variables binarias, $[\mathbf{X}_i]_{i=1,2,\ldots,k}$, k vectores $(n \times 1)$, que representen la pertenencia a cada una de las k clases de **X**, y buscar unas puntuaciones b_i tales que maximicen la correlación (bien al cuadrado, bien en valor absoluto) de **Y** con la

[13] VTotal = VExplicada + VNExplicada ó SCTotal = SCEntre + SCDentro

combinación lineal $\hat{\mathbf{Y}} = \sum_{i=1}^{k} b_i \mathbf{X}_i$: Maximizar $\left| r\left(\mathbf{Y}, \hat{\mathbf{Y}}\right) \right|$ o $r^2\left(\mathbf{Y}, \hat{\mathbf{Y}}\right)$, resulta que $b_i = \dfrac{\sum_{j=1}^{n_i} y_{ij}}{n_i}$ para $i = 1,2,\ldots,k$. Por lo tanto, para estas puntuaciones, se cumple que $\hat{y}_{ij} = \overline{y}_{i\cdot}$. Si ahora calculamos el coeficiente de correlación maximizado al cuadrado resultante:

$$r^2 = \left[\frac{\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{\cdot\cdot}\right)\left(\hat{y}_{ij} - \overline{y}_{\cdot\cdot}\right)}{\sqrt{\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2\right)\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(\hat{y}_{ij} - \overline{y}_{\cdot\cdot}\right)^2\right)}} \right]^2 = \frac{\left[\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{\cdot\cdot}\right)\left(\hat{y}_{ij} - \overline{y}_{\cdot\cdot}\right)\right]^2}{\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2\right)\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(\hat{y}_{ij} - \overline{y}_{\cdot\cdot}\right)^2\right)} =$$

$$= \frac{\left[\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{\cdot\cdot}\right)\left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)\right]^2}{\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2\right)\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2\right)} = \frac{\left[\sum_{i=1}^{K} n_i \left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2\right]^2}{\left(\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2\right)\left(\sum_{i=1}^{K} n_i \left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2\right)} =$$

$$= \frac{\sum_{i=1}^{K} n_i \left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2}{\sum_{i=1}^{K}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2} = \eta^2$$, obtenemos que, $r^2\left(\mathbf{Y}, \hat{\mathbf{Y}}\right) = \eta^2$.

- **Cualitativa-Cualitativa**: partimos de dos variables, **X** e **Y**, cualitativas con p y q clases respectivamente. Sea N = (n_{ij}) la tabla de contingencia $p \times q$ que las resume. Pueden obtenerse medidas de asociación entre filas y columnas a partir del estadístico *ji*-cuadrado que se utiliza usualmente para contrastar independencia:

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q} \left(n_{ij} - \frac{n_{i\cdot} \times n_{\cdot j}}{n_{\cdot\cdot}} \right)^2 \Bigg/ \left(\frac{n_{i\cdot} \times n_{\cdot j}}{n_{\cdot\cdot}} \right), \tag{3.5}$$

si llamamos $\phi^2 = \dfrac{\chi^2}{n_{\cdot\cdot}}$, Cramér (1946) propone como medida de asociación el siguiente coeficiente que estandariza al estadístico ϕ^2 para un rango entre 0 y1:

$$C = \left\{ \frac{\phi^2}{\min\{(p-1),(q-1)\}} \right\}^{1/2}. \tag{3.6}$$

El significado de C según la interpretación desprendida por la propia *ji*-cuadrado es el siguiente:

$$\chi^2 = \sum_{i=1}^{p}\sum_{j=1}^{q}\left(n_{ij} - \frac{n_{i\cdot} \times n_{\cdot j}}{n_{\cdot\cdot}} \right)^2 \Big/ \left(\frac{n_{i\cdot} \times n_{\cdot j}}{n_{\cdot\cdot}} \right) = n_{\cdot\cdot}\left[\left(\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_{ij}^2}{n_{i\cdot} \times n_{\cdot j}} \right) - 1 \right],$$

$$C^2 = \frac{\chi^2}{s \times n_{\cdot\cdot}} = \frac{1}{s}\left[\left(\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_{ij}^2}{n_{i\cdot} \times n_{\cdot j}} \right) - 1 \right],$$

$\Rightarrow$ si hay independencia: $n_{ij} = \dfrac{n_{i\cdot} \times n_{\cdot j}}{n_{\cdot\cdot}}$, $\chi^2 = 0$,

$$C^2 = \frac{1}{s}\left[\left(\sum_{i=1}^{p}\sum_{j=1}^{q} \frac{n_{i\cdot}^2 \times n_{\cdot j}^2}{n_{\cdot\cdot}^2 \times n_{i\cdot} \times n_{\cdot j}} \right) - 1 \right] = \frac{1}{s}\left[\left(\frac{1}{n_{\cdot\cdot}^2} \sum_{i=1}^{p}\sum_{j=1}^{q} n_{i\cdot} \times n_{\cdot j} \right) - 1 \right] =$$
$$= \frac{1}{s}\left[\left(\frac{1}{n_{\cdot\cdot}^2} \sum_{i=1}^{p} n_{i\cdot} \sum_{j=1}^{q} n_{\cdot j} \right) - 1 \right] = \frac{1}{s}\left[\frac{n_{\cdot\cdot}^2}{n_{\cdot\cdot}^2} - 1 \right] = \frac{1}{s}[0] = 0$$

$\Rightarrow$ si hay dependencia: $n_{ij} = 0$ para $i \neq j$, $\chi^2 = s \times n_{\cdot\cdot}$,

$$C^2 = \frac{1}{s}\left[\left(\sum_{i=j} \frac{n_{ij}^2}{n_{i\cdot} \times n_{\cdot j}} \right) - 1 \right] = \frac{1}{s}\left[\left(\sum_{i=1}^{p} \frac{n_{ii}^2}{n_{i\cdot}^2} \right) - 1 \right] = \frac{1}{s}\left[\left(\sum_{i=1}^{p} 1 \right) - 1 \right] = \frac{(p-1)}{(p-1)} = 1$$

En bibliografía especializada como Andenberg (1973) pp. 70-92, encontramos otras variedades como:

Coeficiente de Tschuprow:

$$T = \left\{ \frac{\phi^2}{\{(p-1)(q-1)\}^{1/2}} \right\}^{1/2} \tag{3.7}$$

que propone como factor de normalización de ϕ^2 la media geométrica de $(p - 1)$ y $(q - 1)$.

Coeficiente de Pearson:

$$P = \left\{ \frac{\phi^2}{1+\phi^2} \right\}^{1/2} = \left\{ \frac{\chi^2}{n_{..} + \chi^2} \right\}^{1/2} \tag{3.8}$$

conocido de manera extendida en la bibliografía como *coeficiente de contingencia* [Pearson (1966)].

Si se desea tener en cuenta la ordinalidad de al menos una de las dos variables categóricas implicadas en una tabla de contingencia, nos referimos a Agresti (1984) pp. 156-179. Además de toda la conocida variedad de medidas basadas en rangos, como puede ser el coeficiente de correlación de Kendall o el de correlación de Spearman.

Por otro lado, podemos construir, al igual que para (3.4), dos conjuntos de variables binarias que definan las clases de ambas variables cualitativas: $\left[\mathbf{X}_i\right]_{i=1,2,\ldots,p}$ e $\left[\mathbf{Y}_j\right]_{j=1,2,\ldots,q}$, $p+q$ vectores de dimensión $(n\times 1)$, y buscar unos coeficientes a_i, $i=1,2,\ldots,p$, y b_j, $j=1,2,\ldots,q$ que conformen las combinaciones lineales: $\tilde{\mathbf{X}} = \sum_{i=1}^{p} a_i \mathbf{X}_i$, $\tilde{\mathbf{Y}} = \sum_{j=1}^{q} b_i \mathbf{Y}_i$, tales que maximicen $\left|r\left(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right)\right|$ o $r^2\left(\tilde{\mathbf{X}}, \tilde{\mathbf{Y}}\right)$,

$$r = \frac{\sum_{i=1}^{p}\sum_{j=1}^{q} n_{ij}\left(a_i - \bar{a}\right)\left(b_j - \bar{b}\right)}{\sqrt{\left(\sum_{i=1}^{p} n_{i.}\left(a_i - \bar{a}\right)^2\right)\left(\sum_{j=1}^{q} n_{.j}\left(b_j - \bar{b}\right)^2\right)}} \text{ siendo } \bar{a} = \frac{\sum_{i=1}^{p} n_{i.} a_i}{n_{..}}, \quad \bar{b} = \frac{\sum_{j=1}^{q} n_{.j} b_j}{n_{..}}.$$

La solución a este problema viene dada por la primera de las correlaciones canónicas, r_1. Nos referimos a Cuadras (2003) para la solución numérica de la maximización y correspondiente descomposición singular.

Las correlaciones canónicas nos permiten descomponer la asociación en diferentes dimensiones ortogonales: $\chi^2 = n_{..}\sum_{i=1}^{s} r_i^2$ donde $s = \min\{p,q\} - 1$. Por esta igualdad, tenemos que $\phi^2 = \sum_{i=1}^{s} r_i^2$, de este modo, la interpretación desprendida por las correlaciones canónicas positivas respecto al coeficiente C [Cramér (1946)] es la siguiente:

$$C = \left\{ \frac{\sum_{i=1}^{s} r_i^2}{s} \right\}^{1/2} . \tag{3.9}$$

En un problema con N variables, hay $\binom{N}{2} = \frac{1}{2}N(N-1)$ diferentes pares a comparar. Si la medida de asociación de un par es 0.72 y para otro 0.59, significa que el primer par está más asociado que el segundo. Por supuesto cada tipo de medida refleja la asociación en un contexto particular. Cuando disponemos de variables mixtas y analizamos las asociaciones 2 a 2, lo hacemos con el objetivo de realizar comparaciones. Para ello deben ser utilizadas medidas "compatibles", o cuyo significado sea similar. Andenberg (1973) pp. 96-97 propone las siguientes como "compatibles":

- Cuantitativa-Cuantitativa: el coeficiente de correlación, ρ.
- Cuantitativa-Cualitativa: el coeficiente de correlación calculado con puntuaciones optimas asignadas a cada clase, η.
- Cualitativa-Cualitativa: correlación canónica entre filas y columnas utilizando puntuaciones óptimas para los dos conjuntos de clases, r_1.

En las aplicaciones del capítulo 5 calculamos y comparamos las medidas descritas en este apartado.

3.2. Análisis estadístico multivariante

Las medidas de asociación nos permiten conocer la relación variable a variable con la siniestralidad, pero el objetivo es la obtención de un conjunto "equilibrado" de variables de tarifa. Si seleccionamos separadamente las variables que una a una están más asociadas con el riesgo, es posible que en el conjunto de variables seleccionadas resultante dispongamos de información redundante o bien que no tengamos incorporadas variables que de manera conjunta con otras resulten significativas. Por lo que se hace necesario realizar el estudio conjunto teniendo en cuenta a la vez todos los factores potenciales del riesgo e "idealmente" todas sus interacciones. Para ello hacemos uso de técnicas de análisis estadístico multivariante.

Algunas de las técnicas nos serán útiles además para seguir cubriendo las siguientes etapas hasta la formación de los grupos de tarifa exclusivos y exhaustivos, y/o hasta la posterior estimación de la prima pura para cada asegurado. Cada técnica de análisis estadístico multivariante busca unos objetivos, por ello se basa en un cierto tipo de datos y tiene implícitas una serie de hipótesis, y su aplicación implica un mayor o menor esfuerzo computacional. Podemos clasificar las técnicas que utilizaremos en el trabajo de diferentes formas, por ejemplo: según su filosofía o según las fases de tarificación que nos permitan cubrir.

Según la filosofía las dividimos en tres bloques: técnicas de predicción mediante modelos de regresión, técnicas de agrupación de individuos mediante análisis cluster y técnicas de clasificación de individuos mediante análisis discriminante.

Las técnicas de regresión consisten en la estimación de la respuesta a partir de una serie de variables explicativas o predictores. Disponemos de una gran variedad de modelos a los que adaptar nuestro tipo de datos. Su aplicación a la selección de variables de tarifa es la siguiente: dado un modelo concreto, acorde con nuestros datos, buscaremos mediante un proceso de selección de predictores la "mejor" combinación de ellos para la estimación del riesgo, y éstos pasarán a ser el conjunto de variables de tarifa. Con estas técnicas podemos, si nos interesa, realizar ya una estimación acurada de las primas puras. En este capítulo veremos con detalle los modelos lineales generalizados por ser la técnica más actual utilizada en la bibliografía actuarial para la selección en tarificación *a priori*. También estudiaremos en el capítulo siguiente el modelo de regresión basada en distancias como propuesta de herramienta alternativa para cubrir esta fase.

Los métodos de análisis cluster sirven para la formación de grupos homogéneos de individuos. Los métodos pueden ser [Andenberg (1973); Carrasco y Hernán (1993); Hartigan (1975); Hawkins y Kass (1982a); Cuadras (1996); Lebart, Morineau y Fénelon (1985); Sierra (1986)]: jerárquicos / no jerárquicos; aglomerativos / divisivos; monotéticos / politéticos. En nuestro estudio destacamos únicamente la aplicabilidad de los métodos jerárquicos aglomerativos y divisivos politéticos.

- *Jerárquicos*:

  Aglomerativos: empiezan con las clases básicas (indiviuo a individuo) y las van fusionando para formar subclusters. El punto común de partida lo constituye una matriz de distancias entre individuos que se calcula a partir de la experiencia de siniestralidad (univariante o multivariante) objeto de estudio. Éstos son de utilidad exclusivamente para la formación de clases de tarifa sólo si cada individuo representa una clase de tarifa inicial, por ejemplo, si queremos agrupar por zonas, cada individuo deberá ser una provincia. Así su aplicabilidad es limitada y no es adecuado para la selección de variables. Dentro de los jerárquicos aglomerativos también clasificamos al método de Ward [Byron, Morgan y Ray (1995); Campbel (1986); Ward (1963); Ward y Hook (1963)]. Éste se separa del resto porque no se basa en una función de distancias, sino en la minimización de la varianza dentro de los grupos que va formando. Nosotros lo recomendamos exclusivamente para la discretización de variables continuas.

  Divisivos: empiezan con el conjunto completo de individuos que forman un solo cluster y van particionando sucesivamente en clases más finas. Dentro de ellos englobamos principalmente las técnicas politéticas de segmentación que veremos con detalle en este capítulo, pues han sido utilizadas en el año 1995 por UNESPA [Sánchez (1997); UNESPA (1995)] en la segmentación de la cartera de RC de automóviles en España de la estadística común. Éstas son un caso especial de análisis cluster que también puede ser considerado como técnica de regresión, pues necesita de una variable respuesta y de un conjunto de predictores. El resultado final del análisis de segmentación son unos segmentos terminales que nos resumen los grupos de tarifa a partir de una clasificación no cruzada y que tiene en cuenta, aunque de un modo jerárquico, el efecto de interacción entre predictores.

- *No jerárquicos*: son métodos que optimizan un funcional objetivo fijado un número de clusters. Los que como objetivo tienen alguna variedad de minimización de la varianza dentro de los

grupos a formar, pueden ser utilizados unidimensionalmente, al igual que el método de cluster jerárquico de Ward, en la discretización de variables continuas.

El análisis discriminante clasifica a los individuos en dos o más poblaciones previamente establecidas según los valores de siniestralidad y posteriormente con un proceso de selección de predictores escogemos aquellos que "mejor" discriminan a las poblaciones [McLachlan (1992)]. Es usual distinguir dos poblaciones, la que no conlleva riesgo y la que conlleva riesgo (extrapolable al caso de más de dos poblaciones). Y lo usual es basarse en la experiencia del número de siniestros, así la población sin riesgo será la que no tenga siniestros y la de riesgo la que tenga al menos un siniestro. Al seleccionar las variables que mejor discriminan las poblaciones lo que cubrimos es el paso de selección de variables de tarifa. Por lo tanto el análisis discriminante no es una técnica predictiva en el sentido de la tarificación, aunque si nos permite clasificar a un nuevo individuo en una población concreta según los valores que tome en las variables de tarifa. En cualquier caso deberemos asignarle posteriormente de algún modo la prima pura.

Según las fases del proceso de tarificación *a priori* que nos permiten cubrir, podemos dividir las técnicas en predictivas, que en principio cubrirán todo el proceso de tarificación, y no predictivas, que cubrirán tan sólo alguna fase:

Técnicas predictivas:

- *Modelos de regresión*: destacamos especialmente el modelo lineal generalizado y la regeresión basada en distancias, la cual estudiamos con detalle en el capítulo 4. Ambos modelos nos van a permitir cubrir cada una de las fases del proceso de tarificación hasta la estimación de las primas puras. Incluimos aquí los modelos de credibilidad basados en técnicas de regresión que nos permiten realizar una estimación de la siniestralidad a partir de unos grupos homogéneos de riesgo. Aunque cabe notar que precisan ser combinados previamente con alguna metodología de selección de variables que nos indique cuales serán los grupos de tarifa homogéneos iniciales.

- *Técnicas de segmentación*: por ejemplo, CHi-squared Automatic Interaction Detector (CHAID), EXtended Automatic Interaction Detector (XAID), THeta Automatic Interaction Detector (THAID), etc. Éstas cubren todas las fases, aunque su predicción está limitada a las clases ya existentes de los factores categóricos seleccionados y la única opción para la estimación de la prima pura es alguna media de los grupos terminales.

Técnicas no predictivas:

- *Análisis cluster jerárquico aglomerativo*: únicamente nos permite cubrir la formación de clases de tarifa factor a factor de un modo algo crítico, pues cada individuo debe representar una unidad lógica en la agrupación. *Análisis cluster no jerárquico*: lo utilizaremos unidimensionalmente para la discretización de variables continuas una a una, junto con el método jerárquico de Ward (nos referimos al apartado 3.6. para el detalle).
- *Análisis discriminante*: nos es útil únicamente para la selección de variables y si es el caso para la formación de los grupos de tarifa que mejor discriminen las poblaciones.

3.3. Metodologías en la bibliografía actuarial

Existe una considerable literatura que discute, en términos generales, la filosofía a seguir por el actuario en el caso de selección de variables de tarifa en un proceso de tarificación *a priori* no vida. Un *survey* clásico que incluye algunas metodologías de selección es van Eeghen, Greup y Nijssen (1983).

A modo de cita, tenemos:

Regresión:

- Modelos lineales generalizados: Agsaa (1977); Albrecht (1983a,b); Andrade y Silva (1989); Boj, Claramunt y Fortiana (2001); Boj, Claramunt, Fortiana y Vidiella (2002); Brockman y Wright (1992); Hipp (2000); López y López de la Manzanara (1996, pp. 137-139); Stroinski (1986); Stroinski y Currie (1989); Zehnwirth (1994)
    - Caso particular para predictores de tipo continuo mediante regresión lineal ordinaria: Lemaire (1977,1979,1985)
    - Caso particular para predictores de tipo categórico mediante regresión lineal ordinaria: Hallin (1977)
- Modelos de credibilidad: Cabral y García (1977); Cohen, Durpin y Levy (1986); Bühlmann (1967,1974,1999); Nelder y Verrall (1997); Sundt (1987)

Discriminante: Beuthe y van Namen (1975); Masure (1978); Prokkola y Romppainen (1992a,b)

Cluster:

- Cluster: Byron, Morgan y Ray (1995); Campbel (1986); Loimaranta, Jacobson y Lonka (1980)
- Segmentación: Calatayud y Martínez (1997); Hawkins (1997); UNESPA (1995); Sánchez (1997); Pérez (2001); Boj, Claramunt y Fortiana (2001)

Otros (incluyendo variedades de los anteriores, como redes neuronales, *fuzzy cluster* o modelos de estadística *Bayesiana*): Beirlant, Derveaux, de Meyer, Goovaerts, Labie y Maenhoudt (1991); Boskow y Verrall (1994); Conger (1987); Derrig y Ostaszewski (1995); Grünig (1975); Hallin (1977); Hallin y Ingenbleek (1981); Hooge (1974); Hutchinson y Rowell (1986); Jewell (1975); Picech y Pessoni (1998); Pitkänen (1975); Ramachandran (1975); Reid (1975); Schmiter y Straub (1975); Taylor (1989).

3.4. Análisis de segmentación

El *Análisis de Segmentación* (*AS*) es una técnica estadística de cluster jerárquico divisivo que trabaja sobre datos tipo regresión. Las variables independientes son categóricas, de tipo nominal u ordinal, y la variable dependiente puede ser cuantitativa o categórica. Se utiliza con fines exploratorios y descriptivos, con el objetivo básico de encontrar una clasificación de la población en grupos capaces de describir la variable dependiente de la mejor manera posible. En el caso de variables de tipo cualitativo, los análisis estadísticos usuales se limitan a producir y examinar todas las tabulaciones cruzadas que se consideran de interés, lo que resulta limitado, y en muchas ocasiones sólo sirve para identificar relaciones que *a priori* ya resultaban evidentes. El AS reduce la complejidad del problema, rechazando tabulaciones cruzadas no significativas, detectando automáticamente los mejores predictores y creando subgrupos potencialmente explicativos de la variable dependiente. Permite conocer qué variables son útiles para describir la variable dependiente, qué categorías de un predictor son homogéneas respecto a la variable dependiente y cual es el efecto conjunto de dos o más predictores.

Puede ser utilizado para la predicción directamente a partir del árbol resultante. Y adicionalmente es útil como paso previo en la aplicación de otras técnicas especializadas para datos cualitativos como el

análisis de correspondencias. En la aplicación 1 del capítulo 5 del trabajo lo utilizamos para formar los grupos homogéneos de que parte un modelo de credibilidad en la estimación de primas. Del mismo modo lo podríamos utilizar para formar los grupos de tarifa que entrarían a modo de predictores en un modelo de regresión, sin tener que realizar entonces un proceso de selección de variables.

3.4.1. Técnicas de Detección Automática de la Interacción

En muchas investigaciones en las que interviene una variable respuesta, el efecto sobre ésta, de una variable explicativa, depende del nivel de otra u otras variables explicativas; es decir, de si existe interacción entre ellas. La presencia de interacción exige un tratamiento adecuado, por eso resulta fundamental la detección de este fenómeno. Las técnicas de segmentación que utilizamos en el trabajo pertenecen a la familia de métodos estadísticos denominada AID (Automatic Interaction Detection). Todas las técnicas de AID, cuyo objetivo básico es detectar la existencia de interacción en un modelo de predicción, operan de un modo secuencial y tratan de dividir el conjunto de individuos objeto de estudio en grupos homogéneos (segmentos) mutuamente excluyentes y exhaustivos, en los cuales se pueda describir la relación entre los predictores y la variable respuesta sin que ésta sea enmascarada por efecto de la interacción. Podemos encontrar históricamente diferentes tipos de AID, dependiendo de la naturaleza de la variable dependiente:

Si la variable dependiente es de tipo CUANTITATIVO:

- ALGORITMO DE MORGAN Y SONQUIST: Este algoritmo considera una variable dependiente de tipo cuantitativo y varias variables independientes dicotómicas, y propone divisiones binarias sucesivas utilizando como criterio la reducción en la suma de cuadrados no explicada, y lo que hace es que de todas las divisiones binarias factibles del grupo de observaciones, tomando como base cada predictor binario, debe hallarse aquélla que produzca la mayor reducción en la suma de cuadrados residual, y ésta será la particionada en dos subgrupos. El criterio de parada que utiliza es el siguiente: -Particionar si la reducción obtenida es de al menos el 1% de la suma de cuadrados residual total, sino se busca un próximo grupo para ser subdividido, -Y de entre los grupos formados seleccionar aquél para el cual la suma de cuadrados residual sea máxima, es decir, el más heterogéneo, siempre que ésta sea mayor que el 2% de la suma de cuadrados residual. El proceso se detiene sólo si ningún grupo explica más del 2% de la suma de cuadrados residual total.

- ALGORITMO XAID (eXtended AID): Es una extensión del anterior, en el cual se utiliza el estadístico F del análisis de la varianza. De este algoritmo existe una versión no paramétrica que utiliza el estadístico de Kruskal-Wallis [Worsley (1977)].

Si la variable dependiente es de tipo CUALITATIVO:

- ALGORITMO THAID (THeta AID): Esta técnica fue propuesta por Messenger y Mandell (1972), y descrita en detalle por Morgan y Messenger (1973). Es un algoritmo para variable dependiente cualitativa que produce segmentaciones binarias, utilizando como criterio, maximizar el número de observaciones en cada categoría modal. La idea consiste en calcular la probabilidad de éxito al predecir el valor de **Y** por medio de la categoría modal correspondiente, dado el predictor $\mathbf{F}_p$.

- ALGORITMO CHAID (CHi-square AID): Este algoritmo fue propuesto inicialmente por Kass (1980). Supone una variable dependiente de tipo cualitativo y unas variables independientes de tipo también cualitativo. Utiliza el test *ji*-cuadrado para contrastar independencia en las distintas fases del proceso. Es considerado como un ALGORITMO GENERAL DE SEGMENTACIÓN [Escobar (1992)], y por ello todos los anteriores pueden ser considerados como casos particulares de éste. Lo analizamos con detalle en los siguientes sub-apartados.

3.4.1.1. Algoritmo general de segmentación

El análisis de segmentación particiona unos datos sucesivamente en grupos cada vez más pequeños, basados en los valores de las diferentes categorías de los predictores. Inicialmente divide la población en dos o más grupos diferentes basados en las categorías del "mejor" predictor, para después dividir cada uno de estos grupos en grupos más pequeños basados en otras variables predictoras seleccionadas en su caso como las "mejores" para los correspondientes nodos. El proceso sigue hasta que no quedan predictores significativos. El resultado son los grupos finales -segmentos terminales- reflejados en un diagrama de árbol o dendograma fácil de interpretar, donde los casos pueden clasificarse fácilmente en el segmento apropiado tan sólo conociendo las categorías que le corresponden en las variables predictoras que discriminan los segmentos. El dendograma puede ser utilizado para realizar predicciones y/o para entender la importancia de las interacciones entre los diferentes predictores. Así, se trata de un proceso iterativo que puede representarse en un diagrama de flujo [Ramírez (1995) pp. 23] que podemos observar en el anexo 3.1.

Supongamos que tanto la variable respuesta, **Y**, como los predictores, $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_p$, son variables cualitativas:

Fase de agrupación de categorías: En esta fase se agrupan las categorías sin influencia significativa en el patrón de respuesta de la variable dependiente; es decir, se agrupan las categorías de los predictores cuando éstos tienen un perfil similar en la variable dependiente. Para ello se utiliza el estadístico *ji*-cuadrado; se agrupan todas las categorías para las cuales el valor del test *ji*-cuadrado resultante de cruzar la variable respuesta con las categorías elegidas del predictor en estudio, sea no significativo. El proceso se repite con las nuevas categorías agrupadas para analizar si proceden, o no, nuevas fusiones. El proceso termina cuando todas las categorías son significativamente diferentes, o cuando se han agrupado todas las categorías, en cuyo caso el predictor se elimina. Los pares posibles de categorías de un predictor que intervendrán en la fase de agrupación dependen de la naturaleza del predictor. Pueden considerarse los siguientes tipos de predictores:

- Predictor libre: será aquel del cual podremos combinar todas las categorías con todas en la fase de agrupación óptima de categorías de un predictor. Se tratará, en general, de predictores de naturaleza categórica nominal.

- Predictor monótono: será aquel del cual sólo podremos agrupar categorías contiguas. Se tratará, en general de predictores categóricos con un orden en sus escalas (categóricos ordinales), o bien de los predictores continuos discretizados en clases, que por lo tanto también poseerán un orden natural.

  El término -monótono- induce a pensar que el predictor crece o decrece monótonamente con la variable respuesta, pero ésto, aunque es usual, no es necesario para definir a un predictor como tal. Si se tiene un predictor en la escala ordinal, es bueno definirlo primero como libre y observar si la agrupación *a posteriori* ha sido razonablemente monótona y finalmente decidir si lo definimos como monótono o libre.

- Predictor flotante: será una variante del predictor monótono, tal que su escala es monótona excepto para una sola clase flotante cuya posición en la escala monótona es desconocida. Así que esta clase puede ser agrupada con cualquier otra de la escala. Usualmente la clase flotante es la que contendrá los valores *missing*.

Respecto a los predictores flotantes, en el caso de predictores ordinales, se contempla el caso en que los valores *missing* de la clase flotante sean *missing* aleatoriamente o el caso en que formen una clase capaz de ser informativa. Si el predictor es missing aleatoriamente, entonces la categoría flotante tiende a ser agrupada con otras categorías cercanas al medio de la parte ordinal dc la escala. Pero si la clase flotante es predictiva, entonces la categoría flotante puede quedar sola o agrupada en alguno de los extremos de dicha escala. Respecto a la información *missing* en los predictores libres, no es necesaria ninguna consideración, lo único que hay que hacer es crear una clase extra que podrá ser agrupada o no con cualquier otra.

Como curiosidad, nos referimos a capítulo 7 pp. 213 de Ramírez (1995): "CHAID y el análisis de coordenadas principales" donde se muestra como es posible utilizar el análisis en coordenadas principales para obtener una representación gráfica en términos de distancias para solventar alternativamente la fase de agrupación en el caso de predictores libres.

Fase de selección del mejor predictor: Una vez seleccionadas las categorías portadoras de información nos encontramos con varios predictores potencialmente explicativos cuyas categorías son significativamente diferentes respecto a la variable dependiente. Se busca el mejor predictor que será aquel para el cual obtengamos una mayor asociación con la variable respuesta; es decir, aquel para el cual obtengamos un p-valor más pequeño o el mayor valor para el coeficiente de asociación elegido. El mejor predictor para segmentar el grupo es el que mejor discrimina a los sujetos según la variable dependiente.

Fase de segmentación: Si el mejor predictor es significativo a un nivel previamente establecido, se realizará la segmentación de la población considerando tantos segmentos como categorías tenga el predictor elegido. Para cada segmento se repite el proceso y se realizan nuevas segmentaciones hasta que no haya predictores significativos en ninguno de los grupos. Si no aparecen predictores significativos, el grupo se considera como grupo terminal. El proceso de segmentación termina cuando todos los grupos son terminales.

Finalización del proceso: Si no se pusieran otras limitaciones al proceso, éste terminaría cuando no hubiesen predictores significativos en ninguno de los grupos. En ese caso, probablemente el estadístico *ji*-cuadrado se obtendría a partir de tablas poco ocupadas, con la problemática que esto

conlleva[14]. De esta manera nos podríamos encontrar con una gran cantidad de grupos terminales, de tamaño muy pequeño, los cuales resultarían difíciles de interpretar. Es conveniente por lo tanto, limitar el proceso de segmentación mediante la introducción de ciertos controles, los cuales son conocidos como "*Filtros del proceso*":

Significación de categorías: Es el nivel de significación utilizado en la fase de agrupación de categorías para ver si dos categorías tienen un perfil similar, es decir, si no son significativamente diferentes, se compara su significación.

Significación de un predictor: Es el nivel de significación utilizado en la fase de selección del mejor predictor para verificar si un predictor es significativo.

Filtros de asociación: Se esteblece una asociación mínima entre la variable dependiente y el predictor para considerarlo como potencial candidato para realizar la segmentación. Los filtros de asociación pueden utilizarse solos o en combinación con los filtros de significación de predictores, siendo esta última, más aconsejable. Una buena estrategia a seguir en la fase de selección sería escoger aquél con mayor significación (menor valor de p), y descartarlo si el coeficiente de asociación no es superior a un valor mínimo establecido por el usuario, el cual por supuesto, depende del coeficiente elegido. Una ventaja que tienen los filtros de asociación sobre los de significación es que resultan menos sensibles al número total de individuos.

Tamaño Antes: Se establece un tamaño mínimo para que un grupo pueda segmentarse. Esto quiere decir que si un grupo tiene menor número de individuos que el prefijado por el "tamaño antes", no se segmenta y se declara como terminal.

Tamaño Después: Se establece un tamaño mínimo para que un subgrupo sea formado. Por lo tanto, si alguno de los grupos formados en la segmentación de un grupo tiene menos número de individuos que el "tamaño después", la segmentación es descartada.

Los dos filtros de tamaño tienen como objetivo evitar que se formen grupos pequeños o grupos no balanceados. No existen unos valores prefijados de estos filtros, pues dependerán del número total de

[14] "El test *ji*-cuadrado es apropiado sólo cuando ninguna frecuencia esperada es menor que 5"
"Si menos del 20% de las casillas de una tabla tienen frecuencias esperadas menores que 5, el test *ji*-cuadrado es válido siempre que la frecuencia esperada mínima no sea inferior a 1"

individuos. Una posibilidad es, por ejemplo, fijar el "tamaño antes" como un porcentaje del número de individuos, y el "tamaño después" como un porcentaje del "tamaño antes".

Filtro de Nivel: Se establece un número máximo de niveles de segmentación. Una segmentación con un solo nivel resulta útil pero demasiado simple. Por otro lado, una segmentación con muchos niveles puede resultar compleja y difícil de interpretar.

3.4.1.2. El contraste de independencia

En CHAID el contraste de independencia *ji*-cuadrado entre dos variables categóricas con I y J categorías respectivamente, con frecuencias observadas en cada celda de la tabla de contingencia f_{ij}, y con frecuencias esperadas bajo el supuesto de independencia $\hat{f}_{ij}$,

$$\chi^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(f_{ij}-\hat{f}_{ij}\right)}{\hat{f}_{ij}} \sim \chi^2_{(I-1)(J-1)} \tag{3.10}$$

se utiliza para:

i) Verificar si dos categorías de un predictor difieren significativamente con respecto a la variable dependiente, y para

ii) Seleccionar el mejor predictor en un grupo, y determinar si tiene poder discriminativo suficiente como para realizar una segmentación.

Directamente se trabaja con el valor de probabilidad "*p*-valor" desprendido del contraste. El p-valor es la probabilidad de que la relación observada entre un predictor concreto y la variable dependiente ocurriese si el predictor y la variable dependiente fuesen estadísticamente independientes. Por ejemplo, si *p*-valor = 5% (0.05), significará que la relación observada entre el predictor y la variable dependiente ocurrirá nada más el 5% de la veces si las variables fuesen independientes. Diremos que el *p*-valor será estadísticamente significativo si el *p*-valor $\leq \alpha$ (= nivel de significación predeterminado). En la fase de selección del mejor predictor, el "mejor" predictor será aquel que tenga el menor *p*-valor y que sea significativo. En la fase de agrupación de categorías de un predictor, se agruparán dos categorías, si al mirar la relación de éstas con la variable dependiente, (formaremos una tabla cruzada de la variable dependiente con las dos categorías del predictor en cuestión), salen no

significativas, y empezaremos agrupando aquel par de categorías que nos dé el menor *p*-valor no significativo. Es decir, el mejor predictor será el de más dependencia con la variable dependiente, y la mejor pareja de clases de un predictor concreto para agrupar, será aquella que sea menos dependiente con la variable dependiente, que nos vendrá a decir que los valores de la variable explicativa no tienen relación de dependencia con las dos clases separadas y por lo tanto podemos proceder a su agrupación.

Así, de manera genérica, contrastaremos:

$$H_0 : \text{Independencia} \quad \text{versus} \quad H_1 : \text{Dependencia}$$

p-valor significativo → Rechazamos H_0

p-valor no significativo → Aceptamos H_0 (Rechazamos H_1)

Ajuste de Bonferroni

Cuando en un predictor se ha realizado una combinación de categorías, debe realizarse un ajuste en el cálculo de la significación para la fase de selección del mejor predictor. Sus *d* categorías finales han sido obtenidas después de un proceso de agrupación de las *c* categorías iniciales, y ello afecta a la significación del predictor. El *ajuste de Bonferroni* propuesto por Kass (1980) constituye una aproximación a la significación real. Tal ajuste consiste en multiplicar el *p*-valor del estadístico resultante del contraste de independencia entre la variable dependiente y las categorías finales del predictor en estudio, por una cantidad *C* igual al total de agrupaciones posibles de las *c* categorías iniciales en *d* categorías fusionadas. El valor de la constante *C* varía según el tipo de predictor y por tanto según el procedimiento en la fase de agrupación:

Opción monótona: $C = \binom{c-1}{d-1}$

Opción flotante: $C = \binom{c-2}{d-2} + d \cdot \binom{c-2}{d-1}$

Opción libre: $C = \sum_{i=0}^{d-1} (-1)^i \frac{(d-1)^c}{i!(d-i)!}$

La aproximación de Bonferroni a la significación real resulta en muchas ocasiones conservadora, es decir, demasiado exigente para el rechazo de la hipótesis nula. Como consecuencia se declaran independencias falsas y se descartan segmentaciones que deberían producirse. Por el contrario si no se

realiza ningún ajuste en el *p*-valor, se ocasiona el efecto contrario, se rechazan independencias verdaderas produciendo segmentaciones erróneas. En Ramírez (1995) se discute en detalle este problema y se propone un procedimiento alternativo que consiste en una modificación del citado ajuste, y al cual llama *Bonferroni modificado*. Los estudios de simulación que lleva a cabo ponen de manifiesto que el ajuste de Bonferroni no es tan conservador como parece. Los dos ajustes, el clásico y el modificado producen en general resultados similares: si el predictor es monótono o flotante, los resultados son casi idénticos; si el predictor es libre y con pocas categorías (entre 3 y 5) el procedimiento clásico parece comportarse mejor, si el número de categorías está entre 6 y 8, el clásico parece hacerse más conservador que el modificado, y para más de 8 categorías ninguno de los procedimientos parece controlar adecuadamente el rechazo de independencias verdaderas, produciendo segmentaciones erróneas.

Paradoja de Simpson

Cuando se trabaja con 3 o más variables y se dispone, por tanto, de una tabla trifactorial o multifactorial, no es posible analizar la información examinando cada una de las tablas simples bifactoriales. La información resultante de los análisis parciales puede ser contradictoria, dos variables pueden ser independientes marginalmente y no serlo en presencia de otras variables, a las que llamaremos variables de confusión. Este hecho, es conocido en la literatura como la *paradoja de Simpson* [Simpson (1951)]. El CHAID presupone que es posible colapsar[15] en todas las variables y no siempre es así. El CHAID está basado en contrastes de asociación marginales, pero no lleva implícito ningún paso que garantice que tiene sentido la colapsabilidad en que se basa. En Ávila (1996) se demuestra que el algoritmo CHAID no es capaz de detectar la paradoja de Simpson, produciendo resultados erróneos en presencia de la misma. Se da la incongruencia de que un procedimiento de detección automática de la interacción, no la detecta precisamente porque existe. Demuestra que el AS sólo es válido cuando no existe relación entre las variables explicativas. Propone un algoritmo basado en contrastes de independencia condicionada y en las condiciones de colapsabilidad, que es válido también en el caso de que las regresoras estén relacionadas, lo cual supone un importante avance en la práctica. Nos remitimos a Dorado (1998) donde encontramos un estudio en profundidad de las condiciones en que tiene sentido la colapsabilidad en que se basa el CHAID.

[15] Colapsar en una variable: estudiar por separado las tablas marginales resultantes.

3.4.2. Software utilizado

En el trabajo hacemos uso del módulo CHAID del programa SPSS [Magidson (1993a)], y del CHAID y XAID que encontramos programados en Fortran 77 en http://www.stat.umn.edu/users/FIRM/index.html bajo el nombre de FIRM (*Formal Inference-based Recursive Modeling*) [Hawkins (1997)]. Existen bastantes alternativas en programación de algoritmos de segmentación, pero lo usual es que los paquetes estadísticos no lo lleven implementado por defecto; si lo incorporan, son módulos separados, como por ejemplo el SPSS CHAID. La mayoría de programas están realizados por los propios autores que han profundizado en estas técnicas. A modo de ejemplo, en la Universidad de Salamanca encontramos tres tesis doctorales [Ramírez (1995); Ávila (1996); Dorado (1998)] sobre AS, en las que se implementa CHAID con las modificaciones académicas pertinentes. Otra técnica de segmentación, que nosostros no utilizamos en el trabajo, es la denominada CART (*Classification And Regression Trees*) propuesta inicialmente por Breiman, Friedman, Olshen y Stone (1993). Ésta es bastante utilizada en el ámbito de segmentación de mercados y márketing junto con CHAID [Haughton y Oulabi (1983); Thrasher (1991)], pero por el momento no ha sido aplicada en tarificación. CART sí suele estar incorporado en paquetes estadísticos estándar como pueden ser S-Plus y Statistica.

En el trabajo distinguimos el tipo de algoritmo dependiendo de dos factores: la naturaleza de la variable dependiente y la manera de calcular los p-valores de la fase de agrupación de categorías y de selección del mejor del predictor:

1) **SPSS CHAID nominal**: Realiza un CHAID para respuesta categórica nominal. Respecto a los p-valores las opciones son las siguientes:

- En el cálculo de los p-valores, podemos optar por utilizar la χ^2 estándar de Pearson o *Likelihood Ratio*. Primeramente vamos a describir el modo de calcular las frecuencias esperadas: Se basa en la teoría de los modelos logarítmico-lineales[16], que de hecho son un caso particular del modelo lineal generalizado que veremos en detalle en este capítulo [Agresti (1984) pp. 244]. Dada una tabla de contingencia con frecuencias observadas f_{ij}, sea A la fila del predictor con I categorías, y sea B la columna de la variable dependiente con J categorías, entonces:

$$H_1 : \ln\left(\frac{f_{ij}}{z_{ij}}\right) = \lambda + \lambda_i^A + \lambda_j^B + \lambda_{ij}^{AB} \quad \text{con} \quad \sum_{i=1}^{I}\lambda_i^A = \sum_{j=1}^{J}\lambda_j^B = \sum_{i=1}^{I}\lambda_{ij}^{AB} = \sum_{j=1}^{J}\lambda_{ij}^{AB} = 0$$

$$H_0 : \lambda_{ij}^{AB} = 0 \quad \text{para} \quad \begin{matrix} i = 1,2,\ldots,I \\ j = 1,2,\ldots,J \end{matrix}$$

donde $z_{ij} = \frac{1}{w_{ij}}$, y w_{ij} es un peso especificado. En el caso de no especificar ponderación $w_{ij} = 1$, (cabe notar que no debemos confundir la ponderación con la frecuencia que nos servirá para reducir el volumen de datos, en especial si las combinamos).

Para calcular el *p*-valor no ajustado, se necesitan calcular las frecuencias esperadas $\hat{f}_{ij}$ y para ello es necesario estimar los parámetros del modelo completo, con efectos principales e interacciones. Una vez el algoritmo de cálculo[17] ha convergido, ya se puede calcular el estadístico para ver la bondad del ajuste de las frecuencias esperadas. Podemos elegir entre la χ^2 estándar de Pearson (3.10):

$$\chi^2(H_0) = \sum_{i=1}^{I}\sum_{j=1}^{J}\frac{\left(f_{ij} - \hat{f}_{ij}\right)^2}{\hat{f}_{ij}} \sim \chi^2_{(I-1)(J-1)}$$

o el *Likelihood Ratio*:

$$L^2(H_0) = 2\sum_{i=1}^{I}\sum_{j=1}^{J} f_{ij} \ln\left(\frac{f_{ij}}{\hat{f}_{ij}}\right) \sim \chi^2_{(I-1)(J-1)} \qquad (3.11)$$

- Podemos optar por el ajuste de Bonferroni en la fase de selección del mejor predictor.

2) **SPSS CHAID ordinal**: Realiza un CHAID para respuesta categórica ordinal. En este caso se supone que las categorías de la variable dependiente están ordenadas y lo tiene en cuenta.

- Para su tratamiento, también hace uso de los modelos logarítmico-lineales pero a través de los modelos *linear by linear* [Agresti (1984) pp. 79] que contemplan el hecho de que las dos variables analizadas (fila y columna) puedan estar ordenadas [Goodman (1979)].

[16] Los modelos logarítmico-lineales aplican un modelo lineal a los logaritmos de las frecuencias en cada nivel de combinación de variables [Agresti (1984,1990), Christensen (1990) para una descripción detallada].
[17] Generalmente, por procedimientos máximo-verosímiles.

Concretamente hace uso del modelo de efectos fila (que supone que la variable ordinal es la columna –la respuesta– y que la fila es nominal –el factor–). Respecto a los *p*-valores, utiliza el contraste **Y**–*association* [Magidson (1992,1993b)] con el correspondiente *Likelihood Ratio*. Veamos en qué consiste el test **Y**–*association*:

$$H_1 : \ln\left(\frac{f_{ij}}{z_{ij}}\right) = \lambda + \lambda_i^A + \lambda_j^B + x_i\left(y_j - \bar{y}\right) \quad \text{con} \quad \sum_{i=1}^{I}\lambda_i^A = \sum_{j=1}^{J}\lambda_j^B = \bar{x} = 0$$

donde

y_j es la puntuación asociada a la *j*-ésima categoría de *B* (la variable dependiente)

x_i es el coeficiente desconocido a estimar para y_j

$\bar{y}$ es la media de las puntuaciones de la variable dependiente

Entonces, la hipótesis nula de independencia es:

$$H_0 : x_1 = x_2 = \cdots = x_J$$

La estimación bajo el método ordinal comporta dos pasos: el primero es estimar por máxima verosimilitud las frecuencias esperadas bajo H_1. Estas estimaciones preservan la **Y**-asociación de la tabla inicial, pero eliminan cualquier otro tipo de dependencia. Posteriormente, para contrastar la independencia, se calcula el *Likelihood Ratio* siguiente:

$$L^2\left(H_0\right) = 2\sum_i\sum_j f'_{ij} \ln\left(\frac{f'_{ij}}{\hat{f}_{ij}}\right) \sim \chi^2_{(I-1)} \tag{3.12}$$

donde

f'_{ij} son las frecuencias estimadas bajo H_1, y $\hat{f}_{ij}$ son las frecuencias esperadas bajo independencia H_0. Como vemos, en este modelo los grados de libertad dependen del número de puntuaciones pre-especificadas de **Y**, pues sólo requiere de un parámetro para describir la asociación.

- Podemos optar por el ajuste de Bonferroni en la fase de selección del mejor predictor.

3) **FIRM nominal (CATFIRM)**: Realiza un CHAID para respuesta categórica nominal. Respecto al *p*-valor, podemos elegir en el programa lo siguiente:

- χ^2 de Pearson con posibilidad de modificación para intentar solucionar el efecto de los ceros de las tablas dispersas: pregunta por el valor de la constante A a añadir al denominador del estadístico χ^2. Si ponemos 0, nos dará la χ^2 de Pearson estándar. Si le damos un valor diferente de 0 calculará:

$$\frac{(observadas - esperadas)^2}{esperadas + A} \tag{3.13}$$

de este modo se reducirá la significación. Esto implicará que estará menos inclinado a realizar particiones con un número pequeño de casos. Si hay alguna justificación teórica para esta modificación la ignoramos, pero en la práctica para tablas dispersas desalenta la formación de agrupaciones en las cuales las frecuencias esperadas sean pequeñas. El valor de A ha de ser pequeño (por ejemplo entre 0.5 y 1), sino podemos tener serias distorsiones en la significación estadística de los valores χ^2.

- Además, se puede elegir entre una aproximación asintótica a la distribución *ji*-cuadrado para el cálculo del *p*-valor o una distribución exacta [Mielke y Berry (1985)]. Al utilizar la distribución exacta, que es recomendable, se consiguen *p*-valores más seguros y formales para tablas con frecuencias pequeñas, puesto que la distribución asintótica de estadístico χ^2 de Pearson, se deteriora cuando las frecuencias en la tabla de contingencia son pequeñas. Los procedimientos de segmentación no pueden utilizarse en conjuntos pequeños de datos, pues a medida que se va segmentando no sólo hay una pérdida de potencia del test, sino que la distribución del estadístico, que es válida sólo asintóticamente, se hace cada vez más imprecisa.

- Hawkins y Kass (1982b) pp. 282-285, proponen dos ajustes para el *p*-valor de la fase de selección del mejor predictor: el de Bonferroni, ya explicado, y la "comparación múltiple" [de los cuales también encontramos detalle en Cuadras (1991) pp. 320]. Ambos ajustes proporcionan cotas conservativas para el *p*-valor, por lo que el programa escoge el mínimo valor de ambos como representación de la significación del predictor en estudio.

4) **FIRM continua (CONFIRM)**: Realiza un XAID para variable respuesta cuantitativa, esta opción **no** es la misma que la del SPSSCHAID ordinal con el test **Y**-association.

- A diferencia del CHAID, el XAID se basa en las F de Fisher y las t de Student procedentes del análisis de la varianza a la hora de calcular los p-valores. En la F para la selección del mejor predictor y en la t para la agrupación de categorías de un predictor. Recordemos la descomposición del ANOVA para una respuesta **Y** continua y un predictor **X** categórico con k niveles (3.3). De ella se desprende la siguiente tabla:

Fuentes de variación	Suma de cuadrados	Grados de libertad	Varianzas
Entre grupos (VE)	$\sum_{i=1}^{k} n_i \left(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot}\right)^2$	$q-1$	$\hat{s}_e^2 = \dfrac{VE}{q-1}$
Interna, no explicada o residual (VNE)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{i\cdot}\right)^2$	$n-q$	$\hat{s}_R^2 = \dfrac{VNE}{n-q}$
TOTAL (VT)	$\sum_{i=1}^{k}\sum_{j=1}^{n_i} \left(y_{ij} - \overline{y}_{\cdot\cdot}\right)^2$	$n-1$	$\hat{s}_y^2$

Tabla 3.1. Tabla ANOVA.

Si estamos en la fase de agrupación de categorías,

$H_0 : \mu_i = \mu_j$,

$$t = \frac{\overline{y}_{i\cdot} - \overline{y}_{j\cdot} - \left(\mu_i - \mu_j\right)}{\hat{s}_R \sqrt{\dfrac{1}{n_i} + \dfrac{1}{n_j}}} \sim t_{(n-k)} \tag{3.14}$$

Observamos que el número de grados de libertad es $n-k$ en lugar de $n_i + n_j - 2$, pues la estimación de $\hat{s}_R$ se realiza con todas las observaciones.

Si estamos en la fase de selección del mejor predictor, una vez hemos agrupado sus categorías, y suponiendo ahora que k es el número total de categorías finales:

$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$,

$$F = \frac{\hat{s}_e^2}{\hat{s}_R^2} \sim F_{(k-1),(n-k)} \tag{3.15}$$

- Aquí, puesto que también analizamos la significación de los predictores después de haber realizado una reagrupación de sus categorías, tenemos la posibilidad de ajuste de Bonferroni y de comparación múltiple, y el programa escogerá el mínimo valor de ambos como representación de la significación del predictor en estudio. El programa nos permite introducir información externa adicional para la estimación de $\hat{s}_R$ de los denominadores de los test en ambas fascs, tanto del valor que puede que conozcamos *a priori* como de los grados de libertad implicados, si no le indicamos de modo expreso que utilice la estimación con las categorías agrupadas (que es el estadístico anterior), por defecto escogerá la estimación realizada de la tabla antes de colapsar sus categorías.

3.4.3. Aplicación actuarial

Hace unos años se empezaron a utilizar las técnicas de segmentación, fundamentalmente en el seguro del automóvil, como una herramienta de toma de decisiones dc las entidades [Calatayud y Martínez (1997); Pérez (2001)]. Una aplicación concreta de CHAID la tenemos en la segmentación de la cartera común de automóviles de UNESPA [UNESPA (1995)] elaborada por Sánchez (1997). Las variables explicativas que se consideraron teniendo en cuenta que "eran las que mejor explicaban desde el punto de vista estadístico la siniestralidad" fueron: categoría-clase del vehículo; zona de riesgo; antigüedad del carnet de conducir; edad del conductor; sexo del conductor. Y como variable de siniestralidad la prima media de riesgo, pues se trataba de una estadística común, dividida en 11 categorías y conseguida como producto entre el importe medio de los siniestros y el número medio de siniestros. El resultado final del análisis proporcionó 159 segmentos terminales. Posteriormente, se calculó una prima media de riesgo para cada segmento utilizando la información original, pues éste era el objetivo básico de tal análisis, una tarificación *a priori* lo más ajustada posible[18].

En la literatura se desarrollan muchas técnicas de segmentación distintas y es el investigador quien decide cual de ellas utilizará, calibrando las ventajas e inconvenientes, el tipo de datos de que dispone y el tipo de variable que quiere explicar. Nosotros, en el capítulo 5, utilizamos 3 tipos de algoritmo ya programados: el típico CHAID para variable respuesta categórica nominal (SPSS CHAID nominal y

[18] Cabe notar, el programa utilizado fue el SPSS CHAID. En principio, tal y como sería adecuado, parece que fue utilizado el algoritmo ordinal, pues el resultado descriptivo visual es la media, lo que entra en contradicción con la pantalla de opciones mostrada en la página 57, donde se especifica algoritmo nominal!!!.

CATFIRM), el CHAID para variable respuesta categórica ordinal (SPSS CHAID ordinal) y el XAID para respuesta continua (CONFIRM).

En el caso de la tarificación, vamos a disponer siempre de una respuesta cuantitativa, por lo que, en principio, el algoritmo apropiado será el XAID. De entre el resto de opciones, la más aproximada a la realidad, tal y como veremos empíricamente en la aplicación 1 del capítulo 5, será la del SPSS CHAID ordinal, seguida del CATFIRM y el SPSS CHAID nominal.

Respecto a los predictores, en cualquier caso, deberán ser de tipo categórico (nominales u ordinales), resultando conveniente que tengan pocas modalidades. La utilización de más de 10 categorías, por ejemplo, puede acarrear problemas tanto desde el punto de vista práctico como desde el punto de vista teórico. En el caso de predictores cuantitativos, éstos deben someterse a un proceso adecuado de recodificación para convertirlos en categóricos. Puesto que el análisis de segmentación no es robusto, es muy sensible a pequeños cambios en los datos, debemos ir con sumo cuidado. Con los programas de SPSS CHAID, tal discretización debemos realizarla *a priori* e introducir los datos ya codificados. Con los programas de D. M. Hawkins, FIRM, tenemos la posibilidad de entrar la variable con sus valores continuos y el programa realiza una discretización basada en formar 10 grupos de aproximadamente igual frecuencia, sin contar la clase de *missing* en el caso de que la haya; y si no queremos este criterio para la discretización o bien otro número de clases para el predictor, debemos realizarla a parte. En el apartado 3.6. de este capítulo veremos maneras alternativas de discretizar variables continuas, distinguiendo si se trata de respuestas o predictores. Los programas de SPSS CHAID permiten como mucho 31 categorías diferentes y los de FIRM como mucho 16.

En los resultados visuales, si la variable dependiente es cuantitativa la descriptiva mostrada en los segmentos terminales básicamente es la media y la desviación típica (que será el caso del CONFIRM, y se le suma el SPSS CHAID ordinal, aunque la respuesta sea categórica ordinal), y si la variable dependiente es cualitativa la descriptiva se refiere al porcentaje de individuos en cada una de las categorías (que será el caso del CATFIRM y SPSS CHAID nominal).

Siempre es de interés partir de una base de datos con información desagregada; por un lado, para la estimación de la siniestralidad a partir de los datos originales en cada segmento terminal y, por otro, para la correcta discretización de la respuesta y de los predictores en los casos oportunos:

- Para el CONFIRM: introduciremos directamente la respuesta continua. Si la información disponible es agregada entonces a cada individuo le corresponderá la media de su clase y podremos

aprovechar la variable frecuencia para la introducción de datos. En cualquier caso la estimación de la siniestralidad en cada segmento terminal la realizaremos con la información más aproximada posible.

- Para el SPSS CHAID ordinal: primeramente discretizaremos la respuesta continua con algún criterio adecuado, como por ejemplo el método de cluster de Ward, que detallaremos en el apartado 3.6. El algoritmo nos permite en como mucho 31 niveles; para cada nivel realizaremos alguna media de la siniestralidad original, y ésta será la puntuación que asignaremos de manera común a todos los individuos del nivel correspondiente; y finalmente, para la estimación de la siniestralidad en cada segmento terminal, realizaremos la media con los datos originales. Si partimos de información agregada, nos será más difícil realizar la discretización inicial de la respuesta, y las puntuaciones serán menos precisas.

- Para el SPSS CHAID nominal y el CATFIRM: necesitaremos, al igual que en el caso anterior, tener la respuesta discretizada. En principio parece lógico escoger el número máximo de categorías que nos permita el algoritmo (31 para SPSS CHAID nominal y 16 para CATFIRM) pero para este algoritmo nominal a medida que bajemos de nivel en el árbol, las tablas de contingencia formadas para el cálculo de los p-valores, pueden resultar dispersas, así que debemos equilibrar la pérdida de información con el hecho de que los p-valores calculados en las diferentes etapas tengan sentido. Una vez tenemos las diferentes categorías, tan sólo es necesario asignar un código a cada clase. De los segmentos terminales, podemos proceder como en los dos casos anteriores, estimando la siniestralidad con la media de los valores originales.

Finalmente, deberemos tener presente que la segmentación de la cartera discriminará de forma importante las primas de cada nodo terminal, de forma que, en determinados grupos, la prima puede resultar impagable. Las primas máxima y mínima deben estar en un rango de variación aceptable, por lo que en algunos casos será conveniente repartir la diferencia entre todo el colectivo.

3.5. Modelo lineal generalizado

3.5.1. Descripción del modelo

Nos referimos a los libros Dobson (2001), McCullagh y Nelder (1989), y en castellano a López y López de la Manzanara (1986, pp. 125-145) para una descripción detallada del MLG. Adicionalmente

a la bibliografía referenciada en algún momento del trabajo respecto al MLG en la tarificación tenemos: Ajne (1975,1980,1986); Bennet (1978); Berg (1980); Chang y Fairley (1979); Fairley y Tomberling (1981); Holler, Sommer y Trahair (1999); Jørgensen y Paes de Souza (1994); Jung (1968); Murphy, Brockman y Lee (2000); Renshaw (1993); Smyth y Jørgensen (To appear); Toniolo y Schmitter (1998); ... y redireccionamos a: http://www.statsci.org/glm/bibliog.html donde podemos encontrar bibliografía selecta que incluye la citada.

Vamos a ver, en este apartado, las nociones básicas que nos ayudarán a entender su uso en las aplicaciones prácticas del capítulo 5.

Supongamos la variable aleatoria $\mathbf{Y}_{(n\times 1)}$, con (y_i) para $i=1,2,\ldots,n$ observaciones independientes, que recogen la siniestralidad y juegan el papel de variable respuesta en la regresión, y supongamos los predictores o factores potenciales de la estructura de riesgo $\mathbf{F}_1,\mathbf{F}_2,\ldots,\mathbf{F}_P$, vectores $(n\times 1)$: (f_{ij}) para $i=1,2,\ldots,n$ y $j=1,2,\ldots,P$. Recordemos el modelo clásico de regresión lineal por mínimos cuadrados ordinarios, en el que suponemos una distribución del error ε_i Normal centrada y con varianza constante, $\varepsilon_i \sim N(0,\sigma^2)$. La relación lineal de la respuesta con la estructura sistemática dada por los predictores es:

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j f_{ij} + \varepsilon_i \tag{3.16}$$

Así, tenemos observaciones independientes $y_i \sim N(\mu_i,\sigma^2)$, con esperanza:

$$E[y_i] = \mu_i = \beta_0 + \sum_{j=1}^{p} \beta_j f_{ij} \tag{3.17}$$

y varianza constante:

$$Var[y_i] = \sigma^2. \tag{3.18}$$

En el MLG seguimos teniendo (y_i) para $i=1,2,\ldots,n$ observaciones independientes de la respuesta, unos errores centrados $E[\varepsilon_i]=0$, y un predictor lineal determinista al que simbolizamos por η_i:

$$\eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j f_{ij}\ . \tag{3.19}$$

Las dos extensiones respecto al modelo clásico son:

1) La distribución de **Y** no tiene porqué ser la Normal. Consideramos en primer lugar que puede ser cualquier distribución derivada de la familia exponencial [McCullagh y Nelder (1989) p.28]. Estas distribuciones se caracterizan por tener la función de densidad o de probabilidad en un punto de la forma:

$$f(y_i;\theta_i,\phi_i)=\exp\left\{\frac{y_i\theta_i-b(\theta_i)}{a(\phi_i)}+c(y_i,\phi_i)\right\} \tag{3.20}$$

para funciones especificadas $a(\cdot)$, $b(\cdot)$ y $c(\cdot)$. Donde θ_i se denomina parámetro canónico y ϕ_i parámetro de dispersión. Puede deducirse fácilmente a partir de la formulación más general que:

$$E[y_i]=\mu_i=b'(\theta_i) \tag{3.21}$$

$$Var[y_i]=b''(\theta_i)a(\phi_i) \tag{3.22}$$

Por lo tanto la varianza es el producto de dos componentes:

- la primera, $b''(\theta_i)$, depende sólo de θ_i (y por tanto de la esperanza μ_i). A esta componente se la denomina *función de varianza* y se explicita su dependencia respecto de la esperanza: $b''(\theta_i)=V(\mu_i)$

- la segunda, $a(\phi_i)$, depende sólo del parámetro de dispersión ϕ_i y usualmente adopta la forma $a(\phi_i)=\frac{\phi}{w_i}$, con parámetro de dispersión constante para todas las observaciones, ϕ, y unos pesos especificados *a priori*, w_i, que varían de observación a observación.

Por lo que la varianza la rescribimos como:

$$Var[y_i]=\frac{\phi V(\mu_i)}{w_i} \tag{3.23}$$

donde ϕ es el parámetro de dispersión, $V(\mu_i)$ es la función de varianza y w_i es el posible peso especificado *a priori* de la observación *i*. Notamos que suponiendo $\phi=1$, los recíprocos de los

pesos pueden reinterpretarse como parámetros de escala no constantes: $1/w_i = \phi_i$.

2) La respuesta está ligada con el predictor lineal a través de una función F:

$$E[y_i] = \mu_i = F(\eta_i) = F\left(\beta_0 + \sum_{j=1}^{p} \beta_j f_{ij}\right) \tag{3.24}$$

Para tener despejada la respuesta, deberemos hacer la función inversa de F, $g = F^{-1}$, a la que denominamos *función de enlace (link function)*, pues es la que nos enlazará la respuesta con el predictor lineal. A la función de enlace, g, le exigimos que sea monótona y diferenciable:

$$g\left(E[y_i]\right) = g(\mu_i) = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j f_{ij} \tag{3.25}$$

Existen, para algunas distribuciones de la familia exponencial, funciones de enlace "naturales", también denominadas canónicas. Para estos enlaces canónicos se produce que el parámetro canónico coincide con el predictor lineal: $\theta(\mu_i) = \eta_i$.

Pero, en general, podemos optar por modelizar con cualquier otro que no sea el canónico. Es usual utilizar uno derivado de la familia de enlaces paramétricos:

$$\eta_{iu} = g(\mu_{iu}) = \begin{cases} \mu_{iu}^{\lambda} & \text{para} \quad \lambda \neq 0 \\ \log(\mu_{iu}) & \text{para} \quad \lambda = 0 \end{cases} \tag{3.26}$$

Así, en el MLG, tenemos dos extensiones respecto al modelo de regresión lineal clásico, una respecto a la distribución del error plasmada en $Var(\mathbf{Y})$, que podrá proceder de cualquiera de las de la familia exponencial y no tiene por qué ser la Normal; y otra respecto a la función de enlace, $g(\boldsymbol{\mu})$, que debe ser una función monótona diferenciable y no tiene por qué ser la Identidad. En la tabla 3.2 del anexo 3.2 se recogen las distribuciones más conocidas que forman parte de la familia exponencial definida en (3.20), junto con el enlace canónico que llevan asociado.

Al aplicar el MLG podemos elegir la distribución del error y la función de enlace. La utilización del enlace canónico correspondiente a cada distribución tiene la ventaja de simplificar la formulación, pero no tiene porqué implicar que sea el más adecuado para unos datos particulares. Si el objetivo es seleccionar un modelo, la simplicidad de la función de enlace no debe sustituir a la calidad del ajuste

como criterio. En cuanto a los valores que puede tomar la respuesta, **Y**, y centrándonos en el caso de la tarificación, será más apropiado utilizar una distribución u otra según analicemos las cuantías de los siniestros o el número de siniestros por póliza. Por ejemplo [Brockman y Wright (1992); Coutts (1984a); Haberman y Renshaw (1996,98); Hipp (2000); Mack (1991)], si analizamos la cuantía de los siniestros será apropiado utilizar una distribución Gamma o una Gaussiana Inversa preferiblemente a una Normal, que no toman valores negativos y tienen asimetría positiva, y si analizamos el número de siniestros será más adecuado utilizar una distribución de Poisson, una Binomial o una Binomial Negativa.

La estimación de los parámetros β_j se realiza mediante la maximización del logaritmo de la función de verosimilitud, que es:

$$l = \sum_{i=1}^{n} l_i = \sum_{i=1}^{n} \left\{ \frac{y_i \theta_i - b(\theta_i)}{a(\phi_i)} + c(y_i, \phi_i) \right\} \quad . \tag{3.27}$$

Cabe notar que suponiendo una distribución del error Normal y la Identidad como función de enlace, obtenemos como caso particular la solución por MCO del modelo clásico.

La versión más general del MLG no exige que la distribución del error de **Y** pertenezca a la familia de distribuciones exponenciales caracterizadas por (3.20). En esta versión más general seguimos teniendo (y_i) para $i = 1, 2, \ldots, n$ observaciones independientes de la respuesta para las cuales se conocen sólo los dos primeros momentos:

$$E[y_i] = \mu_i \tag{3.28}$$

$$Var[y_i] = \frac{\phi V(\mu_i)}{w_i} \tag{3.29}$$

siendo $V(\mu_i)$ una función de varianza especificada, ϕ el parámetro de dispersión también especificado (en general positivo), y w_i los posibles pesos *a priori* de las observaciones.

En este contexto más amplio, aplicable a distribuciones que no pertenezcan a la familia exponencial, la estimación de los parámetros β_j se realiza maximizando los logaritmos de las funciones de cuasi-verosimilitud, que son:

$$q(\mathbf{y};\boldsymbol{\mu}) = \sum_{i=1}^{n} q_i = \sum_{i=1}^{n} w_i \int_{y_i}^{\mu_i} \frac{y_i - s}{\phi V(s)} ds \tag{3.30}$$

Tal maximización nos lleva a resolver el sistema de ecuaciones lineales:

$$\sum_{i=1}^{n} w_i \frac{y_i - \mu_i}{\phi V(\mu_i)} \frac{\partial \mu_i}{\partial \beta_j} = 0 \quad \forall j \tag{3.31}$$

mediante algún método numérico. Para este caso general de la familia exponencial, las quasi-verosimilitudes juegan el papel de verosimilitudes.

En el MLG la variabilidad no explicada por un modelo M (fijada una función de enlace y una distribución del error) se plasma en la *desvianza escalada* $D^*(\mathbf{y};\hat{\boldsymbol{\mu}})$. Si $L(\mathbf{y};\hat{\boldsymbol{\mu}})$ denota la función de verosimilitud del modelo M y $L(\mathbf{y};\mathbf{y})$ la función de verosimilitud del modelo saturado[19], entonces:

$$D^*(\mathbf{y};\hat{\boldsymbol{\mu}}) = -2\log\left(\frac{L(\mathbf{y};\hat{\boldsymbol{\mu}})}{L(\mathbf{y};\mathbf{y})}\right) \tag{3.32}$$

ésta se hace menor a mayor número de número de predictores incluidos en el modelo, hasta llegar a explicar la variabilidad total de los datos. En la tabla 3.3 del anexo 3.2, detallamos la expresión que toma en algunos de los casos particulares de la familia exponencial, pero en términos de desvianzas no escaladas:

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = \phi D^*(\mathbf{y};\hat{\boldsymbol{\mu}}). \tag{3.33}$$

Podemos escribir las desvianzas en términos de las cuasi-verosimilitudes descritas en (3.30):

$$D(\mathbf{y};\hat{\boldsymbol{\mu}}) = \sum_{i=1}^{n} d_i = \sum_{i=1}^{n} 2w_i \frac{y_i - s}{V(s)} d(s) = -2\phi q(\mathbf{y};\hat{\boldsymbol{\mu}}). \tag{3.34}$$

Así, para la construcción de las desvianzas asociadas al modelo especificado, tan sólo necesitamos conocer los dos primeros momentos.

[19] El modelo saturado es el que tiene tantos parámetros como individuos, y por lo tanto se cumple $\hat{\mu}_i = y_i$ para i=1,2,...,n

Es posible, que en este caso general, deseemos realizar inferencias sobre parámetros implicados en la varianza (3.29) que diseñemos. O bien sobre el parámetro de dispersión, ϕ, o bien sobre otros parámetros implicados en la función de varianza, $V(\mu_i)$. Para ello se suele analizar menos dos veces la versión extendida del logaritmo de las cuasi-verosimilitudes,

$$-2q^{+} = \sum_{i=1}^{n} \frac{d_i}{\phi} + \sum_{u=1}^{n} \log\{\phi V(y_i)\}, \tag{3.35}$$

ante cambios infinitesimales del parámetro estudiado.

Si necesitamos estimar el parámetro de dispersión, ϕ, en un modelo con p coeficientes (incluido el término constante) en el predictor lineal, podemos utilizar el denominador de (3.41):

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} d_i \tag{3.36}$$

O alternativamente el estimador de momentos basado en los residuos generalizados de Pearson:

$$\hat{\phi} = \frac{1}{n-p} \sum_{i=1}^{n} w_i \frac{(y_i - \hat{\mu}_i)^2}{V(\hat{\mu}_i)}. \tag{3.37}$$

Usualmente se indica que el modelo más apropiado para unos datos determinados es aquél que nos ofrece una menor desvianza. Como se intuye, tenemos diferentes maneras de reducirla [Millenhall (1999)]:

- si variamos la función de enlace,
- si variamos la distribución del error,
- y/o si variamos los factores de riesgo incluidos en el predictor lineal.

Puesto que nosotros tenemos como objetivo la selección de variables de tarifa, fijaremos un enlace y un error y a partir de aquí realizaremos el proceso de selección: seleccionaremos los factores de riesgo para un modelo dado.

3.5.2. Selección de predictores

Supongamos dos modelos anidados $M_s \subset M_r$:

M_r: con r + 1 parámetros $(\beta_0, \beta_1, \ldots, \beta_s, \beta_{s+1}, \ldots, \beta_r)$

M_s: con s + 1 parámetros $(\beta_0, \beta_1, \ldots, \beta_s)$

Queremos testar si $r-s$ de los $r+1$ parámetros son cero, i.e., si las variables explicativas $\mathbf{F}_{s+1}, \mathbf{F}_{s+2}, \ldots, \mathbf{F}_r$ tienen una influencia significativa en la experiencia de siniestralidad esperada. En otras palabras, testamos si el modelo más pequeño, M_s, describe los datos de manera más adecuada que el modelo mayor, M_r. Sin pérdida de generalidad, testamos para los últimos $r-s$ de los $r+1$ parámetros:

$$\begin{aligned} &H_0 : \beta_{s+1} = \beta_{s+2} = \cdots = \beta_r = 0 \\ &\text{versus} \\ &H_1 : \text{no todos los } \beta_i \ (i = s+1, s+2, \ldots, r) \text{ son } 0 \end{aligned} \tag{3.38}$$

Para el cálculo de los *p*-valores asociados al contraste, podemos:

a) Utilizar la distribución asintótica para la familia exponencial:

a.1) Siguiendo a Albrecht (1983a,b) y a Zehnwirth (1994) entre otros: Si $\hat{\boldsymbol{\mu}}_s$ denota la estimación máximo verosímil y L_s la función de verosimilitud respecto al modelo M_s, y $\hat{\boldsymbol{\mu}}_r$, L_r de igual modo para el modelo M_r, entonces el estadístico razón de verosimilitud, *RV*, para este problema de hipótesis es

$$RV = \frac{L_s(\mathbf{y}; \hat{\boldsymbol{\mu}}_s)}{L_r(\mathbf{y}; \hat{\boldsymbol{\mu}}_r)}, \tag{3.39}$$

y $-2 \cdot \log(RV)$ tiene una distribución asintótica *ji*-cuadrado: χ^2_{r-s}. Que reescrito en términos de desvianzas escaladas de los correspondientes modelos, $D_0^* = -2\log\left(\frac{L_s(\mathbf{y}; \hat{\boldsymbol{\mu}}_s)}{L(\mathbf{y}; \mathbf{y})}\right) \sim \chi^2_{n-s-1}$ y $D_1^* = -2\log\left(\frac{L_r(\mathbf{y}; \boldsymbol{\mu}_r)}{L(\mathbf{y}; \mathbf{y})}\right) \sim \chi^2_{n-r-1}$, tenemos que,

$$D_0^* - D_1^* = -2\log\left(\frac{L_s(\mathbf{y};\boldsymbol{\mu}_s)}{L_r(\mathbf{y};\boldsymbol{\mu}_r)}\right) \sim \chi^2_{r-s} \tag{3.40}$$

a.2) Siguiendo a Agsaa (1977); López y López de la Manzanara (1996, pp. 137-139); Brockman y Wright (1992) entre otros, tenemos que:

$$\frac{(D_0^* - D_1^*)/(r-s)}{D_1^*/(n-r-1)} \sim F_{(r-s,n-r-1)} \tag{3.41}$$

b) Estimar la distribución exacta de los estadísticos mediante simulación, por ejemplo haciendo uso de la metodología *bootstrap* [Diaconis y Efron (1983); Efron y Tibshirani (1993)]. La estimación mediante *bootstrap* de la distribución de probabilidad de estadísticos permite, vía simulación, realizar tests de hipótesis con muestras de tamaño finito, tests que, de otro modo, deberían aproximarse con una distribución asintótica. Véase también Delicado y Placencia (2001) para un estudio detallado del uso de herramientas gráficas y numéricas para resumir los resultados de los estudios de simulación concernientes a los tests de hipótesis.

3.5.2.1. Proceso de selección

Una vez sabemos como contrastar si un conjunto de coeficientes es significativo en la bondad del ajuste de la regresión, se trata de organizar un proceso de selección. Puesto que en general dispondremos de un conjunto elevado de predictores, resultará casi imposible estudiar el ajuste de todos los modelos posibles. Lo usual es optar por un proceso de introducción progresiva, por uno de eliminación progresiva o por uno paso a paso que combine los dos anteriores. En principio resultaría apropiado utilizar un proceso de eliminación progresiva, pero tiene el inconveniente de que ya en el primer paso del proceso debemos tener todas las variables incorporadas en el modelo, por lo que los modelos a estimar son complejos. Esto no ocurre con el de introducción progresiva, pues suponemos que inicialmente no hay ninguna variable seleccionada, vamos introduciendo una a una, y los modelos son en principio simples. Pero corremos el riesgo de detener el proceso con un modelo que incorpore información redundante, olvidándonos de otras combinaciones mayormente significativas. Así parece razonablemente correcto utilizar una combinación de introducción y eliminación, el denominado paso a paso.

Proceso de selección paso a paso

El proceso se inicia suponiendo que no disponemos de ninguna variable incorporada en el modelo, y combina en cada paso una fase de introducción, para ver qué variable es la siguiente a ser introducida, con una de eliminación para comprobar si alguna de las variables remanentes se ha vuelto no significativa. Fijemos un nivel de significación α^* y supongamos que disponemos en total de P factores potenciales de riesgo, $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_P$. Sea k el número de variables seleccionadas resultantes en el paso anterior, $\mathbf{F}(1), \mathbf{F}(2), \ldots, \mathbf{F}(k)$, con $k \leq P$. Definimos, de manera genérica, las fases de introducción y de eliminación para cada paso del proceso del siguiente modo:

Fase de Introducción: Seleccionamos, al menos temporalmente, la variable que al ser introducida explique un mayor porcentaje de variabilidad mediante el menor p-valor de entre las P - k variables aún no seleccionadas a partir de los correspondientes estadísticos $Q\left(\mathbf{F}_p \mid \mathbf{F}(1)\mathbf{F}(2)\cdots\mathbf{F}(k)\right)$. Denotamos por $Q\left(\mathbf{F}_p \mid \mathbf{F}(1)\mathbf{F}(2)\cdots\mathbf{F}(k)\right)$ al estadístico utilizado, bien *ji*-cuadrado (3.40) bien F de Fisher Snedecor (3.41), en la contrastación de si la variable $\mathbf{F}_p$ añade suficiente variabilidad explicada al modelo con $\mathbf{F}(1)\mathbf{F}(2)\cdots\mathbf{F}(k)$ dadas. Así, la variable a incorporar será:

$$\mathbf{F}(k+1) = \left\{ \mathbf{F}_p \,/\, \min_{p \in \{1,2,\ldots,P\} \setminus \{(1),(2),\ldots,(k)\}} p - valor\left(\mathbf{F}_p \mid \mathbf{F}(1)\mathbf{F}(2)\cdots\mathbf{F}(k)\right) \right\}. \tag{3.42}$$

Fase de Eliminación: Si alguna de las variables seleccionadas en las fases de introducción se vuelve no significativa en el modelo es eliminada. Para ello calculamos los p-valores, $p(i)$, de los correspondientes estadísticos $Q\left(\mathbf{F}(i) \mid \mathbf{F}(1), \ldots, \mathbf{F}(i-1), \mathbf{F}(i+1), \ldots, \mathbf{F}(k+1)\right)$ para $i = 1, 2, \ldots, k+1$, y elegimos el máximo p-valor al que nomenclamos $p(m)$,

$$p(m) = máx\left\{p(i)\right\}_{i=1,2,\ldots,k+1} \tag{3.43}$$

Entonces,

→ Si $p(m) < \alpha^* \Rightarrow$ ninguna variable es eliminada

→ Si $p(m) \geq \alpha^* \Rightarrow \mathbf{F}(m)$ es eliminada del modelo,

- Si $(m) \neq (k+1)$ vamos al paso siguiente con el conjunto:

  $\{\mathbf{F}(1),\ldots,\mathbf{F}(m-1),\mathbf{F}(m+1),\ldots,\mathbf{F}(k+1)\}$.

- Si $(m) = (k+1)$ el proceso se detiene con el conjunto resultante de predictores:

  $\{\mathbf{F}(1),\mathbf{F}(2),\ldots,\mathbf{F}(k)\}$.

Cabe destacar que durante un proceso de selección se obtiene información útil tanto de la relación como de la importancia de todas las variables indiferentemente de si finalmente son o no seleccionadas.

3.5.2.2. Validación del modelo resultante

Una vez realizado el proceso de selección, obtenemos como resultado un conjunto de factores seleccionados. Posteriormente, siempre es conveniente estudiar la significación conjunta del conjunto de coeficientes. Para ello, presentamos el siguiente contraste estadístico, que consiste en comprobar si la variabilidad explicada por los predictores seleccionados es suficiente como para que la regresión tenga poder predictivo. Para una regresión con p+1 parámetros, $\beta_0,\beta_1,\beta_2,\ldots,\beta_p$, el contraste es el siguiente:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0 \tag{3.44}$$

Mediante el estadístico F, que no es más que el cociente entre la variabilidad total explicada sobre la total a explicar, tenemos, en términos de desvianzas escaladas:

$$\frac{(D_0^* - D_1^*)/p}{D_1^*/(n-p-1)} \sim F_{(p,n-p-1)} \tag{3.45}$$

Podemos calcular el p-valor asociado bien asintóticamente, bien mediante simulación.

Con este contraste no sólo podemos ver si aceptamos o rechazamos la hipótesis nula, sino también en qué medida. Por lo que esta vía puede ser una manera correcta de comparar dos regresiones con los

mismos predictores pero con diferentes funciones enlaces y/o distribuciones del error, nos quedaríamos con aquélla que nos proporcionara un menor p-valor en el contraste.

Adicionalmente, en las aplicaciones del capítulo 5, plasmamos el *pseudo* R^2 que, en función de las desvianzas escaladas utilizadas en (3.45), se define como:

$$R^2 = \frac{D_0^* - D_1^*}{D_0^*}. \tag{3.46}$$

Es una versión del coeficiente de determinación de la regresión clásica en el caso del MLG. Es una medida acotada entre 0 y 1, y aumenta al añadir regresores en el modelo.

Finalmente, siempre es adecuado realizar algún gráfico de residuos que visualmente nos indique el ajuste del modelo, por ejemplo de,

- residuos de Pearson: $\dfrac{y_i - \hat{\mu}_i}{\left(V(\hat{\mu}_i)/w_i\right)^{1/2}}$
- residuos de desvianza: $signo(y_i - \hat{\mu}_i)\sqrt{d_i}$

  (donde d_i es la i-ésima componente de (3.34))

Aunque más que para una validación formal nos sirve para un análisis de puntos aislados inicial.

3.5.2.3. Codificación de predictores

Respecto a la codificación como *input* de variables cualitativas, tanto ordinales (incluyendo, si es el caso, a las continuas discretizadas) como nominales, en el MLG, debemos crear tantas variables binarias como clases tenga la variable, y tratarlas como cuantitativas en el modelo. Puesto que estimamos un coeficiente para cada clase, podremos ver qué clases tienen los coeficientes más significativos, y podremos realizar procesos de selección tomando a las clases como variables por si solas. Aunque lo usual para la selección de predictores es realizar un proceso con los efectos principales, por lo que cada efecto principal, filosóficamente, vendrá representado por el conjunto

completo de binarias que lo resumen.

La codificación de muchas variables binarias es engorrosa. Si el factor posee un número elevado de clases, las binarias asociadas serán numerosas. Si no sólo queremos estudiar los efectos principales, sino también las interacciones, iremos añadiendo parámetros hasta como mucho construir el modelo saturado, el cual posee tantas variables como individuos. En la aplicación 3 del capítulo 5 describimos como llegar al modelo saturado si tratamos con datos agregados.

Respecto al proceso de selección, hemos empezado definiendo el contraste (3.38) para un conjunto de coeficientes. En (3.42) hemos expresado por $\mathbf{F}(k+1)$ a un factor determinado. Pero si el factor es categórico, no será una sola variable a entrar en el modelo, sino el conjunto de binarias que lo representan. Este hecho debe tenerse en cuenta en los grados de libertad de los estadísticos si hacemos uso de las distribuciones asintóticas (3.40) y (3.41). Por todo ello, consideramos adecuado comparar los *p*-valores, en lugar del estadístico, la desvianza media o la desvianza escalada directamente.

Usualmente dispondremos de un conjunto numeroso de factores de donde seleccionar las variables de tarifa. Si todos son categóricos, puede resultar que dos de ellos posean binarias de algunas clases muy correlacionadas. En este caso deberemos decidir por cual de las dos (o más) binarias "comunes" nos decantamos para evitar problemas de colinealidad[20]. Esto no sólo se acentúa en el primer paso de un proceso de eliminación progresiva en el que los modelos son complejos, también ocurre a medida que los modelos se van complicando en un paso a paso o en uno de introducción.

Cabe notar que discretizar de una manera u otra factores continuos tiene muchísima influencia en los resultados finales, pues ya sólo tenemos en cuenta las binarias y no los pesos lineales de las observaciones continuas.

Para una variable cualitativa nominal, codificaremos una binaria para cada categoría, y una de ellas irá a parar al efecto global de modelo. Para las variables cualitativas ordinales, adicionalmente, tenemos la posibilidad de incorporar la ordinalidad de las clases en las binarias construidas. Utilizamos la aplicación 1 del capítulo 5 con el MLG de distribución del error Normal y enlace logarítmico para ilustrar como en el MLG, el hecho de incluir la ordinalidad en las variables binarias, referentes en este caso a la antigüedad en el puesto laboral, tiene tan sólo una diferencia interpretativa de los coeficientes ya que la estimación final no se ve modificada [Suits (1984)]:

[20] Si las regresoras están relacionadas, los valores estimados presentarán inestabilidad.

Respecto al estado civil, éste es categórico nominal por lo que codificamos una binaria para cada categoría:

$$X_{E1}=\begin{cases}0 & \text{en otro caso}\\ 1 & \text{si } i\in E1\end{cases}\quad X_{E2}=\begin{cases}0 & \text{en otro caso}\\ 1 & \text{si } i\in E2\end{cases}\quad X_{E3}=\begin{cases}0 & \text{en otro caso}\\ 1 & \text{si } i\in E3\end{cases}\qquad (3.47)$$

y hacemos que, por ejemplo, X_{E3} se corresponda con el efecto base β_0.

Respecto a la antigüedad, la trataremos como categórica nominal utilizando, al igual que para el estado, las siguientes binarias disjuntas (dejando X_{A3} como base):

$$X_{A1}=\begin{cases}0 & \text{en otro caso}\\ 1 & \text{si } i\in A1\end{cases}\quad X_{A2}=\begin{cases}0 & \text{en otro caso}\\ 1 & \text{si } i\in A2\end{cases}\quad X_{A3}=\begin{cases}0 & \text{en otro caso}\\ 1 & \text{si } i\in A3\end{cases}\qquad (3.48)$$

Y la trataremos como categórica ordinal si utilizamos la siguiente codificación ordinal para las binarias asociadas a las clases:

$$X_{Ant\geq 2}=\begin{cases}0 & \text{en otro caso}\\ 1 & \text{si antig.}\geq 2\end{cases}\qquad X_{Ant\geq 10}=\begin{cases}0 & \text{en otro caso}\\ 1 & \text{si antig.}\geq 10\end{cases}\qquad (3.49)$$

Si realizamos la estimación, utilizando el estado, con sus respectivas binarias disjuntas, y en cada caso la antigüedad con las binarias disjuntas u ordinales, obtenemos el mismo resultado:

	A1	A2	A3
E1	218.17	291.04	345.22
E2	167.52	223.47	265.07
E3	174.93	233.37	276.81

Lo que sí varía son los coeficientes:

Con las binarias disjuntas,

$$\log(\hat{\mu}_i)=\hat{\beta}_{A3E3}+\hat{\beta}_{A1}X_{A1}+\hat{\beta}_{A2}X_{A2}+\hat{\beta}_{E1}X_{E1}+\hat{\beta}_{E2}X_{E2}$$

$$\log(\hat{\mu}_i)=5.6233-0.4589X_{A1}-0.1707X_{A2}+0.2208\cdot X_{E1}-0.0433X_{E2}$$

$$\hat{\mu}_i = 276.81 \times 0.6320^{X_{A1}} \times 0.8431^{X_{A2}} \times 1.2471^{X_{E1}} \times 0.9576^{X_{E2}}$$

Con la codificación ordinal,

$$\log(\hat{\mu}_i) = \hat{\beta}_{A1E3} + \hat{\beta}_{Ant\geq 2} X_{Ant\geq 2} + \hat{\beta}_{Ant\geq 10} X_{Ant\geq 10} + \hat{\beta}_{E1} X_{E1} + \hat{\beta}_{E2} X_{E2}$$

$$\log(\hat{\mu}_i) = 5.1644 + 0.2882 X_{Ant\geq 2} + 0.1707 X_{Ant\geq 10} + 0.2208 X_{E1} - 0.0433 X_{E2}$$

$$\hat{\mu}_i = 174.93 \times 1.3340^{X_{Ant\geq 2}} \times 1.1861^{X_{Ant\geq 10}} \times 1.2471^{X_{E1}} \times 0.9576^{X_{E2}}$$

Si predecimos $\hat{\mu}_i$, por ejemplo, para la clase E2 del estado, tenemos los siguientes productos de coeficientes en cada caso:

Para los de antigüedad menor a 2 años:

a) Binarias disjuntas: $276.81 \times 0.6320 \times 0.9576 = 167.52$

b) Binarias ordinales: $174.93 \times 0.9576 = 167.52$

Para los de antigüedad entre 2 y 10 años:

a) Binarias disjuntas: $276.81 \times 0.8431 \times 0.9576 = 223.47$

b) Binarias ordinales: $174.93 \times 1.3340 \times 0.9576 = 223.47$

Para los de antigüedad mayor a 10 años:

a) Binarias disjuntas: $276.81 \times 0.9576 = 265.07$

b) Binarias ordinales: $174.93 \times 1.3340 \times 1.1861 \times 0.9576 = 265.07$

Si nos fijamos en los de antigüedad mayor a 10 años, con las binarias ordinales, la interpretación de los coeficientes es más pesada que con las disjuntas: para incluir el efecto de que la antigüedad es mayor a 10 años, hemos de ir multiplicando por los coeficientes de las clases "menores", que nos proporcionan el efecto marginal de pasar de una categoría ordinal a otra, en este caso, incluir que los de antigüedad mayor a 10 también tienen antigüedad mayor a 2.

Si la variable discretizada tiene un número elevado de categorías ordinales, el productorio se incrementa. Es una cuestión simple de interpretación de coeficientes, por lo que si el objetivo es simplificar, manejando un número no muy elevado de coeficientes, vale la pena utilizar uno por clase cuando tratamos al MLG.

En el caso del MLG, el tratamiento de categórico (tanto ordinal como nominal) se realiza mediante la construcción de binarias, que tanto si son disjuntas como ordinales, desembocan en la misma estimación tal y como hemos visto. Esas binarias son tratadas en el modelo como variables cuantitativas.

En el modelo basado en distancias, que presentaremos en el siguiente capítulo, este hecho es muy diferente:

- Si disponemos de un predictor categórico nominal, podemos tratarlo como categórico nominal, o bien como un conjunto binarias disjuntas asociadas a cada clase nominal.

- Si disponemos de un predictor categórico ordinal, podemos tratarlo como categórico nominal, si estamos dispuestos a considerar que las clases son "independientes" entre sí, o como categórico ordinal. Una manera de incluir la ordinalidad de las clases puede ser utilizando binarias con codificación ordinal.

En el caso del MBD se obtienen diferentes resultados según el tratamiento de tales predictores, hecho que ilustramos la aplicación 1 del capítulo 5.

3.5.3. Aplicación actuarial

Se denomina *tarifa aditiva* a la que, partiendo de un grupo de tarifa base, va sumando (o restando) cantidades a éste para calcular la tarifa del resto de grupos contemplados. Y se denomina *tarifa multiplicativa* a la que, partiendo del grupo base, va multiplicando a éste por los tantos por ciento de incremento (o decremento) que correspondan para pasar al resto. La tarifa puede contemplar, o no, muchos grupos dependiendo del número de factores de riesgo incluidos y de su naturaleza.
En el ámbito del MLG, la tarifa aditiva la obtenemos a partir de un modelo que, combinado con cualquier distribución del error, utilice el enlace identidad, y la tarifa multiplicativa a partir de uno que

utilice el enlace logarítmico. Según (3.25), que nos relaciona la respuesta con el predictor lineal, tenemos que para el enlace identidad:

$$\mu_i = \eta_i = \beta_0 + \sum_{j=1}^{p} \beta_j f_{ij} \,. \tag{3.50}$$

Y que para el enlace logarítmico:

$$\log(\mu_i) = \log \eta_i = \log\left(\beta_0 + \sum_{j=1}^{p} \beta_j f_{ij} \right)$$

$$\mu_i = \exp\left(\beta_0 + \sum_{j=1}^{p} \beta_j f_{ij} \right) = e^{\beta_0} \times e^{\beta_1 f_{i1}} \times e^{\beta_2 f_{i2}} \times \cdots \times e^{\beta_p f_{ip}} \,. \tag{3.51}$$

Observamos que, si partimos de la base β_0 ó e^{β_0}, en la aditiva debemos sumar (o restar) las cantidades $\sum_{j=1}^{p} \beta_j f_{ij}$, y que en la multiplicativa debemos multiplicar por las cantidades de incremento (o decremento) $e^{\beta_1 f_{i1}} \times e^{\beta_2 f_{i2}} \times \cdots \times e^{\beta_p f_{ip}}$. En el caso de predictores binarios, partiendo del efecto global, sumaremos β_j o multiplicaremos por e^{β_j}, sólo cuando para la póliza *i* se dé la característica *j*, para $j = 1, \ldots, p$, ya que en tal caso se cumplirá $f_{ij} = 1$.

Suponer una tarifa multiplicativa conlleva dos ventajas importantes en comparación con una aditiva [Brockman y Wright (1992), apéndice B]:

- Por una parte implica "la hipótesis de independencia entre las variables utilizadas" o lo que es equivalente que "no existen términos de interacción" o que "no hay asociación entre las variables" [Zehnwirth (1994, pp. 619)]. Esto nos es de utilidad a la hora de realizar procesos de selección tan sólo con los efectos principales de los factores, ya que se obtiene una estimación ajustada de la siniestralidad sin necesidad de interacciones. De este modo la interpretación del modelo resultante es también más sencilla.

- Adicionalmente, hay evidencias empíricas de que las estimaciones obtenidas tienden a ser positivas, sea cual sea la distribución del error empleada, a diferencia de lo que ocurre con la aditiva.

Los primeros artículos sobre modelización estadística en el seguro del automóvil se centraban especialmente en el número de siniestros prestando menos atención a las cuantías. En los inicios, el debate principal estaba basado en si se debía utilizar un modelo aditivo o uno multiplicativo para

establecer la relación entre el número medio de siniestros y los factores de riesgo. Actualmente, el MLG es considerado como una formalización estadística potente que permite una amplia gama de posibilidades.

Históricamente, en el campo actuarial, se han utilizado especialmente dos casos particulares del MLG: el modelo aditivo clásico de enlace identidad y de error Normal [Lemaire (1977,1979,1985)]; y el modelo multiplicativo, de enlace logarítmico, combinado con una distribución de Poisson [Zehnwirth (1994)]. En este sentido, en Zehnwirth (1994), encontramos al MLG presentado, por una parte, como una extensión del modelo clásico de regresión lineal, y por otra, como la formalización probabilística motivada en los inicios por el modelo de Bailey y Simons [Bailey y Simons (1960); Bailey (1963)], entre otros.

En el capítulo 2, vimos que la cuantía total de los siniestros en términos esperados para cada póliza la calculamos como el producto del número esperado de siniestros por póliza, por la cuantía esperada de un siniestro. Aunque el objetivo es el cálculo de la cuantía total, nos interesa buscar explícitamente los factores que influyen en la cuantía de un siniestro y en el número de siniestros por póliza separadamente, pues predecimos ambas variables mediante dos regresiones independientes.

Cuando tratamos con el número de siniestros, nos encontramos con un átomo importante en cero, cosa que también ocurre con la cuantía total. Para paliar este problema se suele trabajar con datos agregados. Para ello consideraremos que todos los factores son categóricos (discretizando de una manera adecuada a los continuos) y agregaremos la información en la tabla cruzada de todos ellos. Pasaremos a tener tantas observaciones diferentes como el producto del número de clases de todos los factores menos las combinaciones vacías. Ahora, en clases donde la mayoría eran ceros individualmente, el valor que resume a la variable aleatoria de la celda sube debido a que habremos incorporado algunos siniestros. Aunque lo deseable es poder utilizar la información desagregada, en el caso de analizar el número de siniestros es más adecuado agregarla, es preferible la pérdida de información al átomo de distorsión.

Una vez tenemos la información agregada podemos trabajar con el número de siniestros directamente, teniendo en cuenta que la estimación se corresponderá con el número esperado de siniestros para el conjunto de pólizas de la celda. O bien, trabajar con la *frecuencia de siniestralidad*, que es lo usual, es decir, con el número medio de siniestros por póliza en cada celda. Para el tratamiento de la frecuencia de siniestralidad, ponderamos los datos en el modelo con el número de pólizas con las que se han

realizado las medias. Vamos a ver en este apartado con detalle su tratamiento respecto al MLG [Brockman y Wright (1992)].

Nomenclatura:

n = total de pólizas de la cartera

m = total de celdas en las que hemos agregado

e_u = expuestos o número de pólizas de la celda u para $u = 1,...,m$, por lo que $\sum_{u=1}^{m} e_u = n$

N_u = variable aleatoria número de siniestros de la celda u para $u = 1,...,m$

n_u = dato sobre el número de siniestros empírico de la celda u para $u = 1,...,m$. Realización de N_u

S_u = variable aleatoria coste total de los siniestros de la celda u para $u = 1,...,m$

s_u = dato sobre el coste total empírico de los siniestros de la celda u para $u = 1,...,m$. Realización de S_u

N_{iu} = variable aleatoria número de siniestros de la póliza i de la celda u para $i = 1,\ldots,e_u$, $u = 1,...,m$, por lo que $N_u = \sum_{i=1}^{e_u} N_{iu}$

X_{iu} = variable aleatoria coste del siniestro i de la celda u para $i = 1,\ldots,N_u$, $u = 1,...,m$, por lo que $S_u = \sum_{i=1}^{N_u} X_{iu}$

Empíricamente entonces,

- La frecuencia de siniestralidad o número medio de siniestros por póliza en la celda u, y la correspondiente ponderación, w_u, las obtenemos como:

$$\frac{n_u}{e_u}, \quad w_u = e_u$$

- La media del coste de un siniestro para las pólizas de la celda u, y la correspondiente ponderación, w_u, las obtenemos como:

$$\frac{s_u}{n_u}, \quad w_u = n_u$$

Teniendo en cuenta que la clasificación u se refiere en cada regresión al cruce de los factores utilizados. Después de dos procesos de selección, uno para el número y otro para las cuantías, obtendremos las variables de tarifa asociadas en cada caso, que pueden o no coincidir.

3.5.3.1. Número de siniestros

La distribución básica utilizada para modelizar el número de siniestros es la de Poisson. Como ya hemos indicado anteriormente trabajamos con datos agregados por celda u. Respecto de la siniestralidad en cada celda podemos considerar en primer lugar la situación ideal que se corresponde con las tres hipótesis siguientes:

H1) Hipótesis de Poisson: suponemos que el número de siniestros de cada póliza, N_{iu}, para $i=1,\ldots,e_u$ y $u=1,\ldots,m$, se distribuye Poisson: $N_{iu} \sim \text{Poisson}(\lambda_{iu})$

H2) Hipótesis de homogeneidad: suponemos que hemos utilizado los factores de riesgo de tal forma que nos ofrecen una homogeneidad perfecta dentro de las celdas, es decir, $\lambda_{iu}=\lambda_u$ para $i=1,\ldots,e_u$ en cada celda u, por lo que: $N_{iu}=NI_u \sim \text{Poisson}(\lambda_u)$ para $i=1,\ldots,e_u$.

H3) Hipótesis de independencia: suponemos que las pólizas son independientes.

A partir de las hipótesis anteriores se desprende lo siguiente,

- *Número de siniestros por póliza*, N_{iu}, para $i=1,\ldots,e_u$ y $u=1,\ldots,m$, se distribuye Poisson:

$$N_{iu} \sim \text{Poisson}(\lambda_u) \text{ con } E[N_{iu}]=Var[N_{iu}]=\lambda_u$$

  Terminología del MLG:

$$\mu_{iu}=E[N_{iu}]=\lambda_u \quad V(\mu_{iu})=\mu_{iu} \qquad \phi=1 \qquad w_{iu}=1 \tag{3.52}$$

- *Número de siniestros por celda*, $N_u=\sum_{i=1}^{e_u} NI_{iu}$, $u=1,\ldots,m$, se distribuye también exactamente Poisson:

$$N_u \sim Poisson(e_u \cdot \lambda_u) \text{ con } E[N_u]=Var[N_u]=e_u \cdot \lambda_u$$

  Terminología del MLG:

$$\mu_u=E[N_u]=e_u\lambda_u \qquad V(\mu_u)=\mu_u \qquad \phi=1 \qquad w_u=1 \tag{3.53}$$

- Dividiendo por e_u construimos una nueva variable aleatoria: $Y_u=\frac{N_u}{e_u}$, a la que llamamos *frecuencia de siniestralidad*. Para ella tenemos que

$$E[Y_u] = \lambda_u \qquad Var[Y_u] = \frac{\lambda_u}{e_u}$$

Esta variable aleatoria, Y_u, es tal que, excepto por el peso e_u que es una cantidad conocida, la esperanza es igual a la varianza. En la terminología del MLG Y_u tiene una estructura de error Poisson con pesos *a priori* $w_u = e_u$ y parámetro de escala $\phi = 1$:

$$\mu_u = E[Y_u] = \lambda_u \qquad V(\mu_u) = \mu_u \qquad \phi = 1 \qquad w_u = e_u \tag{3.54}$$

En la práctica no se cumple estrictamente ninguna de las tres hipótesis iniciales. Vamos a ver como afecta el no cumplimiento de cada una de ellas.

H2) Hipótesis de homogeneidad

Supongamos que relajamos la hipótesis de homogeneidad, es decir, λ_{iu} no tiene porqué coincidir con λ_u. A partir de las propiedades de las variables aleatorias condicionadas, tenemos que:

$$E[N_{iu}] = E\left[E(N_{iu} \mid \lambda_{iu})\right] = E[\lambda_{iu}]$$

$$Var[N_{iu}] = E\left[Var(N_{iu} \mid \lambda_{iu})\right] + Var\left[E(N_{iu} \mid \lambda_{iu})\right] = E[\lambda_{iu}] + Var[\lambda_{iu}]$$

Cuando las clases son homogéneas λ_{iu} es constante e igual a λ_u, entonces $E[\lambda_{iu}] = \lambda_u$ y $Var[\lambda_{iu}] = 0$. En tal caso $N_{iu} \mid \lambda_u \sim \text{Poisson}(\lambda_u)$, y reproducimos la situación ideal (3.54).

Para el caso heterogéneo $Var[N_{iu}] > E[N_{iu}]$ y el modelo tiene sobredispersión. Actuarialmente es usual recoger la sobredispersión considerando que λ_{iu} es una variable aleatoria. Suelen hacerse dos supuestos:

a) que λ_{iu} sigue una distribución Gamma

b) que λ_{iu} sigue una distribución Gaussina Inversa

a) Si λ_{iu} sigue una distribución Gamma con media y varianza:

$$E[\lambda_{iu}] = \lambda_u \qquad Var[\lambda_{iu}] = \frac{\lambda_u^2}{h_u},$$

el número de siniestros de una póliza cualquiera elegida aleatoriamente de la celda u, $N_{iu} = NI_u$ para $i = 1,\ldots,e_u$, tiene una distribución Binomial Negativa[21] con media y varianza:

$$E[NI_u] = \lambda_u \qquad Var[NI_u] = \left(1 + \frac{\lambda_u}{h_u}\right) \cdot \lambda_u$$

N_u es la suma de e_u variables aleatorias mixtas de Poisson con la misma variable de mixtura, por lo que la suma es también una variable aleatoria mixta de Poisson (en concreto Binomial Negativa) de esperanza la suma de las esperanzas y varianza la suma de las varianzas:

$$E[N_u] = e_u\lambda_u \qquad Var[N_u] = \left(1 + \frac{\lambda_u}{h_u}\right)(e_u\lambda_u)$$

Para poder utilizar los datos agregados ponderados de frecuencia de siniestralidad, definimos al igual que en el caso Poisson una nueva variable aleatoria, $Y_u = \frac{N_u}{e_u}$, para la que

$$E[Y_u] = \lambda_u \qquad Var[Y_u] = \left(1 + \frac{\lambda_u}{h_u}\right)\frac{\lambda_u}{e_u}$$

En este caso esta nueva variable aleatoria, es tal que, excepto por el peso e_u que es una cantidad conocida, tiene los dos primeros momentos como los de una distribución Binomial Negativa. En la terminología del MLG Y_u tiene una estructura de error Binomial Negativa con pesos *a priori* $w_u = e_u$ y parámetro de escala $\phi = 1$:

$$\mu_u = E[Y_u] = \lambda_u \qquad V(\mu_u) = \left(1 + \frac{\mu_u}{h_u}\right)\mu_u \qquad \phi = 1 \qquad w_u = e_u \tag{3.55}$$

Podemos rescribir la varianza de la frecuencia de siniestralidad como $Var[Y_u] = \phi_u \frac{\lambda_u}{e_u}$, donde $\phi_u = \left(1 + \frac{\lambda_u}{h_u}\right)$, de forma que la heterogeneidad provoca respecto del caso homogéneo de Poisson dos efectos: los parámetros de escala son diferentes entre las celdas y en su conjunto son mayores que 1.

Siguiendo a Brockman y Wright (1992) podemos prescindir del hecho de que ϕ_u dependa de u y trabajar razonablemente con la hipótesis de un parámetro de escala constante, ϕ, en una estructura

[21] La demostración puede encontrarse en manuales básicos como por ejemplo Panjer y Willmot (1992).

de error de Poisson con pesos *a priori* $w_u = e_u$, y parámetro de escala constante $\phi > 1$. Es decir, trabajaremos con el caso Poisson sobredisperso:

$$\mu_u = E[Y_u] = \lambda_u \qquad V(\mu_u) = \mu_u \qquad \phi > 1 \qquad w_u = e_u \tag{3.56}$$

b) Si λ_{iu} sigue una distribución Gaussiana Inversa con media y varianza:

$$E[\lambda_{iu}] = \lambda_u \qquad Var[\lambda_{iu}] = \frac{\lambda_u^3}{h_u}$$

El número de siniestros de una póliza cualquiera elegida aleatoriamente de la celda u, $N_{iu} = NI_u$ para $i = 1,\ldots,e_u$, sigue una distribución Poisson-Gaussiana Inversa[22] con media y varianza:

$$E[NI_u] = \lambda_u \qquad Var[NI_u] = \left(1 + \frac{\lambda_u^2}{h_u}\right)\lambda_u$$

N_u es la suma de e_u variables aleatorias mixtas de Poisson con la misma variable aleatoria de mixtura, por lo que la suma es también una variable aleatoria mixta de Poisson (en concreto Poisson-Gaussiana Inversa) de esperanza la suma de las esperanzas y varinaza la suma de las varianzas:

$$E[N_u] = e_u\lambda_u \qquad Var[N_u] = \left(1 + \frac{\lambda_u^2}{h_u}\right)(e_u\lambda_u)$$

Para poder utilizar los datos agregados de frecuencia de siniestralidad, definimos de nuevo $Y_u = \frac{N_u}{e_u}$, para la que

$$E[Y_u] = \lambda_u \qquad Var[Y_u] = \left(1 + \frac{\lambda_u^2}{h_u}\right)\frac{\lambda_u}{e_u}$$

En este caso esta nueva variable aleatoria, es tal que, excepto por el peso e_u que es una cantidad conocida, tiene los dos primeros momentos de una distribución Poisson-Gaussiana Inversa. En la terminología del MLG Y_u tiene una estructura de error Poisson-Gaussiana Inversa con pesos *a priori* $w_u = e_u$ y parámetro de escala $\phi = 1$:

$$\mu_u = E[Y_u] = \lambda_u \qquad V(\mu_u) = \left(1 + \frac{\mu_u^2}{h_u}\right)\mu_u \qquad \phi = 1 \qquad w_u = e_u \tag{3.57}$$

[22] De nuevo la demostración puede encontrarse en Panjer y Willmot (1992).

Del mismo modo que en el caso a) rescribimos la varianza de la frecuencia de siniestralidad como $Var[Y_u] = \phi_u \frac{\lambda_u}{e_u}$, donde $\phi_u = \left(1 + \frac{\lambda_u^2}{h_u}\right)$, y operamos en la práctica considerando un $\phi_u = \phi > 1$, con el caso Poisson sobredisperso (3.56).

En el límite, cuando $h_u \to \infty$, ambos casos van a parar al caso Poisson homogéneo (3.54).

Estos dos casos, Poisson-Gammma y Poisson-Gaussiana Inversa, se corresponden con el modelo más general en el que componemos una función de varianza especial, (3.29), que no surge como caso particular de la familia exponencial (3.20). Observamos que adicionalmente la función de varianza depende de los parámetros h_u para $u = 1,\ldots,m$. No es sencillo decidir con qué valores vamos a operar. Si suponemos que el parámetro es constante para todas las celdas, h, una posibilidad es realizar inferencia utilizando la versión extendida de las cuasi-verosimilitudes (3.35) sobre el parámetro, y quedarnos con el que nos ofrezca una menor desvianza [Renshaw (1994)]. Pero en la práctica es demasiado complejo. Tal y como ya hemos dicho, se puede trabajar sin pérdida de generalidad con el caso Poisson sobredisperso. En tal caso, una estimación del parámetro de dispersión, $\hat{\phi}$, la podemos obtener calculando (3.36), que realiza un promedio con las desvianzas del modelo, o calculando (3.37) que realiza un promedio con los residuos de Pearson en lugar de los de desvianza.

H1) Hipótesis de Poisson

En la práctica la hipótesis de Poisson para el número de siniestros por póliza N_{iu} también puede ser violada. Si suponemos que cada póliza genera los siniestros mediante un proceso de Poisson, estamos suponiendo que la intensidad se mantiene constante para todo el período de observación. Pero, después de la ocurrencia de un siniestro, la intensidad del riesgo suele disminuir, bien porque el vehículo está en reparación durante un tiempo, o bien porque el conductor circula con mayor precaución. Por lo tanto la intensidad no se mantiene constante para el período tal y como es requerido.

Contemplemos el caso extremo en el que, después de un siniestro, la intensidad del riesgo es cero para el resto del período. En tal caso, N_{iu} siempre será 0 ó 1. Supongamos que cada póliza de la celda u o bien tiene un siniestro con probabilidad p_u, o bien no tiene siniestros, lo que ocurre con probabilidad complementaria $1 - p_u$. Así, $N_{iu} = NI_u$ es una variable aleatoria Bernouilli. Entonces, siendo las

pólizas independientes, $N_u = \sum_{i=1}^{e_u} NI_u$ sigue una distribución Binomial(e_u, p_u) con media y varianza:

$$\mu_u = E[N_u] = e_u p_u \qquad Var[N_u] = (1-p_u) e_u p_u$$

Por lo que la frecuencia de siniestralidad, $Y_u = \frac{N_u}{e_u}$, tiene esperanza y varianza:

$$E[Y_u] = p_u \qquad Var[Y_u] = (1-p_u)\frac{p_u}{e_u}$$

En este caso extremo, en la terminología del MLG, Y_u tiene una estructura de error Binomial con pesos *a priori* $w_u = e_u$ y parámetro de escala $\phi = 1$:

$$\mu_u = E[Y_u] = p_u \qquad V(\mu_u) = (1-p_u)p_u \qquad \phi = 1 \qquad w_u = e_u \tag{3.58}$$

Podemos rescribir al modelo como uno de estructura del error de Poisson con pesos *a priori* $w_u = e_u$, y con parámetro de escala no constante reinterpretado como un factor de decrecimiento $\phi_u = (1-p_u)$. Puesto que p_u será una cantidad pequeña, las diferencias entre las probabilidades de las celda serán también pequeñas, por lo que podemos suponer razonablemente un parámetro de escala constante a lo largo de las celdas y casi cercano a 1. Deducimos de ello, que el caso Binomial sale como un caso infradisperso, $\phi < 1$, del Poisson ponderado:

$$\mu_u = E[Y_u] = \lambda_u \qquad V(\mu_u) = \mu_u \qquad \phi < 1 \qquad w_u = e_u \tag{3.59}$$

La distribución Binomial ha recibido poca atención en el contexto actuarial. Ello es debido, creemos, a su interpretación. Además del indicado, otro enfoque de la distribución Binomial es el siguiemte:

Supongamos que p_u es la probabilidad de que una póliza de la celda *u* tenga al menos un siniestro, y $1-p_u$ la probabilidad correspondiente a que no tenga siniestros. En este caso, siendo las pólizas independientes, N_u sigue también una distribución Binomial. Con esta interpretación N_u no nos indica realmente el número de siniestros, sino más bien el número de pólizas que sufren algún siniestro. Beirlant, Derveaux, de Meyer, Goovaerts, Labie y Maenhoudt (1991) estudian este caso trabajando directamente sobre N_u, en combinación con el enlace "log-odds",

$$\eta_u = \log\left(\frac{\mu_u}{e_u - \mu_u}\right) = \log\left(\frac{p_u}{1-p_u}\right) = \sum_{j=1}^{P} f_{uj}\beta_j \ ,$$

en aplicación al seguro del automóvil con datos de Bélgica.

H3) Hipótesis de independencia

La hipótesis de independencia entre riesgos no tiene porqué cumplirse [Brockman y Wright (1992)]. Usualmente en un accidente se ven envueltos varios vehículos, por lo que los riesgos no son independientes. Pero una cartera concreta está compuesta tan sólo por una porción de todos esos riesgos, por lo que este efecto puede considerarse casi despreciable. Notamos que en la hipótesis de independencia no incluimos el efecto que pueden suponer por ejemplo unas condiciones meteorológicas desfavorables, ya que este hecho haría subir las medias de siniestralidad, λ_u, simultáneamente, y aquí hablamos de la independencia mutua de las variaciones aleatorias referentes a las medias individuales λ_{iu} para todo i y para todo u.

Detallando más, supongamos que los asegurados puedan verse envueltos en los mismos accidentes. Esto significa respecto del número de siniestros que: $\mathrm{cov}\left(N_{iu}, N_{ju}\right) > 0 \quad \forall i \text{ y } \forall u$, y por tanto $Var\left(N_u\right) > \sum_{i=1}^{e_u} Var\left(N_{iu}\right)$. Así, $Var\left(N_{iu}\right) = \phi \cdot \lambda_u \Rightarrow Var\left(N_u\right) > \phi \cdot e_u \cdot \lambda_u$. Este hecho de nuevo lo interpretamos para la frecuencia de siniestralidad, $Y_u = \frac{N_u}{e_u}$, como que el parámetro de escala desconocido sea incrementado, y como consecuencia estamos ante el caso sobredisperso (3.59). Pero el incremento del parámetro de dispersión, ϕ, achacado a esta causa para una cartera determinada es pequeño, casi despreciable como ya se ha comentado.

Respecto a la función enlace en el caso Poisson

Lo usual es combinar la distribución de error Poisson con el enlace canónico logarítmico, el cual nos ofrece una tarifa multiplicativa, pero siempre podemos optar por modelizar con cualquier otro derivado de la familia de enlaces paramétricos (3.26).

3.5.3.1.1. Test de dispersión en el caso Poisson

Acabamos de ver que la validez del modelo Poisson (3.54) para la frecuencia de siniestralidad puede ser violada por diferentes causas:

- Por no cumplirse la hipótesis de Poisson, **H1**: que en el caso extremo va a parar a la distribución Binomial, y cuyo incumplimiento implica infra dispersión en el modelo
- Por no cumplirse la hipótesis de homogeneidad, **H2**: en este caso la varianza es mayor que la media y estamos ante una sobre dispersión (si modelizamos la aleatoriedad de las medias surgen los casos Poisson-Gamma y Poisson-Gaussiana Inversa)
- Por no cumplirse la hipótesis de independencia, **H3**: se refleja en una sobredispersión que en la mayoría de los casos, puede considerarse casi despreciable

Esto se refleja en el parámetro de dispersión. Una posibilidad razonable es realizar el supuesto de que el parámetro es común a todas las celdas. Siempre podemos obtener una estimación del parámetro de dispersión común a partir de (3.36) ó de (3.37). Tal estimación puede salir mayor o menor que 1. Puesto que es una estimación común para las celdas esto no implica, lógicamente, que para todos los perfiles ocurra la misma infra o sobre dispersión estimada. Tal estimación puede servirnos de guía pero en ningún caso nos informará exactamente a qué es debida.

Greene (1999) pp. 806-808 presenta tres estadísticos formales para contrastar la dispersión en el contexto del modelo Poisson: uno basado en un modelo de regresión, otro basado en un contraste de momentos condicionales, y un tercero basado en multiplicadores de Lagrange de un modelo alternativo. El procedimiento más sencillo en la práctica es el primero, el cual presentamos a continuación.

Contraste de dispersión en el modelo de Poisson basado en un modelo de regresión

$$H_0 : Var[y_i] = E[y_i]$$
$$H_1 : Var[y_i] = E[y_i] + \alpha g(E[y_i])$$

consiste en regresar $z_i = \frac{(y_i - \hat{\mu}_i)^2 - y_i}{\hat{\mu}_i \sqrt{2}}$, (con $\hat{\mu}_i$ la predicción de la i-ésima observación utilizando el modelo Poisson), o bien sobre la variable constante, o bien sobre $\hat{\mu}_i$ (sin utilizar entonces término

constante). Para contrastar H_0 frente a H_1, basta contrastar si el coeficiente resultante de la regresión es significativo.

Encontramos un ejemplo de su uso en Greene (1999) pp. 806, sobre unos datos de McCullagh y Nelder (1989). Se obtiene que no se rechaza la hipótesis nula, i.e., que los datos no presentan sobre o infra dispersión. Sin embargo McCullagh y Nelder afirman que son sobredispersos basándose simplemente en que la desviación típica de **Y** es 1.3 veces su media.

Cabe notar que **Y** será o bien el número de siniestros N_{iu} para datos desagregados, o bien N_u para datos agregados, o bien la frecuencia de siniestralidad $Y_u = N_u / e_u$.

Siguiendo a Albrecht (1983a)

Supongamos el MLG $\lambda_u = F(\beta_0 + \beta_1 + \cdots + \beta_r)$, realizado sobre la variable aleatoria frecuencia de siniestralidad, Y_u, con estructura de error Poisson con pesos *a priori* $w_u = e_u$ y parámetro de escala constante $\phi = 1$, (3.54), que utiliza $r+1 \leq m$ predictores, y la función de enlace $g = F^{-1}$. Según Albrecht la validez de tal modelo puede ser violada porque:

1) El número de siniestros individual, N_{iu}, no se distribuye Poisson para alguna celda, o bien todos los N_{iu} siguen una Poisson, pero las celdas no son homogéneas, es decir, no todos los e_u de la celda u tienen una media $\lambda_{iu} = \lambda_u$. Esto es, la situación ideal descrita en el apartado anterior deja de cumplirse violando la hipótesis de Poisson, **H1**, o la hipótesis de homogeneidad, **H2**.

2) La función de enlace utilizada, $g = F^{-1}$, no es la apropiada.

Para contrastar 1) propone el test de varianza de Fisher, que es un test de homogeneidad. Este test es superior al test de ajuste *ji*-cuadrado de bondad del ajuste para la distribución de Poisson. Para la clase u, $u = 1,\ldots,m$, el estadístico es el siguiente:

$$d(u) = \sum_{i=1}^{e_u} \frac{\left(N_{iu} - \bar{N}_u\right)^2}{\bar{N}_u} \sim \chi^2_{e_u - 1} \tag{3.60}$$

donde $\bar{N}_u = \dfrac{N_u}{e_u}$. Observamos que la homogeneidad de una clase tan sólo puede ser contrastada de este modo si se dispone de información desagregada.

Para contrastar 2), una primera posibilidad es validar el modelo comparando el número total de siniestros empíricos de cada clase con el número esperado a partir del modelo Poisson a contrastar, por lo que propone el estadístico:

$$Q_d = \sum_{u=1}^{m} \frac{\left(N_u - e_u \hat{\lambda}_u\right)^2}{e_u \hat{\lambda}_u} = \sum_{u=1}^{m} e_u \frac{\left(\bar{N}_u - \hat{\lambda}_u\right)^2}{\hat{\lambda}_u} \sim \chi^2_{m-(r+1)} \tag{3.61}$$

Observamos que este test sólo requiere información agregada.

Una investigación más detallada haciendo uso de información individual o desagregada sería como sigue. Definimos:

$$Q = \sum_{u=1}^{m} \sum_{i=1}^{e_u} \frac{\left(N_{iu} - \hat{\lambda}_u\right)^2}{\hat{\lambda}_u} \sim \chi^2_{n-(r+1)} \tag{3.62}$$

entonces, si el valor de Q es significativamente grande, lo podemos achacar a 1) ó a 2). Ahora lo dividimos en dos estadísticos independientes que se aproximan a *ji*-cuadrados:

$$Q = Q_w + Q_d = \sum_{u=1}^{m} \sum_{i=1}^{e_u} \frac{\left(N_{iu} - \bar{N}_u\right)^2}{\hat{\lambda}_u} + \sum_{u=1}^{m} e_u \frac{\left(\bar{N}_u - \hat{\lambda}_u\right)^2}{\hat{\lambda}_u}$$

La heterogeneidad puede ser contratada aproximadamente vía Q_w de manera conjunta para todas las celdas:

$$Q_w = \sum_{u=1}^{m} \sum_{i=1}^{e_u} \frac{\left(N_{iu} - \bar{N}_u\right)^2}{\hat{\lambda}_u} \sim \chi^2_{n-m} \tag{3.63}$$

Respecto al punto 2), éste puede ser también testado a partir de:

$$F = \frac{Qd \big/ m-(r+1)}{Qw \big/ n-m} \sim F_{(m-(r+1),n-m)} \tag{3.64}$$

de manera análoga que en el análisis de la varianza para la validez de un modelo de regresión.

Ahora bien Albrecht indica que si queremos testar dos enlaces alternativos:

$$\lambda_u = F_1(\beta_0, \beta_1, \ldots, \beta_r)$$
versus
$$\lambda_u = F_2(\beta_0, \beta_1, \ldots, \beta_r)$$

no existe nada formal, pero siempre podemos calcular Q_d para cada enlace y compararlos. Y ya como extremo comenta que, para un detalle mucho mayor, podríamos estudiar qué enlace es mejor para cada clase u, $u = 1, \ldots, m$.

3.5.3.2. Cuantía por siniestro

En el caso de las cuantías por siniestro no vamos a tener ningún inconveniente en utilizar la información desagregada. Tal y como ya indicamos en el apartado 3.5.1, basándonos en el rango de las cuantías, nos va a interesar utilizar distribuciones de rango positivo y con asimetría positiva como la Gamma o la Gaussiana Inversa frente a la Normal. Para ello podemos hacer uso de la familia paramétrica

$$V(\mu_{iu}) = \mu_{iu}^{\zeta} \tag{3.65}$$

para la distribución del error. Respecto a la función de enlace podemos utilizar, al igual que para el número de siniestros algún enlace derivado de la familia paramétrica (3.26).

Observamos que con la familia paramétrica de distribuciones del error (3.65) obtenemos:

$\Rightarrow$ para $\zeta = 2$ la estructura de error Gamma

$\Rightarrow$ para $\zeta = 3$ la estructura de error Gaussiana Inversa

Jørgensen (1987) demuestra que para $\zeta \geq 2$ obtenemos distribuciones de rango positivo y con asimetría positiva que es lo que nos interesa para las cuantías. Notamos que para $\zeta = 0$ obtenemos la estructura de error Normal, y para $\zeta = 1$ la estructura de error Poisson.

Así, si analizamos la variable aleatoria cuantía de un siniestro, X_{iu} $\forall i$ y $\forall u$:

$$\mu_{iu} = E[X_{iu}] = m_u \qquad V(\mu_{iu}) = \mu_u^{\zeta} \qquad \hat{\phi} \qquad w_i = 1 \tag{3.66}$$

Y si analizamos el coste medio de los siniestros por celda, $Y_u = \frac{S_u}{n_u}$:

$$\mu_u = E[Y_u] = m_u \qquad V(\mu_u) = \mu_u^{\zeta} \qquad \hat{\phi} \qquad w_u = n_u \tag{3.67}$$

Combinado en cada caso con algún enlace procedente de la familia (3.26).

3.5.4. Software utilizado

Para los cálculos del capítulo 5 referentes al MLG se ha hecho uso de dos programas:

- El programa S-PLUS 2000 (http://www.splus.com). Este programa permite bastantes combinaciones de error y enlace. Incluye el modelo general que utiliza las cuasi-verosimilitudes permitiendo especificar el parámetro ζ de la función de varianza de la familia paramétrica (3.65). El programa soporta un número elevado de casos.

- El paquete glmlab (http://www.sci.usq.edu.au/staff/dunn/glmlab/glmlab.html), complemento para MATLAB. Este paquete permite especificar adicionalmente el parámetro de dispersión, tanto para el caso Poisson como para el caso Binomial. El usuario puede predeterminar uno constante, igual o diferente de 1, y sino el programa lo estima mediante la desvianza media. Además contempla la familia de enlaces paramétricos (3.26), pudiendo especificar el usuario la λ deseada. Con MATLAB no tenemos restricción en el número de casos.

Hay otros paquetes estadísticos y lenguajes de programación posibles que incluyen el tratamiento del MLG, por ejemplo:

LEM (http://www.kub.nl/faculteiten/fsw/organisatie/departementen/mto/software2.html); MINITAB; STATGRAPHICS; BMDP; STATA; STATISTICA; SPSS; SAS; la NAG de Fortran con actualmente la versión GLIM Release 4.1 (http://www.nag.co.uk/); GENSTAT;... Los más completos son el SAS y el GLIM. Del resto, hay algunos que están limitados o en los posibles modelos a estimar (como pueden ser sólo toda la gama de modelos log-lineales) o en los resultados a mostrar (análisis de los residuos o resultados gráficos).

De cara a las compañías aseguradoras son necesarios programas capaces de computar ágilmente con grandes volúmenes de datos. Hace unos años el programa más asociado al campo actuarial era EMBLEM. De éste, Michael J. Brockman y Tom Wright realizaron una presentación en el *General Insurance Convention & ASTIN Colloqium* de 1998 celebrado en Glasgow. Pero en la actualidad existen muchas más alternativas de paquetes construidos por empresas especializadas o consultoras. Suele utilizarse el programa SAS como base por ser una alternativa potente que permite la programación de algoritmos haciendo uso de las funciones incorporadas. Por ejemplo, la consultora Tillinghast-Tower Perrin vende el paquete *Tscore v 8.0*. Éste está programado para realizar el estudio de tarificación a partir de una base datos en formato SAS, haciendo uso tan sólo del modelo Poisson para el número de siniestros y del Gamma para las cuantías, ambos combinados con el enlace logarítmico. Además incorpora otras particularidades necesarias para la tarificación como son la agrupación de zonas a partir de los datos estudiados y la discretización previa de los factores continuos. Finalmente el programa permite guardar un informe de los resultados obtenidos en *excel*. El problema es que se presenta como un programa independiente y por lo tanto cerrado a las amplias alternativas que permite el SAS.

3.6. Criterios de discretización de variables continuas

Nosotros denominamos *discretización de variables continuas*, al cambio de escala de intervalo[23] a ordinal[24]. Éste suele ser el cambio de escala más utilizado. Se trata de partir de una variable cuantitativa, dividirla en clases según algún criterio y convertirla en categórica ordinal. Una vez se tienen las clases construidas, se suelen asignar unas puntuaciones ordinales a las clases, por ejemplo la media aritmética de los valores continuos de la clase o el punto medio de la clase, los cuales sustituyen a los valores originales. El motivo principal es el de poder utilizar técnicas estadísticas que como *input* necesitan de variables categóricas.

Cuando discretizamos una variable continua se nos plantean dos problemas cruciales: el *número de clases* en que dividiremos a la variable, y la *amplitud de los intervalos* de las clases.

[23] Indiferentemente intervalo o ratio.
[24] En general, para cambios de escala, no se distingue entre continua y discreta numérica, como es el caso del número de siniestros, pues siempre partimos de un número finito de puntos.

Respecto al número de clases *k*

Si partimos de n observaciones de la variable original continua, algunos autores proponen los siguientes criterios para decidir sobre k:

- Siguiendo a Gutiérrez, Rodríguez y Santos (1995) pp. 22-35:

$$k \approx 2\sqrt{n}\text{, o bien más aconsejable } k \approx 1+3.3\ln(n).$$

- Siguiendo a Neter y Wasserman (1970) pp. 207-213:

  k debería estar entre 4 y 20, teniendo en cuenta que escogeremos un número mayor de clases cuanto mayor sea el número de observaciones n.

- Siguiendo a Domènech (1977) pp. 231-275: a título indicativo aconseja el siguiente cuadro,

Número de observaciones, n:	Número de clases, k:
8	4
16	5
32	6
64	7
128	8
256	9
512	10
1024	11
2048	12

- Siguiendo a van der Laan (1988) pp. 196-199:

$$\text{Si } \alpha = \text{nivel de significación dado, } \alpha = \int_{-\infty}^{d} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} dx\text{, entonces}$$

$$k = 2\sqrt{\frac{2(n-1)^2}{d^2}} \quad \text{para } n \geq 42.$$

Respecto a la amplitud de los intervalos

Podemos realizar amplitudes iguales o desiguales. En general será más correcto realizarlas desiguales para tener en cuenta la naturaleza de los datos.

Si queremos amplitudes iguales tan sólo hemos de dividir el rango de la variable entre el número de clases a formar. Pero puede ocurrir que en algunas clases tengamos pocos valores, o bien que el rango de variación de los valores de cada clase sea muy diferente.

Podemos evitar un número desigual de observaciones organizando clases equiprobables, es decir, clases con el mismo número de observaciones.

Podemos evitar que el rango de variación de los valores de cada clase sea desigual si utilizamos alguna técnica que como objetivo tenga la formación de clases con una mínima pérdida de información en la reducción de la escala. En Andenberg (1973) pp. 30-51 encontramos con detalle diferentes alternativas: métodos de análisis cluster jerárquicos (los basados en una función de distancias entre individuos y el método de Ward), no jerárquicos (la extensa gama de "k-means"), y métodos basados en análisis discriminante (la función discriminante lineal y el método de Cochran y Hopkins). En el trabajo estudiamos las alternativas de cluster, para el detalle del discriminante nos remitimos a la referencia citada.

Los métodos de análisis cluster no jerárquicos forman los grupos optimizando un funcional objetivo para un número k de clusters preespecificado. Concretamente nos son de interés unidimensionalmente los que como objetivo tienen alguna variedad de minimización de la varianza interna de los grupos a formar, como es toda la gama de "*k-means*".

Dentro de los métodos jerárquicos el de mayor interés es el método de Ward [Ward (1963); Ward y Hook (1963)]: éste empieza con n grupos, cada uno con una observación (observación a observación); en cada paso junta dos grupos cuya combinación da el menor incremento en la varianza dentro, SCD; continua hasta que ha juntado n-1 veces y por lo tanto sólo queda 1 grupo. Ésta técnica "tiende" a dar las particiones que minimizan SCD para cada número de grupos de n hasta 1. Con Ward no es necesario fijar previamente k, podemos observar las diferentes agrupaciones y quedarnos con la que nos interesa.

Según la situación que se pretende solventar, nos interesará aplicar una metodología u otra. Las dos situaciones concretas en el caso de la tarificación son:

a) Discretización de la respuesta continua (cuantía de un siniestro o cuantía total) en caso de utilizar CHAID (tanto SPSS CHAID nominal, CATFIRM, como SPSS CHAID ordinal), y

b) Discretización de predictores continuos para cualquiera de las técnicas de segmentación vistas en el trabajo. Además del caso en que se pretenda agregar la información disponible.

En el caso a), puesto que la respuesta continua es la experiencia de siniestralidad, nos va a interesar peder la mínima información. Para ello podemos utilizar algún método de cluster no jerárquico o el método de Ward, que es el que en el trabajo se recomienda, ambos unidimensionalmente. Para decidir cual es el número k de clases a formar, en principio sería adecuado escoger el número máximo de categorías que nos permita el algoritmo (31 para los programas del SPSS y 16 para los de D. M. Hawkins), pero hay que pensar que a medida que bajamos de nivel en el árbol, las tablas formadas para el cálculo de los p-valores, resultarán dispersas, así que debemos ir con cuidado para que los p-valores calculados tengan sentido.

En el caso b), nos va a interesar realizar la discretización teniendo en cuenta la relación del predictor con la experiencia de siniestralidad objeto de estudio. Por ejemplo, si tomamos la edad del conductor, será de interés separar al colectivo de 18 a 20 años (que no representará a lo mejor ni un 1% de la cartera) debido a su alta siniestralidad. Los métodos de cluster que utilizan un criterio externo son las técnicas de segmentación. Así, la propuesta es utilizar el algoritmo XAID, (que recordemos es de respuesta continua) con un solo predictor. El predictor deberá partir de una cierta discretización inicial lo más amplia posible, y a la hora de agrupar tales clases iniciales mediante XAID el predictor debería, en principio, definirse como monótono.

3.6.1. Ejemplo de discretización de respuesta continua

A modo de ejemplo del caso a), vamos a realizar la discretización de la respuesta de los datos descritos en el apartado 2.3.4 y que utilizamos en la aplicación 1 del capítulo 5. Se trata de 401 cuantías de siniestros. Nos interesa discretizarlas para poder realizar en dicha aplicación las correspondientes aplicaciones de CHAID. Respecto a los factores de riesgo, éstos ya son categóricos.

Primero decidiremos sobre el número de clases k:

- Siguiendo a Gutiérrez et al. (1995): $k = 2\sqrt{n} = 2\sqrt{401} \approx 40$, o bien más aconsejable

$k = 1 + 3.3\ln(n) = 1 + 3.3\ln(401) = 20.7 \approx 21$.

- Siguiendo a Neter y Wasserman (1970): k entre 4 y 20, teniendo en cuenta que n = 401, más bien 4.
- Siguiendo a Domènech (1977): unas 10 clases.
- Siguiendo a van der Laan (1988): si $\alpha = 0.05$, $0.05 = \int_{-\infty}^{d} \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\} dx \Rightarrow d = -1.64$,

entonces $k = 2\sqrt{\frac{2(n-1)^2}{d^2}} = 2\sqrt{\frac{2(401-1)^2}{(-1.64)^2}} = 20.7 \approx 21$.

Nos sale que debemos escoger entre aproximadamente 4 y 21 clases. Puesto que se pretende realizar un estudio de sensibilidad respecto a los resultados del AS según se agrupe a la respuesta, se procede a la discretización en 4, 10, 20 y 31 clases, haciendo uso del paquete estadístico SPSS. Se utilizan tres variedades de K-means no jerárquicos y el método de Ward, para confirmar que el método Ward, al menos empíricamente, es el que nos ofrecerá una mayor minimización de la varianza dentro de los grupos, lo que nos asegura una mayor confianza en la estructura de los datos resultantes. Si enumeramos:

1) Ward
2) K-mean: Iterar y clasificar actualizando la media
3) K-mean: Iterar y clasificar sin actualizar la media
4) K-mean: Sólo clasificar

Obtenemos los siguientes resultados de las SCE:

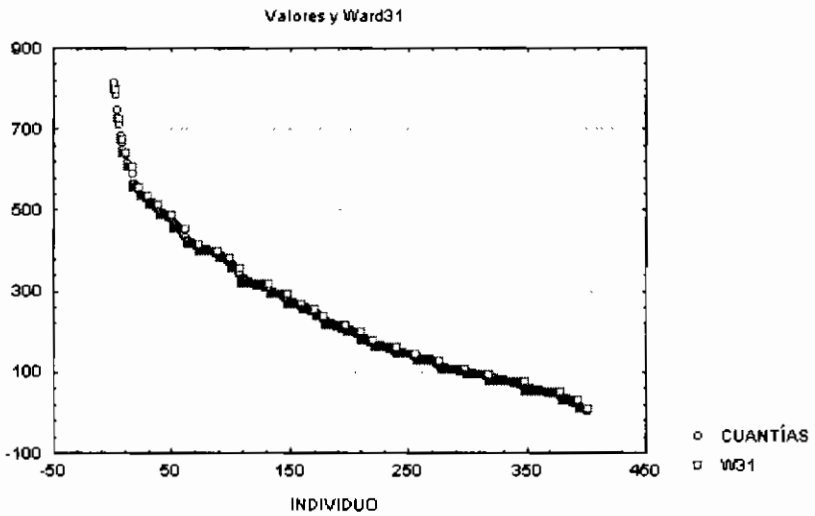
	Opción 1)	Opción 2)	Opción 3)	Opción 4)
4 clases	26957.08	**27061.14**	26826.81	20558.48
10 clases	**29148.08**	28046.24	27847.13	23516.97
20 clases	**29439.28**	29044.82	28943.41	28118.50
31 clases	**29495.96**	29203.96	29087.73	28144.99

La varianza de los datos o SCT es 29531.59, puesto que SCT = SCE + SCD, minimizar la SCD equivale a maximizar la SCE, por lo que nos quedaremos con la agrupación de máximo valor en SCE. Vemos que los máximos valores se obtienen aplicando el método de Ward, excepto para la agrupación

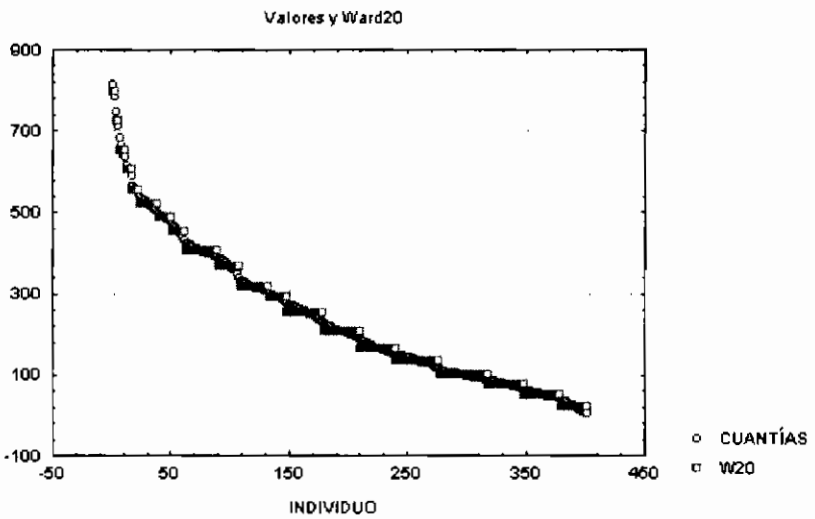
en 4 clases. En general, si no queremos realizar todas las posibilidades, daremos por válido que Ward es una buena opción.

Veamos gráficamente, al utilizar como puntuación las medias de los clusters obtenidos en cada caso, como a medida que disminuimos el número de clases la información también es menos precisa:

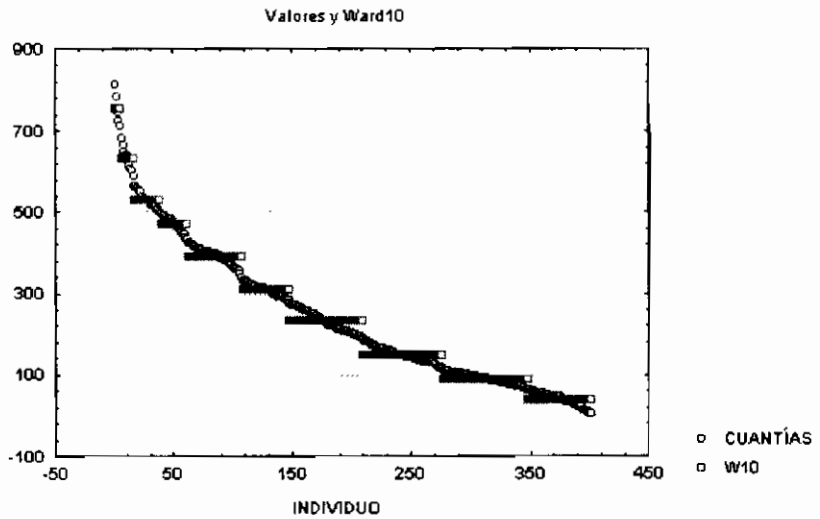
Para 31 clases:



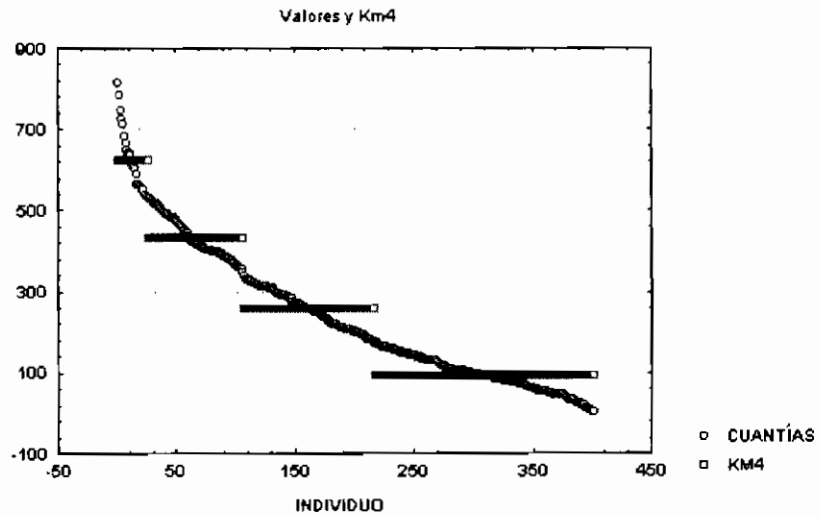
Para 20 clases:



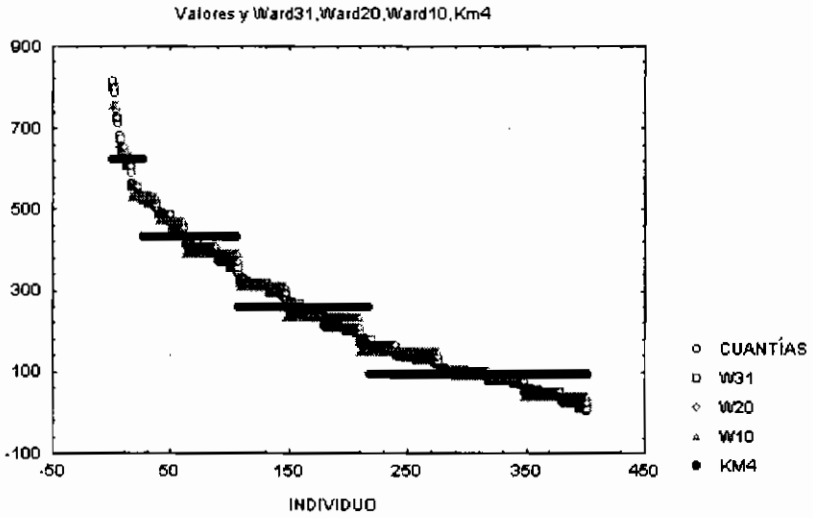
Para 10 clases:



Para 4 clases:



Y conjuntamente:



En la aplicación 1 hacemos uso de las discretizaciones: 4 clases del K-means opción 2), 10 y 31 clases de Ward. Elegimos 4 por ser un número bastante pequeño de clases, 10 para tener un intermedio soportable por los algoritmos disponibles y 31 por ser el número máximo que admiten los programas de SPSS, en especial el algoritmo ordinal. Descartamos la clasifieación de 20, simplemente porque FIRM sólo admite hasta 16 clases y no podríamos realizar comparaciones con CHAID nominal y CATFIRM.

3.6.2. Ejemplo de discretización de predictores continuos

A modo de ejemplo del caso b), realizamos la discretización de los predietores cuantitativos de los datos descritos en el apartado 2.3.2 que hacen referencia a la aplicación 2 del capítulo 5.

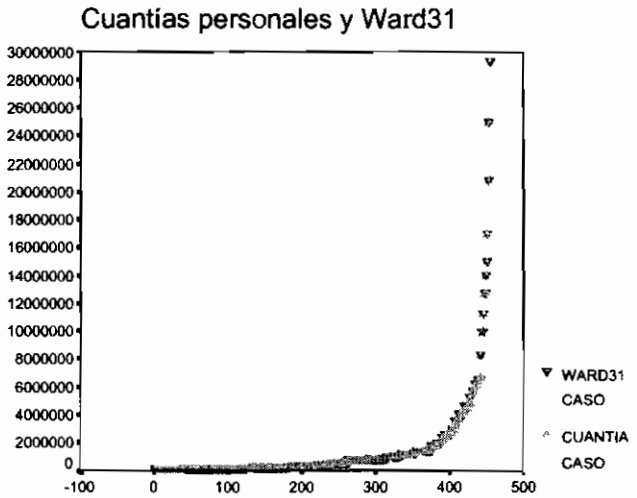
La potencia, la antigüedad del vehículo y el bonus/malus son variables cuantitativas discretas. El valor del vehículo, la antigüedad del carnet y la edad de conductor habitual son cuantitativas continuas. Discretizaremos las variables continuas, y aprovecharemos para agrupar las clases de las discretas con un número elevado de valores como son la potencia y la antigüedad del vehículo.

Previamente calculamos unos descriptivos que nos orientarán en la discretización:

Estadísticos descriptivos

	Rango	Mínimo	Máximo	Media	Desv. típ.
POTENCIA	254	32	286	76.59	27.32
ANTIVEHI	23	0	23	7.44	4.49
VALORVEH	10050	600	10650	1793.48	923.75
ANTICARN	38.02	2.75	40.78	20.1981	8.3150
Edad del cond habi	54.70	22.27	76.97	45.0760	10.6007
MALUS	.700	.000	.700	7.25E-03	5.15E-02

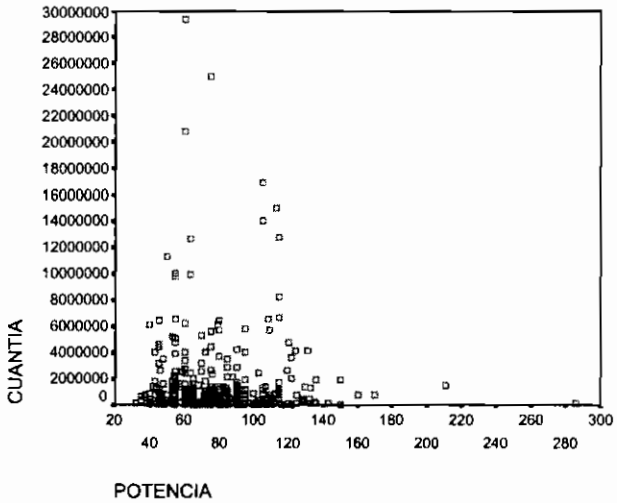
Para cualquiera de las discretizaciones que presentamos a continuación hemos utilizado el CHAID ordinal, procediendo a discretizar las cuantías haciendo uso del método de Ward para 31 categorías de igual modo que en el apartado anterior:



Respecto a los predictores realizamos una discretización inicial amplia de como mucho 31 categorías para cada uno; los definimos como monótonos; fijamos para la fase de agrupación de categorías un nivel de significación bastante amplio de 0.25, para que no nos junte las clases en exceso; no ponemos restricción para que el predictor sea elegido, así que el nivel de significación para la selección del mejor predictor es de 1; y finalmente exigimos un tamaño mínimo para la formación de un grupo tanto de tanto de 6 como de 25. En las tablas presentamos las medias de la discretización de Ward 31, y el número de siniestros correspondiente.

Potencia

La potencia es discreta con un rango de variación de 254. Presentamos un gráfico de su relación con las cuantías, que no nos sirve de mucha ayuda en la decisión de la agrupación inicial:

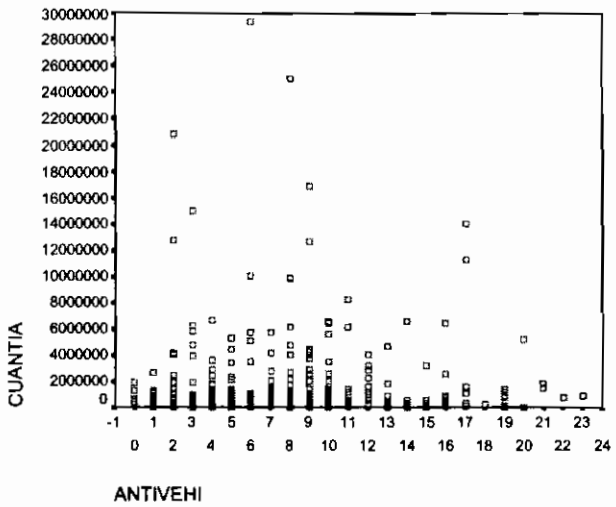


Decidimos construir intervalos iniciales de amplitud 5, obteniendo 16 intervalos. Tanto para el tamaño mínimo de 6 como de 25 obtenemos el siguiente resultado en la agrupación o reducción de clases:

Intervalos iniciales	Segmentación	Media y número
[0,35]	[0,80]	1342033.99 n = 293
[36,40]		
[41,45]		
[46,50]		
[51,60]		
[61,70]		
[71,80]		
[81,90]	[81,100]	863706.62 n = 81
[91,100]		
[101,110]	[101,120]	2086526.19 n = 55
[111,120]		
[121,130]	[121,...]	1049491.04 n = 26
[131,150]		
[151,200]		
[201,250]		
[250,...]		

Antigüedad del vehículo

La antigüedad del vehículo es discreta con un rango de 23, por lo que no tendremos problema en construir los intervalos iniciales de uno en uno:

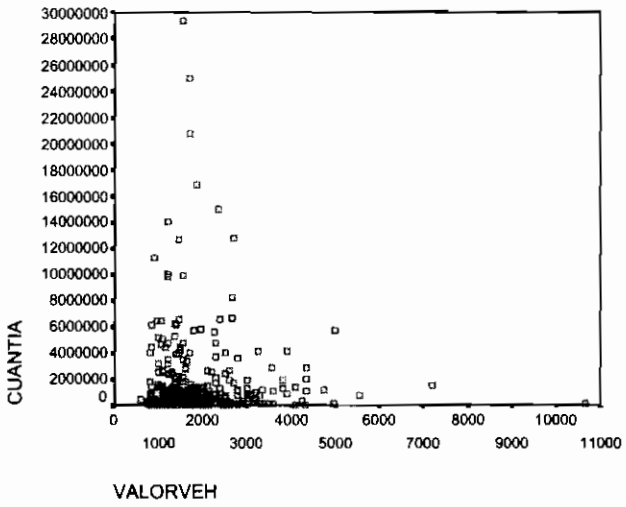


Obtenemos dos discretizaciones diferentes, una para el tamaño mínimo de 6 y otra para el de 25:

Valores iniciales	Segmentación 6	Media y número	Segmentación 25	Media y número
0 = [0,1)	[0,2)	660490.10 n = 30	[0,2)	660490.10 n = 30
1 = [1,2)				
2 = [2,3)	[2,4)	1911612.32 n = 52	[2,4)	1911612.32 n = 52
3 = [3,4)				
4 = [4,5)	[4,8)	1022097.78 n = 171	[4,8)	1022097.78 n = 171
5 = [5,6)				
6 = [6,7)				
7 = [7,8)				
8 = [8,9)	[8,13)	1681488.66 n = 139	[8,13)	1681488.66 n = 139
9 = [9,10)				
10 = [10,11)				
11 = [11,12)				
12 = [12,13)				
13 = [13,14)	[13,17)	831346.73 n = 41	[13,...)	1230122.45 n = 63
14 = [14,15)				
15 = [15,16)				
16 = [16,17)				
17 = [17,18)	[17,19)	3165043.29 n = 9		
18 = [18,19)				
19 = [19,20)	[19,20)	704344.39 n = 6		
20 = [20,21)	[20,...)	1528720.34 n =7		
21 = [21,22)				
22 = [22,23)				
23 = [23,...)				

Valor del vehículo

El valor del vehículo toma los siguientes valores continuos:

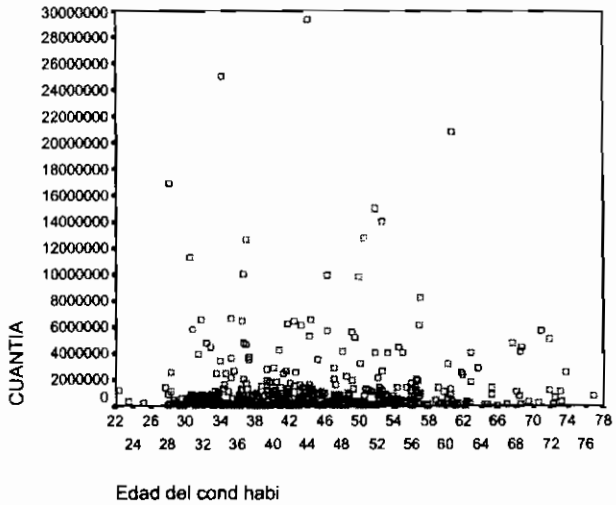


Hemos procedido a la formación de los intervalos iniciales que presentamos en la tabla. Y el resultado ha sido el mismo para los dos tamaños, 6 y 25:

Intervalos iniciales	Segmentación	Media y número
[0,1000)	[...,2000)	1359900.60 n = 329
[1000,1250)		
[1250,1500)		
[1500,1750)		
[1750,2000)		
[2000,2250)	[2000,2250)	533062.24 n =32
[2250,2500)	[2250,3000)	1754004.85 n = 55
[2500,3000)		
[3000,3500)	[3000,...)	1135547.40 n = 39
[3500,4000)		
[4000,4500)		
[4500,5000)		
[5000,6000)		
[6000,8000)		
[8000,10000)		
[10000,...)		

Edad del primer conductor

La edad del conductor principal toma los siguientes valores continuos:

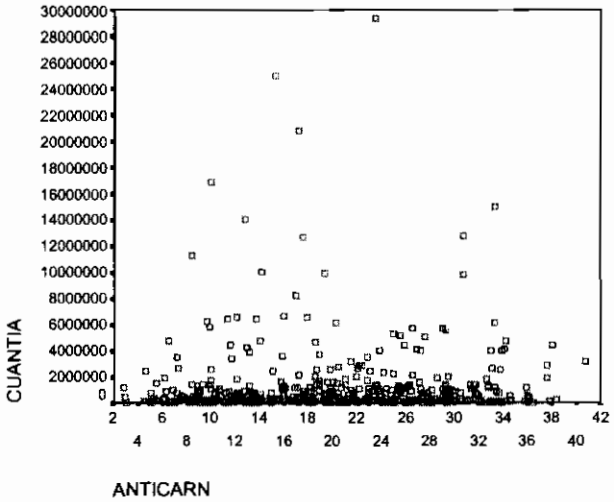


Hemos construido 27 intervalos iniciales, de aproximadamente 2 años de amplitud, excepto para el primero, que va de [18,25), ya que el mínimo valor es de 22.27 años, obteniendo el siguiente resultado:

Intervalos iniciales	Segmentación 6	Media y número	Segmentación 25	Media y número
[18,25)	[...,39)	1347181.84 n = 150	[...,39)	1347181.84 n = 150
[25,27)				
[27,29)				
[29,31)				
[31,33)				
[33,35)				
[35,37)				
[37,39)				
[39,41)	[39,41)	774632.56 n = 35	[39,41)	774632.56 n = 35
[41,43)	[41,47)	1580188.36 n = 78	[41,47)	1580188.36 n = 78
[43,45)				
[45,47)				
[47,49)	[47,49)	795166.81 n = 30	[47,49)	795166.81 n = 30
[49,51)	[49,53)	1632726.26 n = 57	[49,55)	1368207.94 n = 80
[51,53)				
[53,55)	[53,55)	712662.53 n = 23		
[55,57)	[55,...)	1456904.35 n = 82	[55,...)	1456904.35 n = 82
[57,59)				
[59,61)				
[61,63)				
[63,65)				
[65,67)				
[67,69)				
[69,71)				
[71,73)				
[73,75)				
[75,...)				

Antigüedad del carnet del primer conductor

Procedemos como en los otros casos:



Intervalos iniciales	Segmentación 6	Media y número	Segmentación 25	Media y número
[0,3)	[...,9)	902178.92 n = 47	[...,9)	902178.92 n = 47
[3,5)				
[5,7)				
[7,9)				
[9,11)	[9,19)	1640500.88 n = 156	[9,19)	1640500.88 n = 156
[11,13)				
[13,15)				
[15,17)				
[17,19)				
[19,21)	[19,23)	816282.02 n = 85	[19,23)	816282.02 n = 85
[21,23)				
[23,25)	[23,25)	2041638.98 n = 28	[23,25)	2041638.98 n = 28
[25,27)	[25,29)	951173.62 n = 57	[25,29)	951173.62 n = 57
[27,29)				
[29,31)	[29,31)	1624486.33 n = 31	[29,33)	1330538.18 n = 51
[31,33)	[31,33)	874918.55 n = 20		
[33,35)	[33,35)	2257948.87 n = 19	[33,...)	1879908.04 n = 31
[35,37)	[35,37)	419516.59 n = 6		
[37,...)	[37,...)	2143170.16 n = 6		

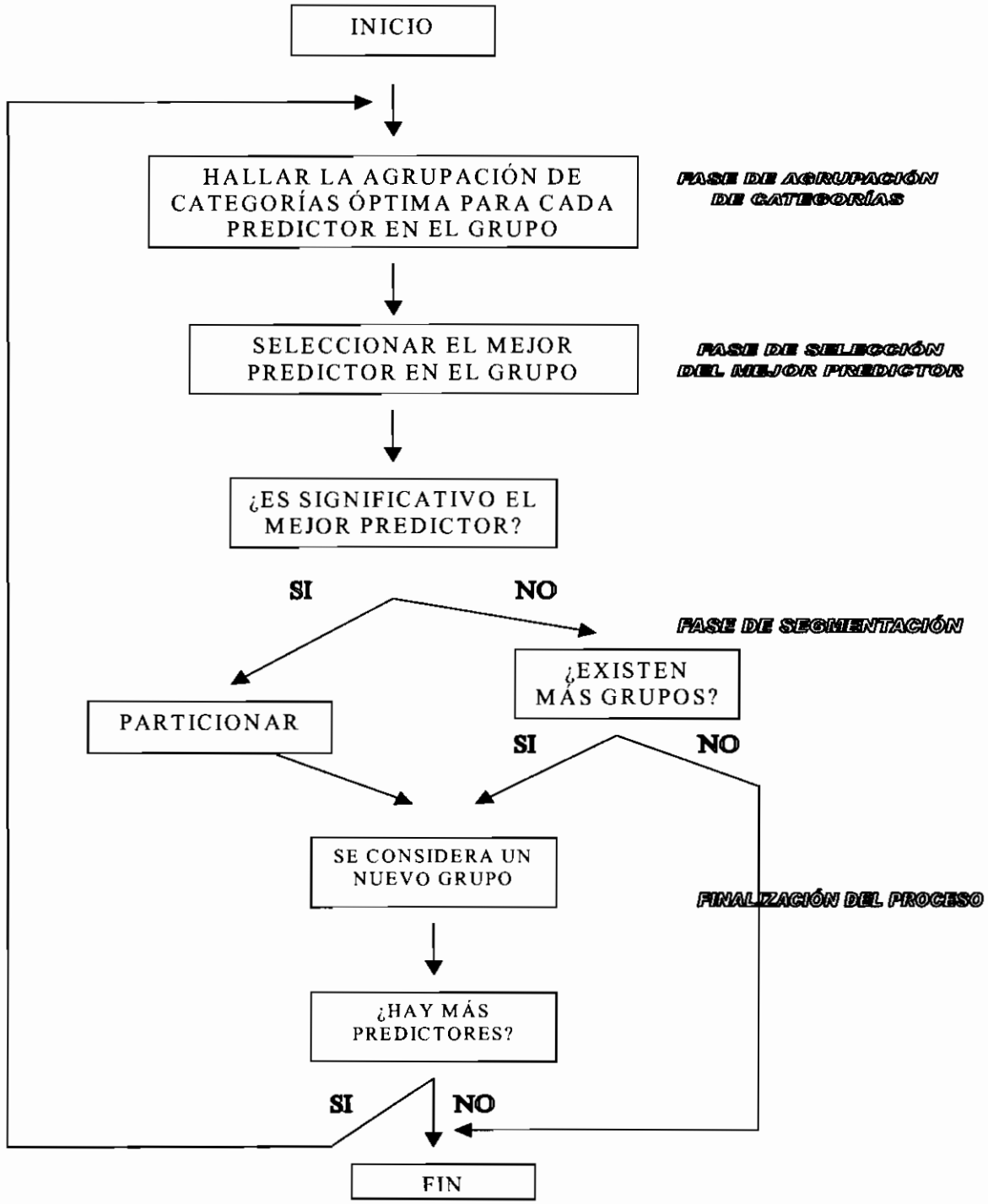
Una vez tenemos las discretizaciones, procedemos a realizar las anovas de las cuantías con cada una de ellas, utilizando los datos originales en lugar de la discretización de Ward para 31 categorías que hemos plasmado en las tablas anteriores. En la siguiente tabla detallamos la medida de asociación (3.4) y los *p*-valores de dichas anovas que encontramos en el anexo 3.3:

η = p-valor =	**Discretización inicial**	**Para un mínimo de 6**	**Para un mínimo de 25**
Potencia	0.139 0.850	0.116 **0.105**	0.116 **0.105**
Valor del Vehículo	0.134 0.836	0.092 **0.282**	0.092 **0.282**
Antigüedad del Vehículo	0.195 0.802	0.167 **0.078**	0.167 0.107
Edad	0.153 0.997	0.102 **0.581**	0.082 0.695
Antigüedad del carnet	0.173 0.758	0.163 0.208	0.114 **0.154**

Observamos que las η de las discretizaciones iniciales, para cualquiera de las variables, son mayores que para las dos segmentaciones, la de 6 y la de 25. Esto es debido al elevado número de clases iniciales. Sin embargo observamos, como era de esperar, que los p-valores de las anovas realizadas con las discretizaciones iniciales, son en cualquier caso mayores. Hemos resaltado en negrilla la discretización que ofrece un p-valor menor para cada variable: Potencia6y25, Valorveh6y25, Antiveh6, Edad6 y Anticarn25.

Éstas nuevas variables cualitativas, son las que deberíamos incluir en el análisis del MLG y de la RBD de la aplicación 2, si en lugar de tratamiento cuantitativo (que es el que hemos dado en la aplicación) decidiésemos dar tratamiento cualitativo a estos predictores. Sin embargo, para la correcta aplicación del AS debemos utilizar las discretizaciones iniciales de los predictores para dejar que el algoritmo junte en cada nivel las clases más adecuadas.

ANEXO 3.1. Diagrama de flujo del algoritmo CHAID



ANEXO 3.2. Tablas de propiedades para casos particulares del MLG

	Normal o Gaussiana $N(\mu,\sigma^2)$	Binomial $B(m,\pi)/m$	Poisson $P(\mu)$	Gamma $G(\mu,\nu)$	Gaussiana inversa $GI(\mu,\sigma^2)$
Rango de y	$(-\infty,+\infty)$	0,1,..,m/m	0,1,2,...,∞	$(-\infty,+\infty)$	$(-\infty,+\infty)$
Peso w	1	1	1	1	1
Parámetro de dispersión: ϕ	σ^2	$1/m$	1	υ^{-1}	σ^2
$b(\theta)$	$\frac{\theta^2}{2}$	$\log(1+e^{\theta})$	$\exp(\theta)$	$-\log(-\theta)$	$-(-2\theta)^{1/2}$
$c(y;\theta)$	$-\frac{1}{2}\left(\frac{y^2}{\phi}+\log(2\pi\phi)\right)$	$\log\binom{m}{my}$	$-\log y!$	$\nu\log(\nu y)-\log(y)$ $-\log\Gamma(\nu)$	$-\frac{1}{2}\left(\log(2\pi\phi y^3)+\frac{1}{\phi y}\right)$
$\mu(\theta)=E(Y;\theta)$	θ	$\frac{e^{\theta}}{(1+e^{\theta})}$	$\exp(\theta)$	$\frac{-1}{\theta}$	$(-2\theta)^{-1/2}$
Función de enlace canónica: $\theta(\mu)$	Identidad: μ	Logit: $\log(\mu/(1-\mu))$	Logarítmico: $\log(\mu)$	Recíproco: $1/\mu$	$1/\mu^2$
Función de varianza: $V(\mu)$	1	$\mu(1-\mu)$	μ	μ^2	μ^3

Tabla 3.2. Principales características.

Distribución:	Desvianzas D:
Normal o Gaussiana $N(\mu,\sigma^2)$	$\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2$
Binomial $B(m,\pi)/m$	$2\sum_{i=1}^{n}\{y_i\log(y_i/\hat{\mu}_i)+(m-y_i)\log[(m-y_i)/(m-\hat{\mu}_i)]\}$
Poisson $P(\mu)$	$2\sum_{i=1}^{n}\{y_i\log(y_i/\hat{\mu}_i)-(y_i-\hat{\mu}_i)\}$
Gamma $G(\mu,\nu)$	$2\sum_{i=1}^{n}\{-\log(y_i/\hat{\mu}_i)+(y_i-\hat{\mu}_i)/\hat{\mu}_i\}$
Gaussiana inversa $GI(\mu,\sigma^2)$	$\sum_{i=1}^{n}(y_i-\hat{\mu}_i)^2/(\hat{\mu}_i^2 y_i)$

Tabla 3.3. Desvianzas.

Anexo 3.3. Anovas de las cuantías con los factores discretizados

Potencia discretizada

Descriptivos

CUANTIA

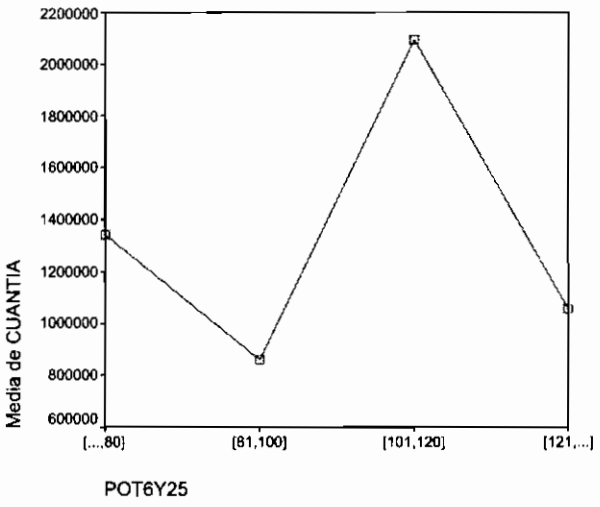
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[...,80]	293	1341088	3084649.24	180207.13	986418.73984	1695758	13080.000	2.9E+07
[81,100]	81	861754.8	1075041.39	119449.04	624043.66853	1099466	13080.000	5825920
[101,120]	55	2092298	4023061.72	542469.53	1004712.071	3179884	13080.000	1.7E+07
[121,...]	26	1054020	1243088.28	243789.67	551925.27726	1556114	30000.000	4167255
Total	455	1330158	2906476.99	136257.74	1062383.896	1597932	13080.000	2.9E+07

ANOVA

CUANTIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	5.174E+13	3	1.72E+13	2.056	.105
Intra-grupos	3.783E+15	451	8.39E+12		
Total	3.835E+15	454			

Gráfico de las medias



Antigüedad del vehículo discretizada

Descriptivos

CUANTIA

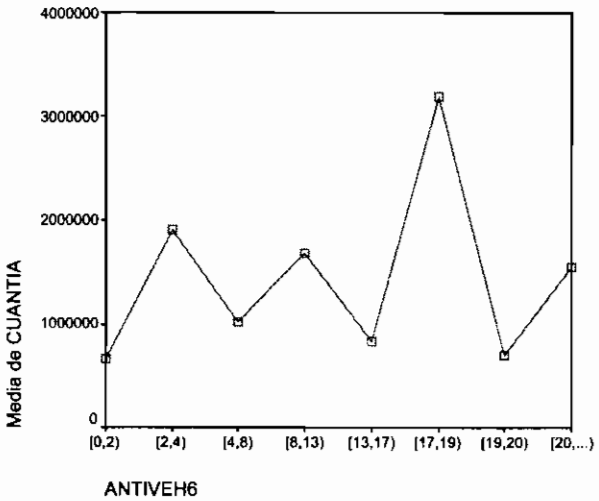
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[0,2)	30	663459.5	723562.859	132103.90	393276.68822	933642.3	13080.000	2680920
[2,4)	52	1903743	3967976.44	550259.33	799051.34425	3008435	13080.000	2.1E+07
[4,8)	171	1021970	2572218.78	196702.49	633676.14006	1410264	13080.000	2.9E+07
[8,13)	139	1679566	3217962.51	272944.02	1139872.633	2219259	13080.000	2.5E+07
[13,17)	41	837205.3	1589472.94	248233.97	335505.70414	1338905	13080.000	6589303
[17,19)	9	3190512	5426653.37	1808884.5	-980782.813	7361807	100000.0	1.4E+07
[19,20)	6	703457.5	547605.517	223559.02	128780.75395	1278134	20000.000	1424960
[20,...)	7	1549452	1714631.58	648069.82	-36317.72674	3135222	34800.000	5228106
Total	455	1330158	2906476.99	136257.74	1062383.896	1597932	13080.000	2.9E+07

ANOVA

CUANTIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1.075E+14	7	1.54E+13	1.841	.078
Intra-grupos	3.728E+15	447	8.34E+12		
Total	3.835E+15	454			

Gráfico de las medias



Valor del vehículo discretizada

Descriptivos

CUANTIA

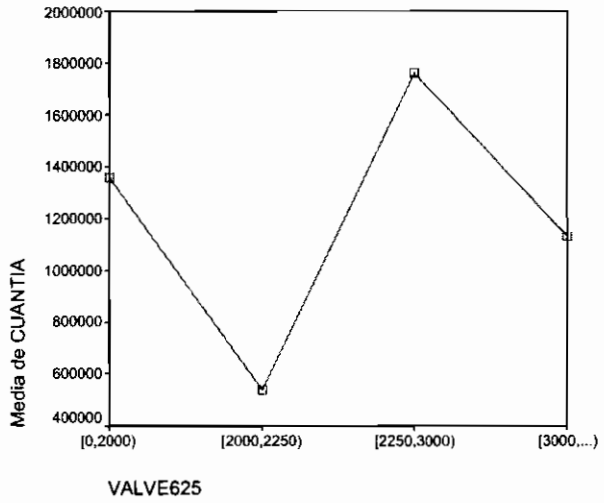
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[0,2000)	329	1359204	3132121.06	172679.42	1019505.558	1698903	13080.000	2.9E+07
[2000,2250)	32	535420.9	671421.509	118691.68	293347.60680	777494.1	13080.000	2635730
[2250,3000)	55	1761056	3068285.41	413727.52	931582.60877	2590530	13080.000	1.5E+07
[3000,...)	39	1129540	1286084.33	205938.31	712639.76580	1546440	22238.000	5730520
Total	455	1330158	2906476.99	136257.74	1062383.896	1597932	13080.000	2.9E+07

ANOVA

CUANTIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	3.227E+13	3	1.08E+13	1.276	.282
Intra-grupos	3.803E+15	451	8.43E+12		
Total	3.835E+15	454			

Gráfico de las medias



Edad discretizada

Descriptivos

CUANTIA

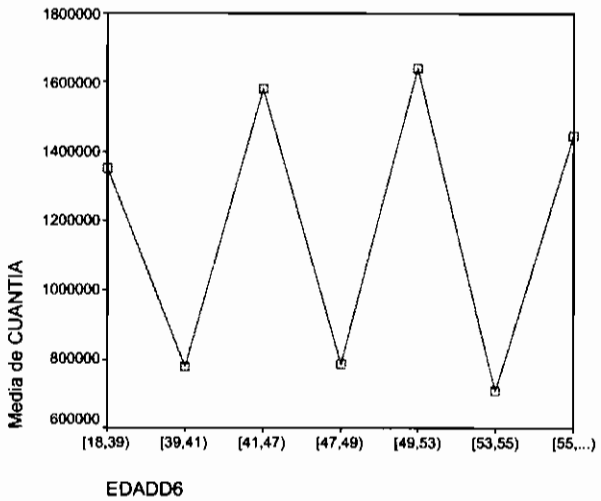
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[18,39)	150	1351091	3079604.87	251448.68	854225.55858	1847957	13080.000	2.5E+07
[39,41)	35	778334.5	984836.732	166467.79	440031.20314	1116638	23000.000	4231688
[41,47)	78	1581147	3695456.24	418428.02	747950.56537	2414344	20000.000	2.9E+07
[47,49)	30	786845.3	940730.226	171753.06	435570.86070	1138120	13080.000	4167255
[49,53)	57	1639891	3391130.34	449166.19	740102.60138	2539679	21000.000	1.5E+07
[53,55)	23	707846.7	1152325.15	240276.40	209543.88831	1206149	37080.000	4415614
[55,...)	82	1446676	2718792.78	300240.44	849291.23594	2044060	13080.000	2.1E+07
Total	455	1330158	2906476.99	136257.74	1062383.896	1597932	13080.000	2.9E+07

ANOVA

CUANTIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	3.998E+13	6	6.66E+12	.787	.581
Intra-grupos	3.795E+15	448	8.47E+12		
Total	3.835E+15	454			

Gráfico de las medias



Antigüedad del carnet discretizada

Descriptivos

CUANTIA

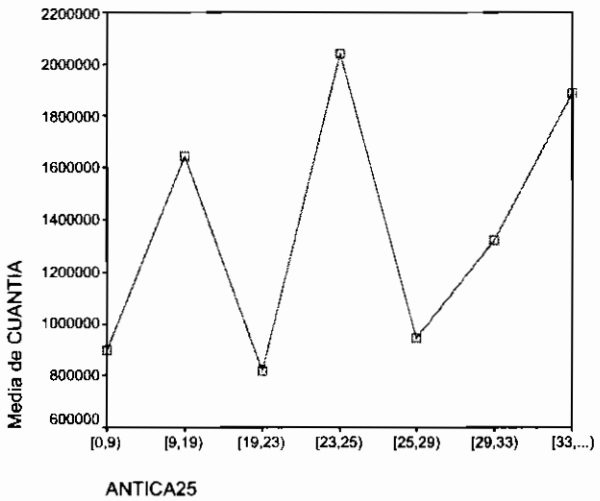
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[0,9)	47	897812.4	1820786.41	265588.99	363209.50870	1432415	13080.000	1.1E+07
[9,19)	156	1644976	3548250.94	284087.44	1083793.091	2206158	13080.000	2.5E+07
[19,23)	85	817057.0	1425337.29	154599.54	509618.89298	1124495	13080.000	9887069
[23,25)	28	2043498	5490472.11	1037601.7	-85485.25846	4172480	20000.000	2.9E+07
[25,29)	57	943788.2	1426775.73	188981.06	565213.65320	1322363	18880.000	5722400
[29,33)	51	1322859	2385151.67	333988.04	652024.51708	1993694	13080.000	1.3E+07
[33,...)	31	1886418	2995833.24	538067.53	787537.81838	2985299	21000.000	1.5E+07
Total	455	1330158	2906476.99	136257.74	1062383.896	1597932	13080.000	2.9E+07

ANOVA

CUANTIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	7.898E+13	6	1.32E+13	1.570	.154
Intra-grupos	3.756E+15	448	8.38E+12		
Total	3.835E+15	454			

Gráfico de las medias



Capítulo 4

Metodología basada en distancias

Muchos métodos estadísticos y de análisis de datos utilizan el concepto geométrico de distancia entre individuos, entre poblaciones, y de un individuo a una población. Esto ocurre especialmente en técnicas de representación de datos (análisis de correspondencias, análisis de coordenadas principales, análisis de proximidades), donde la distancia como medida de diferenciación entre objetos constituye la base fundamental de la representación de los resultados. Las distancias, además, aparecen en muchos otros aspectos de la estadística matemática: contraste de hipótesis, estimación, regresión, análisis discriminante, etc.

Con el objetivo de proponer una herramienta estadística alternativa al resto de metodologías de selección de variables de tarifa, en este capítulo presentamos una metodología de selección de predictores en el modelo de regresión basada en distancias que permite trabajar directamente sobre factores potenciales de riesgo de tipo mixto. Esta regresión basada en distancias fue inicialmente planteada por Cuadras (1989b) y Cuadras y Arenas (1990), y posteriormente desarrollada en Cuadras, Arenas y Fortiana (1996), Cuadras y Fortiana (1998) y Fortiana y Cuadras (1998).

En el presente capítulo recogemos las principales características de la regresión basada en distancias y realizamos las siguientes aportaciones:

- Proponemos la versión ponderada de la regresión basada en distancias. La motivación es que como hemos visto en el capítulo anterior al modelizar el número de siniestros y la cuantía por siniestro se utilizan normalmente datos agregados y por lo tanto ponderados.
- Planteamos el proceso de selección de predictores. Para ello definimos las medidas y tests estadísticos apropiados para la regresión basada en distancias.
- No conocemos las distribuciones de los estadísticos de test para muestras finitas y, ciertamente, sería complicado obtenerlas, incluso aproximadamente. Por ello hemos optado por simularlas, empleando la metodología *bootstrap* que, como veremos, se adapta especialmente bien a las características peculiares de la regresión basada en distancias.

El presente capítulo se estructura en cinco apartados y dos anexos.

En el primer apartado 4.1, mencionamos algunas funciones de distancias entre individuos. Diferenciamos la naturaleza global del conjunto de variables según la escala de medida de las variables: cuantitativa, cualitativa o mixta.

En el apartado 4.2, describimos la regresión basada en distancias. Brevemente, consiste en proyectar la respuesta continua en el espacio euclídeo obtenido mediante escalado multidimensional métrico a partir del conjunto de predictores. La información aportada por los factores de riesgo queda reflejada en una matriz de distancias sobre la que se opera. La regresión basada en distancias es una extensión del modelo clásico de regresión: si la distancia empleada es ℓ^2 y los predictores son cuantitativos se obtiene como caso particular el modelo de regresión lineal por mínimos cuadrados ordinarios. Dividimos el apartado en 4 partes: primeramente describimos resultados básicos de escalado multidimensional métrico, necesarios para la construcción del modelo; pasamos luego a explicar la predicción basada en distancias en general, para centrarnos en el caso de la regresión; en tercer lugar exponemos algunos casos particulares de ésta bien conocidos; y finalmente explicamos como abordar en el modelo el tratamiento de los términos de interacción entre predictores.

Tal y como hemos visto en el capítulo 3, en la tarificación, es usual trabajar con datos agregados y por lo tanto ponderados, especialmente para el tratamiento del número de siniestros mediante la frecuencia de siniestralidad. Por ello, en el apartado 4.3, consideramos la regresión basada en distancias con datos ponderados. Este modelo permite tratar también el caso heteroscedástico. En el subapartado 4.3.1 presentamos resultados básicos del escalado multidimensional métrico ponderado. En 4.3.2 construimos el modelo, mostrando su consistencia con la regresión basada en distancias usual.

En el apartado 4.4, proponemos un método de selección para la regresión basada en distancias. Primeramente definimos las medidas y los tests estadísticos necesarios para la realización del proceso. Posteriormente construimos el proceso de selección paso a paso, de manera análoga a como lo hicimos en el capítulo 3 para el MLG. La estimación de los p-valores la realizamos con la metodología *bootstrap*, a partir de estimaciones por simulación de las distribuciones de probabilidad de los estadísticos de los tests. Los modelos basados en distancias son especialmente adecuados para el empleo de *bootstrap*, pues el hecho que todas las interdistancias entre individuos de un remuestreo aparezcan ya en la matriz de distancias inicial nos permite vectorizar los remuestreos mediante

matrices de multiplicidades, lo que es de gran economía computacional. Para la validación del modelo resultante empleamos diferentes criterios, incluidos los métodos de validación cruzada.

La metodología de selección que proponemos, además de cubrir la fase de selección de variables de tarifa, puede servir, si se desea, para completar la tarificación hasta la estimación de primas, al igual que ocurre con el MLG.

Finalmente, en el apartado 4.5, describimos los programas utilizados para la realización de los cálculos de las aplicaciones del capítulo 5. Éstos programas, cuyo código se encuentra en el anexo 4.2, están implementados con `octave`, por lo que, en principio, no tienen restricción en el número de pólizas, ni en el de factores de potenciales de riesgo.

Para terminar el apartado dedicado a la metodología basada en distancias, observamos que, puesto que la filosofía de la regresión basada en distancias es utilizar las variables latentes que determinan la configuración euclidea a modo de predictores, no disponemos de unos coeficientes de los factores directamente interpretables. Este tipo de problema también se plantearía en los modelos aditivos generalizados [Hastie y Tibshirani (1991)].

4.1. Distancias sobre matrices de datos

Una distancia δ sobre un conjunto (finito o no) Ω, es una aplicación que a cada par de individuos $(i,j)\in\Omega\times\Omega$, le hace corresponder un número real $\delta(i,j)=\delta_{ij}$, que como mínimo cumple las siguientes propiedades básicas: $\delta_{ij}\geq 0$, $\delta_{ii}=0$ y $\delta_{ij}=\delta_{ji}$. En tal caso podemos hablar de *disimilaridad*. Cuando, además, se cumple la desigualdad triangular, $\delta_{ij}\leq\delta_{ik}+\delta_{kj}$, y $\delta_{ij}=0$ si y solo si $i=j$, entonces la distancia es *métrica*.

Si se cumplen las propiedades básicas, la desigualdad triangular, y además podemos encontrar puntos $\mathbf{x}_i=(x_{i1},\ldots,x_{ir})$, $\mathbf{x}_j=(x_{j1},\ldots,x_{jr})$ de $\mathbb{R}^r$ tales que permiten reproducir las distancias originales:

$$\delta_{ij}^2=(\mathbf{x}_i-\mathbf{x}_j)(\mathbf{x}_i-\mathbf{x}_j)^T, \tag{4.1}$$

es decir, que siendo δ_{ij} la distancia ℓ^2 (4.2) entre los puntos $\mathbf{x}_i$, $\mathbf{x}_j$, (Ω,δ) puede representarse mediante el espacio euclídeo $(\mathbb{R}^r,\delta)$, entonces la distancia es *euclídea*.

Nos referimos a Cuadras (1989a) para una clasificación más detallada de las distancias según sus propiedades (distancia ultramétrica, distancia aditiva, divergencia,...). En el trabajo dedicamos especial atención a las distancias euclídeas.

Si Ω es un conjunto finito, que indicaremos como $\Omega=\{1,2,\ldots,n\}$, las distancias δ_{ij} se expresan mediante la matriz simétrica $\boldsymbol{\Delta}$, llamada matriz de distancias sobre Ω,

$$\boldsymbol{\Delta}=\begin{pmatrix}\delta_{11} & \delta_{12} & \cdots & \delta_{1n}\\ \delta_{21} & \delta_{22} & \cdots & \delta_{2n}\\ \cdots & \cdots & \cdots & \cdots\\ \delta_{n1} & \delta_{n2} & \cdots & \delta_{nn}\end{pmatrix} \qquad \delta_{ii}=0,\ \ \delta_{ij}=\delta_{ji}.$$

En el capítulo 3 vimos medidas de asociación lineal entre variables (dos a dos) dado un conjunto de individuos, ahora nos interesa cuantificar las asociaciones entre individuos (dos a dos) implicados en un conjunto de variables. Para ello hacemos uso del concepto de distancia o su dual similaridad.

4.1.1. Distancias sobre datos cuantitativos

Supongamos que cada individuo i de $\Omega=\{1,2,\ldots,n\}$ viene representado por un punto $\mathbf{x}_i=(x_1,x_2,\ldots,x_p)\in\mathbb{R}^p$. La distancia más familiar entre dos individuos i,j es la distancia ℓ^2:

$$d_2(i,j)=\sqrt{\sum_{k=1}^{p}(x_{ik}-x_{jk})^2} \tag{4.2}$$

tal distancia es un caso particular de las distancias ℓ^q de Minkowski

$$d_q(i,j)=\left[\sum_{i=1}^{p}\left|x_{ik}-x_{jk}\right|^q\right]^{1/q} \quad \text{con} \quad 1<q<\infty,$$

que son disimilaridades que verifican la propiedad triangular. No son distancias euclídeas, salvo en el caso $q = 2$. Para $q = 1$ se tiene la denominada distancia "ciudad" $d_1(i,j)=\sum_{k=1}^{p}\left|x_{ik}-x_{jk}\right|$, y en el límite, $d_\infty(i,j)=\max_k\left\{\left|x_{ik}-x_{jk}\right|\right\}$, la denominada distancia "dominante".

La distancia (4.2) tiene varios inconvenientes: no está acotada y no es invariante por cambios de escala, además si las variables implicadas no son estocásticamente independientes esta estructura no queda correctamente reflejada. Para solventarlos se han propuesto variadas modificaciones, nos referimos de nuevo a Cuadras (1989a, 2003) para una discusión extensa.

Una modificación simple consistiría en dividir por el número de variables:

$$d_2(i,j)=\sqrt{\frac{1}{p}\sum_{k=1}^{p}(x_{ik}-x_{jk})^2}.$$

Respecto a la invarianza por cambios de escala, la podemos resolver, por ejemplo, dividiendo cada sumando de la distancia por la correspondiente desviación típica, lo que nos lleva a la distancia de K. Pearson:

$$d_{ij} = \left(\sum_{k=1}^{q} \left(\frac{x_{ik} - x_{jk}}{s_k} \right)^2 \right)^{1/2}.$$

Respecto al supuesto de independencia entre las variables, una solución nos la proporciona la distancia de Mahalanobis,

$$d_m^2(i,j) = \left(\mathbf{x}_i - \mathbf{x}_j \right)^T \boldsymbol{\Sigma}^{-1} \left(\mathbf{x}_i - \mathbf{x}_j \right),$$

que tiene en cuenta las correlaciones entre variables, y por tanto la redundancia existente entre las mismas, además de ser invariante ante cambios de escala. Naturalmente, la podemos definir en poblaciones $(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, es decir, con vector de medias $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$, sin necesidad de normalidad en las poblaciones.

En la tabla 4.1 del anexo 4.1, extraida de Cuadras (1989a) y de Gower y Legendre (1986), presentamos las propiedades métrica y euclídea de algunas distancias. Entre paréntesis encontramos el no cumplimiento de la propiedad en el caso de contemplar valores negativos en las variables. En la tabla, r_k es un número positivo arbitrario que usualmente será el rango, G_k, o la desviación típica, s_k. Notamos que Gower (1971) demuestra que la distancia D3 es euclídea si r_k se corresponde con el rango:

$$d_{ij} = \left(\frac{1}{p} \sum_{k=1}^{p} \frac{\left| x_{ik} - x_{jk} \right|}{G_k} \right)^{1/2}. \tag{4.3}$$

Sobre los criterios que deben seguirse para la elección de la distancia, véase Gower y Legendre (1986).

4.1.2. Distancias sobre datos cualitativos

En muchas aplicaciones conviene trabajar con similaridades, concepto dual del de distancia. Una *similaridad s* sobre un conjunto Ω es una aplicación que a cada $(i,j) \in \Omega \times \Omega$ le hace corresponder un número real $s_{ij} = s(i,j)$ que cumple: $0 \leq s_{ij} \leq s_{ii} = 1$, $s_{ii} = 1$ y $s_{ij} = s_{ji}$.

Cuando Ω es un conjunto finito, tenemos la matriz de similaridades,

$$\mathbf{S} = \begin{pmatrix} s_{11} & s_{12} & \cdots & s_{1n} \\ s_{21} & s_{22} & \cdots & s_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ s_{n1} & s_{n2} & \cdots & s_{nn} \end{pmatrix} \qquad s_{ii} = 1, \ s_{ij} = s_{ji}.$$

La cantidad $s_{ij} = s(i,j)$ es una medida del grado de semejanza entre dos elementos i, j, en el sentido de que si ambos son muy parecidos, entonces s_{ij} se aproxima a 1. El concepto de similaridad es especialmente utilizado cuando sobre Ω se han introducido p características cualitativas, que se asocian a otras tantas variables binarias, que toman el valor 0 si la característica está ausente y el valor 1 si está presente.

Es inmediato pasar de similaridad a distancia y recíprocamente. Dos de las transformaciones básicas son:

$$\delta_{ij} = 1 - s_{ij} \tag{4.4}$$

$$\delta_{ij} = \sqrt{1 - s_{ij}} \ . \tag{4.5}$$

La última es más aconsejable, pues da lugar siempre a una distancia métrica, incluso euclídea para muchas de las similaridades estudiadas.

Pero, en general, una matriz de similaridades puede tener elementos $s_{ii} \neq 1$ en su diagonal, incluso no necesariamente estar comprendida entre 0 y 1. En tal caso se define una transformación más adecuada [Gower (1966)]:

$$\delta_{ij} = \sqrt{s_{ii} + s_{jj} - 2s_{ij}} \ . \tag{4.6}$$

Similaridades con variables binarias

Supongamos que disponemos de p variables binarias. Denominamos a, b, c, d a las frecuencias de $(1,1)$, $(1,0)$, $(0,1)$ y $(0,0)$, respectivamente, sumando $a + b + c + d = p$. Una similaridad s_{ij} es entonces una función de a, b y c.

El criterio a seguir para la elección de un coeficiente de similaridad depende del peso que se desee dar a las frecuencias a, b, c y d. En la tabla 4.2 del anexo 4.1 presentamos algunos coeficientes de similaridad, de los cuales detallamos: el rango, si la matriz resultante de similaridades, **S**, es semi-definida positiva, el cumplimiento o no de las propiedades métrica y euclídea en referencia a la distancia (4.6), y en su caso el autor.

Similaridades con variables cualitativas

Si disponemos de p variables cualitativas, a las cuales pretendemos dar la misma importancia o peso sin tener en cuenta el número de clases implicadas en cada variable, podemos definir coeficientes de coincidencias como sigue. Supongamos que al estudiar la similitud entre los individuo i, j, α_{ij} es el número de coincidencias para las p variables cualitativas, por lo que $p - \alpha_{ij}$ serán las no coincidencias, en tal caso podemos estudiar diferentes combinaciones. La más usual es el coeficiente de coincidencias:

$$s_{ij} = \frac{\alpha_{ij}}{p} \ . \tag{4.7}$$

Esta similaridad tiene un rango entre 0 y 1, y posee tanto la propiedad métrica como euclídea.

También es factible estudiar las propiedades de otras combinaciones como por ejemplo: $s_{ij} = \dfrac{\alpha_{ij}}{p - \alpha_{ij}}$,

$s_{ij} = \dfrac{\alpha_{ij} - (p - \alpha_{ij})}{p}$ y $s_{ij} = \dfrac{\alpha_{ij}}{\alpha_{ij} + \theta(p - \alpha_{ij})}$ en función de θ.

4.1.3. Distancias sobre datos mixtos

Supongamos que disponemos de un conjunto de variables de tipo mixto (mezcla de variables cuantitativas, binarias y cualitativas). En tal caso es apropiado tratar a los diferentes tipos de la manera que les corresponde. Varios autores han estudiado esta casuística [Estabrook y Rogers (1966); Gower (1971); Legendre y Chodorowski (1977)].

Un coeficiente apropiado para el tratamiento de datos mixtos es el coeficiente de similaridad propuesto por Gower (1971):

$$s_{ij} = \frac{\sum_{h=1}^{p_1} \left(1 - \left|x_{ih} - x_{jh}\right| / G_h\right) + a + \alpha_{ij}}{p_1 + (p_2 - d) + p_3} \tag{4.8}$$

donde p_1 es el número de variables cuantitativas, a y d son el número de coincidencias positivas y negativas respectivamente para las p_2 variables dicotómicas, y α_{ij} es el número de coincidencias para las p_3 variables cualitativas. G_h es el rango de la h-ésima variable cuantitativa.

Este coeficiente verifica $0 \le s_{ij} \le 1$, por lo que para pasar de similaridad a distancia podemos aplicar indiferentemente (4.5) ó (4.6). Adicionalmente admite la posibilidad de datos faltantes [Cuadras (2003)].

Observamos, que el coeficiente de similaridad de Gower no es más que la suma de diferentes coeficientes apropiados para cada tipo de variables. Por ejemplo:

- Si sólo disponemos de variables cuantitativas, utilizando (4.5), el coeficiente se reduce a la distancia (4.3).

- Si sólo disponemos de variables binarias, el coeficiente se reduce al coeficiente de Jaccard (Tabla 4.2 del anexo 4.1).

- Y si sólo disponemos de variables cualitativas, el coeficiente se reduce al coeficiente de coincidencias (4.7).

Con esta idea, podemos construir fácilmente otros coeficientes. Será adecuado combinar coeficientes que independientemente posean la propiedad euclídea si queremos que el coeficiente resultante de la suma también la posea. Respecto a los coeficientes para variables cuantitativas, nos va a interesar utilizar aquellos que dividen cada comparación por un factor de normalización antes de sumar. Respecto a los coeficientes de similaridad para binarias y cualitativas, nos va a convenir los de rango [0,1], pues si permitimos valores negativos en el rango, las disimilaridades que obtendremos serán mayores a 1, por lo que adicionalmente debemos escalarlas antes de sumar. Nos referimos a la tabal 6 de Gower y Legendre (1986) para mayor detalle.

4.2. Regresión basada en distancias

La representación de un conjunto finito Ω de objetos, individuos o estímulos, constituye una de las más interesantes aplicaciones de la estadística basada en la topología asociada a una distancia. Las representaciones más usuales de un conjunto finito de elementos son: representación euclídea, representación ultramétrica (en forma de dendograma), representación cuadripolar (en forma de árbol aditivo), y representación de robinson (en forma de árbol piramidal). Nosotros, en el trabajo nos centramos en las posibilidades de la representación euclídea, que son clásicas y numerosas en análisis multivariante. Especialmente vamos a ver su aplicación a la predicción.

4.2.1. Escalado multidimensional métrico

4.2.1.1. Configuración euclídea

La demostración del siguiente teorema de caracterización [Schoenberg (1935)] puede encontrarse en Cuadras (1996), Mardia, Kent y Bibby (1979) y Seber (1984):

TEOREMA 5.1 *Sea* $\Delta = \left(\delta_{ij}\right)$ *una matriz* $n \times n$ *de disimilaridades sobre un conjunto finito* $\Omega = \{1,2,\ldots,n\}$. *Consideremos la matriz* $\mathbf{G} = \mathbf{H}\left(-\frac{1}{2}\Delta^{(2)}\right)\mathbf{H}$, *donde* $\mathbf{H} = \mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T$ *es la matriz centradora de datos, con* $\mathbf{1}_n$ *representando el vector* $n \times 1$ *cuyos elemenmtos son todos iguales a 1, y el superíndice* $^{(2)}$ *simbolizando el producto de Hadamard elemento a elemento,* $\Delta^{(2)} = \Delta \circ \Delta$.

Entonces, Δ *es euclídea si, y sólo si* $\mathbf{G}$ *es semidefinida positiva.*

En caso afirmativo, Ω *puede ser representado por* $\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathbb{R}^r$ *(escritos como vectores fila), siendo r=rango*$(\mathbf{G}) \leq n-1$, *de modo que*

$$\delta_{ij}^2 = \left(\mathbf{x}_i - \mathbf{x}_j\right)\left(\mathbf{x}_i - \mathbf{x}_j\right)^T, \ \forall i,j \in \Omega. \tag{4.9}$$

Si **G** es semidefinida positiva y de rango *r*, entonces existe al menos una matriz **X** de dimensión $n \times r$ tal que

$$\mathbf{G} = \mathbf{X}\mathbf{X}^T. \tag{4.10}$$

Las filas $\mathbf{x}_1,\ldots,\mathbf{x}_n$ de tal matriz satisfacen (4.9). Estas filas y, por extensión, la matriz **X** reciben el nombre de *configuración euclídea* de Δ.

La definición de **G**, aplicando doble centrado a $\left(-\frac{1}{2}\Delta^{(2)}\right)$ asegura que las configuraciones de (4.10) son centradas, es decir, $\sum \mathbf{x}_i = 0$ (en notación matricial $\mathbf{1}^T\mathbf{X} = \mathbf{0}$).

Cuando **X** proviene de la descomposición espectral de $\mathbf{G} = \mathbf{U}\Lambda^2\mathbf{U}^T$, se obtiene la solución clásica del *Escalado multidimensional métrico* (*EMM*) planteada por Rao (1964) y Gower (1966):

$$\mathbf{H}\left(-\frac{1}{2}\boldsymbol{\Delta}^{(2)}\right)\mathbf{H} = \mathbf{U}\boldsymbol{\Lambda}^2\mathbf{U}^T, \tag{4.11}$$

de donde

$$\mathbf{X} = \mathbf{U}\boldsymbol{\Lambda}. \tag{4.12}$$

Las n filas de $\mathbf{X}$ son las *coordenadas principales* de los elementos de Ω respecto a la distancia δ y las r columnas $\mathbf{X}_1,\ldots,\mathbf{X}_r$ de $\mathbf{X}$, llamadas *ejes principales* de la representación, pueden ser interpretadas como componentes principales.

La igualdad

$$\mathbf{X}^T\mathbf{X} = \boldsymbol{\Lambda}^2 \tag{4.13}$$

significa que los ejes principales son incorrelacionados y con varianzas dadas por los elementos de la diagonal principal de $\boldsymbol{\Lambda}^2$, es decir, los valores propios $\lambda_1^2,\ldots,\lambda_r^2$ de $\mathbf{G}$.

La representación además posee la siguiente propiedad óptima: Si $\mathbf{X}_{(k)}$, para $k \le r$, contiene las k primeras columnas de $\mathbf{X}$ (ordenadas de acuerdo con el decrecimiento de los valores propios de $\mathbf{G}$), y $\boldsymbol{\Delta}(k) = \left(d(k)_{ij}\right)$ es la matriz de distancias euclídeas entre las filas de $\mathbf{X}_{(k)}$, entonces

$$\sum_{i,j=1}^{n} d(k)_{ij}^2 = 2n\left(\lambda_1^2 + \cdots + \lambda_k^2\right)$$

es máxima en dimensión $k \le r$.

Del teorema 5.1 obtenemos un criterio para estudiar la euclídianidad de una matriz de distancias $\boldsymbol{\Delta}$, consistente en la comprobación de si la correspondiente matriz $\mathbf{G}$ es semidefinida positiva. Para resultados posteriores cabe notar que la relación matricial entre la matriz de productos escalares, $\mathbf{G}$, y la matriz de distancias al cuadrado, $\boldsymbol{\Delta}^{(2)}$, es:

$$\boldsymbol{\Delta}^{(2)} = \mathbf{g}^T\mathbf{1}_n^T + \mathbf{1}_n\mathbf{g} - 2\mathbf{G} \tag{4.14}$$

donde $\mathbf{g}$ es el vector fila que contiene los elementos de la diagonal principal de $\mathbf{G}$.

Si consideramos la distancia (4.6), y **S** es una matriz de similaridades semidefinida positiva, entonces es inmediato comprobar que la matriz Δ resultante es euclídea: descomponemos $\mathbf{S}=\mathbf{X}\mathbf{X}^T$ y obtenemos $s_{ij}=\mathbf{x}_i\mathbf{x}_j^T$, por lo que $\delta_{ij}^2=s_{ii}+s_{jj}-2s_{ij}=\left(\mathbf{x}_i-\mathbf{x}_j\right)\left(\mathbf{x}_i-\mathbf{x}_j\right)^T$.

Escalado multidimensional no métrico

Supongamos que la matriz de distancias Δ no es euclídea, para solventarlo, deberíamos aproximar δ a una distancia *d*, a fin de que (Ω,δ) admita una representación euclídea aproximada. Para ello hacemos uso de alguna transformación $d=\varphi(\delta)$, donde φ es una función monótona no decreciente, a fin de que se conserve la preordenación de las distancias originales, es decir, $\delta_{ij}\le\delta_{i'j'}\Leftrightarrow d_{ij}\le d_{i'j'}$., y por lo tanto individuos próximos (o lejanos) según δ también estarán próximos (o lejanos) según *d*. En general, la función φ es no lineal. Dos transformaciones algebraicas sencillas especialmente interesantes son:

- transformación aditiva:

$$d_{ij}=\begin{cases}0 & i=j\\ \delta_{ij}+c & i\neq j\end{cases} \tag{4.15}$$

- transformación q-aditiva:

$$d_{ij}^2=\begin{cases}0 & i=j\\ \delta_{ij}^2+c & i\neq j\end{cases} \tag{4.16}$$

El siguente teorema nos proporciona una herramienta en la elección de *c* en cada caso. La demostración se encuentra en las referencias incluidas:

TEOREMA 5.2 *Sea* $\Delta=\left(\delta_{ij}\right)$ *una matriz de distancias no euclídeas, por lo que* ***G*** *tiene valores propios positivos y negativos:* $\lambda_1>\cdots>\lambda_k>0>\lambda_1'>\cdots>\lambda_{k'}'$, *con* $k+k'=n-1$. *Entonces se verifica:*

- *La transformación q-aditiva* (4.16) *con* $c\ge-2\lambda_k'$, *convierte* Δ *en* ***D*** *euclídea* [Lingoes (1971); Mardia (1978)].
- *La transformación aditiva* (4.15) *con* $c\ge\lambda$, *donde* λ *es el mayor valor propio de la matriz no simétrica*

$$\begin{pmatrix} \mathbf{0} & \mathbf{2G} \\ -\mathbf{I} & -\mathbf{4G}_r \end{pmatrix}$$

siendo **G** *la matriz asociada a* Δ *y* $\mathbf{G}_r$ *la matriz asociada a* $\Delta_r = \left(\sqrt{d_{ij}}\right)$, *convierte* Δ *en* ***D*** *euclídea* [Cailliez (1983)].

La mejor transformación q-aditiva es aquella que distorsiona lo menos posible la distancia original. De acuerdo con este criterio, el mejor valor para la constante es $c = -2\lambda_k'$. Las transformaciones aditiva y no lineales son más complicadas. Usualmente los algoritmos de escalado multidimensional utilizan transformaciones no lineales siguiendo criterios de minimización de alguna medida de discrepancia entre la distancia original y la transformada. Por ejemplo, el método de Kruskall consiste en:

1. Fijar una dimensión euclídea p.
2. Transformar la distancia δ_{ij} en la "disparidad" $\hat{\delta}_{ij} = \varphi\left(\delta_{ij}\right)$, donde φ es una función monótona creciente. Las disparidades conservan la preordenación de las distancias.
3. Ajustar una distancia euclídea d_{ij} a las disparidades $\hat{\delta}_{ij}$ de manera que minimice $\sum_{i<j}\left(d_{ij} - \hat{\delta}_{ij}\right)^2$.
4. Asociar a las distancias d_{ij} una configuración euclídeana p-dimensional, y representar a los n objetos a partir de las coordenadas de la configuración.

Para saber si la representación es buena, se calcula la cantidad $S = \sqrt{\frac{\sum_{i<j}\left(d_{ij} - \hat{\delta}_{ij}\right)^2}{\sum_{i<j} d_{ij}^2}}$ denominada "stress", que verifica $0 \le S \le 1$, aunque se suele expresar en tanto por ciento. La representación se considera buena si S no supera el 5%. También es conveniente obtener el diagrama de Shepard, que consiste en representar los $n(n-1)/2$ puntos $\left(\delta_{ij}, d_{ij}\right)$. Si los puntos dibujan una curva creciente, la representación es buena, porque eso quiere decir que conserva bien la preordenación.

4.2.1.2. Configuración para un nuevo punto

Dada la representación del EM r-dimensional $\mathbf{X}$, supongamos que conocemos las distancias

$$\delta_i = \delta(\omega_i, \omega), \quad 1 \le i \le n, \tag{4.17}$$

de un nuevo objeto ω a cada objeto de Ω, y sea $\mathbf{d} = \left(\delta_1^2, \ldots \delta_n^2\right)$ el vector fila de las distancias al cuadrado de (4.17). En principio podemos construir una matriz $(n+1)\times(n+1)$ de distancias al cuadrado

$$\begin{pmatrix} \boldsymbol{\Delta}^{(2)} & \mathbf{d}^T \\ \mathbf{d} & \mathbf{0} \end{pmatrix},$$

y repetir todo el proceso para obtener una nueva solución de EMM. Sin embargo, es posible actualizar la configuración actual añadiendo una fila $\hat{\mathbf{x}}_{n+1}$ a $\mathbf{X}$, formando una configuración euclídea de dimensión $(n+1)\times r$. Gower (1968) demuestra que

$$\hat{\mathbf{x}}_{n+1} = \frac{1}{2}(\mathbf{g} - \mathbf{d})\mathbf{X}\boldsymbol{\Lambda}^{-2}, \tag{4.18}$$

proporciona una representación de ω como punto de $\mathbf{x} \in \mathbb{R}^r$.

Siguiendo el mismo razonamiento, se llega a la fórmula de interpolación

$$\hat{\mathbf{x}}_{n+1} = \frac{1}{2}(\mathbf{g} - \mathbf{d})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}, \tag{4.19}$$

válida para cualquier configuración euclídea centrada.

La fórmula de *añadir un punto* (4.18) ha sido utilizada por Gower (1988) en biplots no lineales, por Gower (1992) en biplots generalizados, por Krzanowski (1994) en análisis canónico generalizado, y por Cuadras y Arenas (1990) y Cuadras, Arenas y Fortiana (1996) en regresión basada en distancias tal y como mostraremos en la siguiente sección del trabajo.

Si disponemos de t nuevos puntos y $\Delta_2^{(2)}$ es la matriz $t \times n$ que contiene las distancias al cuadrado entre esos puntos y los del conjunto original, entonces la matriz $\hat{\mathbf{X}}_2$ de nuevas coordenadas vendrá dada por

$$\hat{\mathbf{X}}_2 = \frac{1}{2}\left(\mathbf{1}_t \mathbf{g} - \Delta_2^{(2)}\right)\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}. \tag{4.20}$$

NOTA: Utilizando (4.14), se comprueba fácilmente que si entramos $\Delta_2^{(2)} = \Delta^{(2)}$ en (4.20), cubrimos la solución $\hat{\mathbf{X}}_2 = \mathbf{X}$.

4.2.2. Predicción basada en distancias

Sea $\mathbf{Y}$ una variable dependiente de un conjunto Ξ de variables, posiblemente de tipo mixto. Supongamos que la observación de Ξ sobre un conjunto Ω de n individuos permite obtener una matriz de datos, $\mathbf{F}$, a partir de la cual construimos una matriz $n \times n$ de distancias, Δ. El esquema de la predicción basada en distancias es como sigue:

$$\left.\begin{array}{c} \Omega \overset{\Xi}{\rightarrow} \Delta \rightarrow \mathbf{X} \\ \Omega \overset{\mathbf{Y}}{\rightarrow} \mathbf{y} \\ \{n+1\} \overset{\Xi}{\rightarrow} \xi_{n+1} \end{array}\right\} \mapsto y_{n+1} = f\left(\mathbf{X}, \mathbf{y}, \xi_{n+1}\right)$$

es decir, la predicción y_{n+1} de $\mathbf{Y}$ para un nuevo individuo $\{n+1\}$ es función de la matriz de configuración euclídea $\mathbf{X}$ obtenida a partir de Δ, del vector $\mathbf{y}$ de observaciones de $\mathbf{Y}$ sobre Ω, y de las observaciones ξ_{n+1} de Ξ sobre $\{n+1\}$.

Se han estudiado tres tipos de problemas [Cuadras y Fortiana (1993a)]:

1. Predecir una variable cuantitativa **Y** como una función de regresión de un conjunto de variables Ξ de tipo mixto.
2. Predecir **Y** cuando la relación con Ξ no es lineal.
3. Predecir **Y**, discreta con *g* estados, como un problema de clasificación, siendo Ξ un conjunto mixto de variables.

Los puntos 1 y 2 se abordan haciendo uso de un modelo de regresión que pasaremos a describir en el siguiente apartado y el punto 3 mediante análisis discriminante [Oliva (1995)] como sigue:

Si **Y** tiene *g* estados que corresponden a las poblaciones $\pi_1,\ldots,\pi_g$, y se dispone de una determinada muestra global $\Omega = \Omega_1 \cup \Omega_2 \cup \cdots \cup \Omega_g$ de tamaño *n*, donde cada Ω_k es un conjunto de n_k individuos de π_k, predecir **Y** para un individuo $\{n+1\}$ equivale a clasificarlo en una de las *g* subpoblaciones. Cuadras (1989b) estudia una regla de clasificación que parte de las *g* funciones discriminantes

$$f_k\left(\{n+1\}\right) = \frac{1}{n_k}\sum_{i=1}^{n_k}\delta_i^2(k) - \frac{1}{n_k^2}\sum_{i=1}^{n_k}\sum_{j=i}^{n_k}\delta_{ij}^2(k) \tag{4.21}$$

donde $\Delta(k) = \left(\delta_{ij}(k)\right)$ es la matriz de distancias de Ω_k, y $\delta_i(k)$ las distancias de $\{n+1\}$ a los individuos de esta submuestra. La regla de clasificación es:

$$\text{[BD] Asignar } \{n+1\} \text{ a } \pi_i \text{ si } f_i\left(\{n+1\}\right) = \min\left\{f_1\left(\{n+1\}\right),\ldots,f_g\left(\{n+1\}\right)\right\}. \tag{4.22}$$

Este método de discriminación goza de buenas propiedades:

- Coincide con el discriminador lineal clásico cuando δ_{ij} es la distancia de Mahalanobis.
- La estimación de la probabilidad de clasificación errónea es fácilmente calculable.
- En caso de conocerse las probabilidades de asignación *a priori*, éstas se pueden incorporar al modelo.
- Puede ser aplicado correctamente a discriminación con variables mixtas.

Numerosos ejemplos de aplicación de [BD], con datos reales y simulados, han sido estudiados por Cuadras (1992).

4.2.2.1. Formulación de la regresión basada en distancias

Pasamos a la descripción de la *Regresión Basada en Distancias* (*RBD*) [Cuadras (1989b); Cuadras y Arenas (1990); Cuadras, Arenas y Fortiana (1996)]. Sea Ω un conjunto de n individuos para los que: sea $\mathbf{Y}$ un vector de dimensión n conteniendo la variable respuesta continua observada en los datos y sea $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_P$ un conjunto de variables explicativas de tipo mixto. Escogemos, en el espacio predictor, una métrica δ de entre las que satisfacen la propiedad euclídea. Calculamos $\boldsymbol{\Delta} = \left(\delta(i,j)\right)$ la matriz $n \times n$ de distancias entre individuos, y a partir de ella la matriz $\mathbf{X}$ de configuración euclídea, a partir de la descomposición (4.10). En este caso, alguna $\mathbf{X}$ tal que $\mathbf{G} = \mathbf{X}\mathbf{X}^T$ sea una configuración euclídea.

Así, dada $\mathbf{X}$ tal que $\mathbf{1}_n^T \mathbf{X} = \mathbf{0}$, $\mathbf{X}\mathbf{X}^T = \mathbf{G}$ y rango($\mathbf{X}$) = r, realizamos la regresión del vector $\mathbf{Y}$ sobre el espacio de las columnas de $\mathbf{X}$, que jugarán el papel de predictores en la regresión:

$$\mathbf{Y} = \boldsymbol{\beta}_0 + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{4.23}$$

$$E[\boldsymbol{\varepsilon}] = 0 \quad E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2 \mathbf{I}_n . \tag{4.24}$$

La estimación de $\boldsymbol{\beta}$ tal que $\left\|\mathbf{Y} - \boldsymbol{\beta}\mathbf{X}\right\|^2 = \min$, como es sabido, viene dada por

$$\hat{\boldsymbol{\beta}}_0 = \mathbf{1}_n \bar{y} \qquad \hat{\boldsymbol{\beta}} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{Y} . \tag{4.25}$$

Sustituyendo,

$$\hat{\mathbf{Y}} = \hat{\boldsymbol{\beta}}_0 + \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{1}_n \bar{y} + \mathbf{P}\mathbf{Y} . \tag{4.26}$$

donde $\bar{y} = \frac{\mathbf{1}_n^T \mathbf{Y}}{n}$ y $\mathbf{P} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ es el proyector ortogonal en el espacio de las columnas de **X**. El proyector puede ser expresado en términos de distancias como sigue: Sea $\mathbf{G} = \mathbf{H}\left(-\frac{1}{2}\boldsymbol{\Delta}^{(2)}\right)\mathbf{H}$, entonces $\mathbf{P} = \mathbf{G}^+\mathbf{G} = \mathbf{G}\mathbf{G}^+$ donde $\mathbf{G}^+$ es la g-inversa de Moore-Penrose. También se cumple que $\mathbf{G}^+ = \mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-2}\mathbf{X}^T$. Ésta es una definición consistente cuando $\boldsymbol{\Delta}$ es una matriz de distancias euclídea, que en el contexto del EMM, se reduce a que **G** sea semidefinida positiva.

4.2.2.2. Predicción para un nuevo individuo

La predicción para un nuevo objeto es como sigue: dado **d**, el vector fila con las distancias al cuadrado en el espacio predictor entre el nuevo objeto $\{n+1\}$ y los otros *n*, (4.17) al cuadrado, y **g** el vector fila $1 \times n$ con la diagonal de **G**, podemos calcular $\hat{\mathbf{x}}_{n+1}$ con la fórmula de interpolación de Gower (4.19) y, a partir de ella la predicción:

$$\hat{\mathbf{y}}_{n+1} = \hat{\beta}_0 + \hat{\mathbf{x}}_{n+1}\hat{\boldsymbol{\beta}}. \tag{4.27}$$

Sustituyendo (4.19) y (4.25) en (4.27),

$$\hat{y}_{n+1} = \bar{y} + \frac{1}{2}(\mathbf{g}-\mathbf{d})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{Y} = \bar{y} + \frac{1}{2}(\mathbf{g}-\mathbf{d})\mathbf{X}\left(\mathbf{X}^T\mathbf{X}\right)^{-2}\mathbf{X}^T\mathbf{Y},$$

i.e.,

$$\hat{y}_{n+1} = \bar{y} + \frac{1}{2}(\mathbf{g}-\mathbf{d})\mathbf{G}^+\mathbf{Y}. \tag{4.28}$$

Observamos que la fórmula (4.28) no depende de la configuración euclídea utilizada, tan sólo de la matriz de distancias.

Si disponemos en general de *t* puntos nuevos y $\boldsymbol{\Delta}_2^{(2)}$ es la matriz $t \times n$ de distancias al cuadrado entre los nuevos puntos y los originales, la estimación basada en distancias, $\hat{\mathbf{Y}}_2$, viene dada por

$$\hat{\mathbf{Y}}_2 = \mathbf{1}_t \bar{y} + \frac{1}{2}\left(\mathbf{1}_t \mathbf{g} - \Delta_2^{(2)}\right)\mathbf{G}^+\mathbf{Y}. \tag{4.29}$$

NOTA: Utilizando (4.14), se prueba fácilmente que si entramos $\Delta_2^{(2)} = \Delta^{(2)}$ en (4.29), cubrimos $\hat{\mathbf{Y}}_2 = \hat{\mathbf{Y}}$.

Observamos que el único problema numérico en la RBD es el cálculo de $\mathbf{G}^+$. Para ello se requiere una diagonalización de tamaño rango(**G**), claramente más rápida que la diagonalización (o descomposición singular) de tamaño *n* que en principio, para el cálculo de las coordenadas (4.12) parecía necesaria. Utilizaremos un *algoritmo de Cholesky* modificado para matrices semidefinidas positivas [Cheng y Higham (1998)]. Sea **G** una matriz simétrica $n \times n$ semidefinida positiva de rango *r*. Entonces, existe una matriz de permutaciones $\mathbf{\Pi}$ tal que $\mathbf{\Pi}^T\mathbf{G}\mathbf{\Pi}$ tiene una única descomposición de Cholesky, la cual es de la forma: $\mathbf{\Pi}^T\mathbf{G}\mathbf{\Pi} = \mathbf{T}\mathbf{T}^T$, $\mathbf{T} = \begin{pmatrix} \mathbf{T}_{11} & 0 \\ \mathbf{T}_{21} & 0 \end{pmatrix}$ donde $\mathbf{T}_{11}$ es una matriz $r \times r$ triangular inferior con elementos diagonales positivos. Entonces, dado el producto $\mathbf{X} = \mathbf{\Pi}\mathbf{T}$, $n \times p$, con $p < r$,

$$\mathbf{G}^+ = \mathbf{\Pi}\mathbf{T}\left(\mathbf{T}^T\mathbf{T}\right)^{-2}\mathbf{T}^T\mathbf{\Pi}^T. \tag{4.30}$$

4.2.3. Casos particulares

Nos referimos a Cuadras y Arenas (1990), Cuadras, Arenas y Fortiana (1996), donde se demuestra que:

Caso 1: Retomando el punto 2 del punto 4.2.2., supongamos que $\mathbf{Y} = f\left(\Xi_1, \Xi_2, \ldots, \Xi_P\right) + \boldsymbol{\varepsilon}$, es decir, **Y** es una función no lineal del conjunto $\Xi = \left(\Xi_1, \Xi_2, \ldots, \Xi_P\right)$ de *P* variables que suponemos contínuas. Sean $\left(\xi_{i1}, \ldots, \xi_{iP}\right)$ y $\left(\xi_{j1}, \ldots, \xi_{jP}\right)$ observaciones sobre un par (*i*,*j*) de elementos de Ω. Adoptando la distancia δ_{ij} valor absoluto:

$$\delta_{ij} = \sqrt{\sum_{h=1}^{P} \left| \xi_{ih} - \xi_{jh} \right|},$$

y aplicando el modelo (4.23), se consigue una buena predicción de **Y** sin necesidad de conocer *f*. Una justificación de esta propiedad del modelo ha sido encontrada por Cuadras y Fortiana (1993b) en términos de polinomios de Tchebychev.

Caso 2: Si las variables explicativas son continuas y la función de distancias utilizada es (4.2), la predicción obtenida mediante RBD y regresión múltiple clásica coinciden.

Caso 3: Si las variables explicativas son categóricas y el coeficiente de similaridad utilizado es el de coincidencias (4.7), la predicción obtenida mediante RBD y regresión múltiple clásica utilizando tantas variables binarias como clases, coinciden.

El caso 2, que nos dice que cuando la distancia empleada es ℓ^2 y los predictores son cuantitativos surge como caso particular el modelo de regresión lineal por mínimos cuadrados ordinarios, nos permite considerar a la RBD como una extensión del modelo clásico de regresión en el ámbito de las distancias.

4.2.4. Términos de interacción en regresión basada en distancias

Una matriz de distancias entre individuos Δ, se calcula a partir de ciertas variables, $\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_P$. Supongamos que nos interesa incluir la cuantificación de las interacciones de esas variables en la distancia. Tenemos dos posibilidades:

- La clásica de incluir una nueva variable $\mathbf{F}_{ij}$ para cada interacción *ij* deseada:
  - Si se trata de dos variables cuantitativas, incluiríamos el producto $\mathbf{F}_{ij} = \mathbf{F}_i \circ \mathbf{F}_j$, al igual que para el MLG.

- Si se trata de dos variables cualitativas podríamos crear una nueva variable con el código cruzado de ambas variables, a la que nomenclamos $\mathbf{F}_{ij} = \mathbf{F}_i * \mathbf{F}_j$, y tratarla como una nueva variable categórica, posibilidad que no ofrece el MLG.
- Y si se trata de la interacción entre una cuantitativa y una cualitativa, podemos pasar por el proceso de creación de variables binarias para las diferentes categorías de la variable cualitativa, al igual que para el MLG: supongamos $\mathbf{F}_i$ cuantitativo y $\mathbf{F}_j$ cualitativo con K categorías, para tener en cuenta la interacción incluiríamos K nuevas variables cuantitativas $\mathbf{F}_{ij_k} = \mathbf{F}_i \circ \mathbf{F}_{j_k}$ para $k = 1,\ldots,K$, siendo $\mathbf{F}_{j_k}$ las binarias asociadas a cada clase.

➢ Otro procedimiento es el siguiente [Fortiana y Esteve (1999a,b); Esteve (2003)]: calculamos para cada variable la matriz de distancias asociada, que denotaremos por $\mathbf{\Delta}_i$ para $i = 1,\ldots,P$, y a su vez los correspondientes productos escalares $\mathbf{G}_i = -\frac{1}{2}\mathbf{H}\mathbf{\Delta}_i^{(2)}\mathbf{H}$, entonces la interacción la incluimos en la matriz total de productos escalares

$$\mathbf{G}_T = \mathbf{G}_1 + \mathbf{G}_2 + \cdots + \mathbf{G}_P. \tag{4.31}$$

del siguiente modo: supongamos que deseamos incluir la interacción *ij*, a la que simbolizamos por $\mathbf{G}_{ij}$, ahora la nueva matriz total de productos escalares se calcula como

$$\mathbf{G}_T = \mathbf{G}_1 + \mathbf{G}_2 + \cdots + \mathbf{G}_P + \mathbf{G}_{ij} \tag{4.32}$$

siendo

$$\mathbf{G}_{ij} = \mathbf{G}_i \circ \mathbf{G}_j. \tag{4.33}$$

En el caso extremo de querer incluir todas las interacciones tendríamos: $\mathbf{G}_T = \mathbf{G} + \mathbf{G} \circ \mathbf{G}$, siendo $\mathbf{G} = \mathbf{G}_1 + \mathbf{G}_2 + \cdots + \mathbf{G}_P$.

El producto de Hadamard aquí empleado es un caso particular de interacción polinómica. El caso general descrito en Esteve (2003) responde a la siguiente formulación:

$$\mathbf{G}_{ij} = \mathbf{G}_i^{1/2}\mathbf{K}\mathbf{G}_j^{1/2} + \mathbf{G}_j^{1/2}\mathbf{K}\mathbf{G}_i^{1/2} \tag{4.34}$$

donde **K** es una matriz de parámetros de dimensión $n \times n$.

Cabe notar que si las matrices iniciales, $\boldsymbol{\Delta}_1,\ldots,\boldsymbol{\Delta}_p$, son euclídeas, las combinaciones (4.32) descritas para la matriz total de productos escalares, **G**, son semi-definidas positivas [Esteve (2003)].

4.3. Generalización al caso heteroscedástico de la regresión basada en distancias

El modelo de RBD se reduce al modelo lineal (4.23), con perturbaciones incorrelacionadas y varianza constante: $E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2 \mathbf{I}_n$. En este apartado vamos a generalizarlo al caso heteroscedástico, en el que asumiremos que la varianza de las perturbaciones, σ^2, no será constante a lo largo de las observaciones, aunque continuaremos asumiendo perturbaciones incorrelacionadas:

$$E[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2 \boldsymbol{\Omega} = \begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \sigma_n^2 \end{bmatrix}. \tag{4.35}$$

Razonamiento heurístico

Para la deducción de las fórmulas generales aplicaremos el siguiente razonamiento heurístico: Deduciremos las fórmulas en el caso particular en el que tenemos υ individuos que son repeticiones (o copias) de m individuos diferentes, con frecuencias absolutas $\upsilon_1,\ldots,\upsilon_m$, tales que $\upsilon = \sum_{i=1}^{m} \upsilon_i$. Porque en esta situación, los cálculo del modelo no ponderado en el que tenemos a los individuos repetidos, con matrices ampliadas de dimensión υ, deben coincidir con los cálculos del modelo ponderado en el que los individuos no están repetidos, con matrices reducidas (de dimensión m) con pesos $\mathbf{w} = (\upsilon_1/\upsilon,\ldots,\upsilon_m/\upsilon)^T$.

Pretendemos deducir mediante este razonamiento las fórmulas de

a) Configuración euclídea de $\boldsymbol{\Delta}$
b) Fórmula de interpolación de Gower
c) Predicción para un nuevo individuo en una regresión basada en distancias con predictores $\boldsymbol{\Delta}$

Consideremos un conjunto Ω ahora de m objetos y una matriz $m \times m$ de distancias euclídea $\boldsymbol{\Delta} = (\delta_{ij})$ sobre Ω, y un vector $\mathbf{w}$ de pesos de dimensión $m \times 1$, tal que $w_j > 0$ para $j = 1,\ldots,m$, y $\sum_{j=1}^{m} w_j = 1$.

Sea la matriz diagonal de pesos $\mathbf{D}_{\mathbf{w}} = diag(\mathbf{w})$, la matriz de w-centrado $\mathbf{K} = \mathbf{I}_m - \mathbf{1}_m \mathbf{w}^T$, y las matrices

$$\mathbf{A} = \left(-\frac{1}{2} \boldsymbol{\Delta}^{(2)} \right) \tag{4.36}$$

$$\mathbf{G} = \mathbf{K}\mathbf{A}\mathbf{K}^T. \tag{4.37}$$

4.3.1. Escalado multidimensional métrico ponderado

4.3.1.1. Configuración euclídea

Configuración euclídea

Partimos de una matriz de distancias euclídea $\boldsymbol{\Delta}$ de dimensión $m \times m$. Una *configuración euclídea* de $\boldsymbol{\Delta}$, es un conjunto de m vectores $\mathbf{x}_1,\ldots,\mathbf{x}_m$ de algún espacio euclídeo E tales que

$$\left\| \mathbf{x}_i - \mathbf{x}_j \right\| = \delta_{ij}. \tag{4.38}$$

La configuración euclídea propiamente no depende del vector $\mathbf{w}$ de pesos.

Centrado de la configuración euclídea

Si queremos centrar la configuración (poner el 0 como origen) de manera que el centroide sea el 0, entonces sí hemos de tener en cuenta los pesos, imponiendo $\mathbf{w}^T\mathbf{X} = 0$.

Configuraciones euclídeas centradas

Sea $\Omega = (\omega_1,\ldots\omega_m)$ el conjunto de m elementos, y sea $\boldsymbol{\Delta}$ la matriz $m \times m$ de distancias euclídea

definida en Ω. Dados m enteros positivos $\upsilon_1,\ldots,\upsilon_m$, consideramos el conjunto $\tilde{\Omega}$, que contiene $\upsilon = \sum_{i=1}^{m} \upsilon_i$ elementos, consistentes en υ_i copias de ω_i, $i = 1,\ldots,m$ y la correspondiente matriz de distancias $\tilde{\Delta}$ de dimensión $\upsilon \times \upsilon$.

La matriz $\mathbf{M}$ de dimensión $\upsilon \times m$ formada por υ_1 filas de $(1,0,\ldots,0)$, υ_2 filas dc $(0,1,\ldots,0)$, ... , υ_m filas de $(0,0,\ldots,1)$, relaciona las dos matrices de distancias:

$$\tilde{\boldsymbol{\Delta}} = \mathbf{M}\boldsymbol{\Delta}\mathbf{M}^T. \tag{4.39}$$

Sea,

$$\tilde{\mathbf{A}} = \left(-\frac{1}{2}\tilde{\boldsymbol{\Delta}}^{(2)}\right) \tag{4.40}$$

$$\tilde{\mathbf{G}} = \mathbf{H}\tilde{\mathbf{A}}\mathbf{H}^T. \tag{4.41}$$

Como se cumple que $\mathbf{HM} = \mathbf{MK}$ y $\tilde{\mathbf{A}} = \mathbf{MAM}^T$, siendo $\mathbf{A} = \left(-\frac{1}{2}\boldsymbol{\Delta}^{(2)}\right)$, resulta que

$$\tilde{\mathbf{G}} = \mathbf{MGM}^T.$$

Una configuración euclídea $\tilde{\mathbf{X}}$ centrada de $\tilde{\Delta}$ es *centrada* si $1_\upsilon^T \tilde{\mathbf{X}} = 0$. Toda configuración euclídea centrada de $\tilde{\Delta}$ cumple $\tilde{\mathbf{X}}\tilde{\mathbf{X}}^T = \tilde{\mathbf{G}}$.

Una configuración euclídea $\mathbf{X}$ de $(\boldsymbol{\Delta}, \mathbf{w})$ es *w-centrada* si $\mathbf{w}^T\mathbf{X} = 0$. Toda configuración euclídea *w*-centrada de $(\boldsymbol{\Delta}, \mathbf{w})$ satisface $\mathbf{XX}^T = \mathbf{G}$.

Si $\tilde{\mathbf{X}}$ es una configuración euclídea centrada de $\tilde{\Delta}$, entonces

$$\tilde{\mathbf{X}} = \mathbf{MX}, \tag{4.42}$$

donde $\mathbf{X}$ es una configuración *w*-centrada de $(\boldsymbol{\Delta}, \mathbf{w})$ y recíprocamente.

Supongamos $\mathbf{X}$ una configuración *w*-centrada, entonces tenemos que:

$$\mathbf{S} = \mathbf{X}^T \mathbf{D}_w \mathbf{X} = \mathbf{X}^T \left(\mathbf{D}_w - \mathbf{w}\mathbf{w}^T \right) \mathbf{X} . \tag{4.43}$$

Observación

$\mathbf{M}$ satisface las siguientes igualdades: $\mathbf{1}_\upsilon^T \mathbf{M} = \upsilon \, \mathbf{w}^T$, $\mathbf{M}^T \mathbf{M} = \upsilon \, \mathbf{D}_w$, $\mathbf{M}\mathbf{1}_m = \mathbf{1}_\upsilon$, $\mathbf{H}\mathbf{M} = \mathbf{M}\mathbf{K}$,

donde $\mathbf{H} = \mathbf{I}_\upsilon - \upsilon^{-1} \mathbf{1}_\upsilon \mathbf{1}_\upsilon^T$ es la matriz de centrado $\upsilon \times \upsilon$ y $\mathbf{K} = \mathbf{I}_m - \mathbf{1}_m \mathbf{w}^T$. También:

$\mathbf{K} = \mathbf{I}_m - \mathbf{1}_m \mathbf{w}^T = \mathbf{D}_w^{-1} \left(\mathbf{D}_w - \mathbf{w}\mathbf{w}^T \right) = \mathbf{D}_w^{-1} \left(\mathbf{I}_m - \mathbf{w}\mathbf{1}_m^T \right) \mathbf{D}_w = \mathbf{D}_w^{-1} \mathbf{K}^T \mathbf{D}_w$, $\mathbf{D}_w \mathbf{K} = \mathbf{K}^T \mathbf{D}_w$ y

$\mathbf{K}\mathbf{K} = \left(\mathbf{I}_m - \mathbf{1}_m \mathbf{w}^T \right) \left(\mathbf{I}_m - \mathbf{1}_m \mathbf{w}^T \right) = \mathbf{I}_m - \mathbf{1}_m \mathbf{w}^T - \mathbf{1}_m \mathbf{w}^T + \mathbf{1}_m \mathbf{w}^T = \mathbf{K}$.

4.3.1.2. Configuración para un nuevo punto

Supongamos que $\upsilon >$ número de columnas de $\tilde{\mathbf{X}}$ y que $m >$ número de columnas de $\tilde{\mathbf{X}}$. Entonces $\tilde{\mathbf{X}}^T \tilde{\mathbf{X}}$ es no singular. Para la deducción de la fórmula de interpolación de Gower aplicaremos el principio heurístico. Sea la fórmula de añadir un punto (4.19):

$$\hat{\mathbf{x}}_{\upsilon+1} = \frac{1}{2} \left(\tilde{\mathbf{g}} - \tilde{\mathbf{d}} \right) \tilde{\mathbf{X}} \left(\tilde{\mathbf{X}}^T \tilde{\mathbf{X}} \right)^{-1} \tag{4.44}$$

donde $\tilde{\mathbf{g}} = \mathbf{g}\mathbf{M}^T$ y $\tilde{\mathbf{d}} = \mathbf{d}\mathbf{M}^T$ (los elementos de $\mathbf{g}$ (de $\tilde{\mathbf{g}}$) son las longitudes al cuadrado de los vectores de la configuración euclídea).

Puesto que $\mathbf{M}^T \mathbf{M} = \upsilon \mathbf{D}_w$, sustituyendo obtenemos:

$$\begin{aligned} \hat{\mathbf{x}}_{\upsilon+1} &= \frac{1}{2} \left(\mathbf{g} - \mathbf{d} \right) \mathbf{M}^T \mathbf{M} \mathbf{X} \left(\mathbf{X}^T \mathbf{M}^T \mathbf{M} \mathbf{X} \right)^{-1} \\ &= \frac{1}{2} \left(\mathbf{g} - \mathbf{d} \right) \upsilon \mathbf{D}_w \mathbf{X} \left(\mathbf{X}^T \upsilon \mathbf{D}_w \mathbf{X} \right)^{-1} \\ &= \frac{1}{2} \left(\mathbf{g} - \mathbf{d} \right) \mathbf{D}_w \mathbf{X} \left(\mathbf{X}^T \mathbf{D}_w \mathbf{X} \right)^{-1} \end{aligned}$$

i.e.,

$$\hat{\mathbf{x}}_{v+1} = \frac{1}{2}(\mathbf{g}-\mathbf{d})\mathbf{D}_w\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-1}. \tag{4.45}$$

Para t nuevos puntos, siendo $\Delta_2^{(2)}$ la matriz $t \times m$ asociada con las distancias al cuadrado entre los nuevos puntos y los originales, la matriz $\hat{\mathbf{X}}_2$ de nuevas coordenadas, similar a (4.20) viene dada por

$$\hat{\mathbf{X}}_2 = \frac{1}{2}\left(\mathbf{1}_t\mathbf{g}-\Delta_2^{(2)}\right)\mathbf{D}_w\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-1}. \tag{4.46}$$

4.3.2. Regresión basada en distancias ponderada

Predicción para un nuevo individuo

Sea $\mathbf{Y}$ el vector de dimensión $m \times 1$ conteniendo la respuesta continua. De nuevo la matriz $\mathbf{M}$ relaciona los dos vectores de respuestas:

$$\tilde{\mathbf{Y}} = \mathbf{M}\mathbf{Y}. \tag{4.47}$$

Siguiendo el principio heurístico, vamos a deducir la fórmula de predicción de un nuevo individuo, (4.27), para la *Regresión Basada en Distancias Ponderada* (*RBDP*) de $\mathbf{Y}$ sobre $\boldsymbol{\Delta}$ con pesos $\mathbf{w}$:

$$\begin{aligned}
\hat{y}_{v+1} &= \overline{y} + \frac{1}{2}\left(\tilde{\mathbf{g}}-\tilde{\mathbf{d}}\right)\tilde{\mathbf{X}}\left(\tilde{\mathbf{X}}^T\tilde{\mathbf{X}}\right)^{-2}\tilde{\mathbf{X}}^T\tilde{\mathbf{Y}} \\
&= \overline{y} + \frac{1}{2}\left(\mathbf{g}-\mathbf{d}\right)\mathbf{M}^T\mathbf{M}\mathbf{X}\left(\mathbf{X}^T\mathbf{M}^T\mathbf{M}\mathbf{X}\right)^{-2}\mathbf{X}^T\mathbf{M}^T\mathbf{M}\mathbf{Y} \\
&= \overline{y} + \frac{1}{2}\left(\mathbf{g}-\mathbf{d}\right)\mathbf{D}_w\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-2}\mathbf{X}^T\mathbf{D}_w\mathbf{Y}
\end{aligned}$$

i.e.,

$$\hat{y}_{v+1} = \overline{y} + \frac{1}{2}\left(\mathbf{g}-\mathbf{d}\right)\mathbf{D}_w\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-2}\mathbf{X}^T\mathbf{D}_w\mathbf{Y}. \tag{4.48}$$

Para t nuevos puntos, siendo $\Delta_2^{(2)}$ la matriz $t \times m$ asociada con las distancias al cuadrado entre los nuevos puntos y los originales, la matriz de estimaciones $\hat{\mathbf{Y}}_2$, similar a (4.29) viene dada por

$$\hat{\mathbf{Y}}_2 = \mathbf{1}_t \bar{y} + \frac{1}{2}\left(\mathbf{1}_t \mathbf{g} - \boldsymbol{\Delta}_2^{(2)}\right)\mathbf{D}_w \mathbf{X}\left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right)^{-2} \mathbf{X}^T \mathbf{D}_w \mathbf{Y}\,. \tag{4.49}$$

Rescribimos (4.49) del siguiente modo:

$$\hat{\mathbf{Y}}_2 = \mathbf{1}_t \bar{y} + \frac{1}{2}\left(\mathbf{1}_t \mathbf{g} - \boldsymbol{\Delta}_2^{(2)}\right)\mathbf{D}_w^{1/2} \mathbf{F}^{+} \mathbf{D}_w^{1/2} \mathbf{Y} \tag{4.50}$$

donde

$$\mathbf{F}^{+} = \mathbf{D}_w^{1/2} \mathbf{X}\left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right)^{-2} \mathbf{X}^T \mathbf{D}_w^{1/2}\,. \tag{4.51}$$

LEMA

La pseudoinversa de Moore-Penrose de $\mathbf{F} = \mathbf{D}_w^{1/2} \mathbf{G} \mathbf{D}_w^{1/2}$ *es* $\mathbf{F}^{+} = \mathbf{D}_w^{1/2} \mathbf{X}\left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right)^{-2} \mathbf{X}^T \mathbf{D}_w^{1/2}$.

Demostración:

1) $\mathbf{F}\mathbf{F}^{+}$ es simétrica:

$$\begin{aligned}
\mathbf{F}\mathbf{F}^{+} &= \mathbf{D}_w^{1/2} \mathbf{G} \mathbf{D}_w^{1/2} \mathbf{D}_w^{1/2} \mathbf{X}\left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right)^{-2} \mathbf{X}^T \mathbf{D}_w^{1/2} \\
&= \mathbf{D}_w^{1/2} \mathbf{X}\left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right)\left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right)^{-2} \mathbf{X}^T \mathbf{D}_w^{1/2} \\
&= \mathbf{D}_w^{1/2} \mathbf{X}\left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{D}_w^{1/2}
\end{aligned}$$

que, evidentemente es simétrica.

2) $\mathbf{F}^{+}\mathbf{F}$ es simétrica:

$$\begin{aligned}
\mathbf{F}^{+}\mathbf{F} &= \mathbf{D}_w^{1/2} \mathbf{X}\left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right)^{-2} \mathbf{X}^T \mathbf{D}_w^{1/2} \mathbf{D}_w^{1/2} \mathbf{G} \mathbf{D}_w^{1/2} \\
&= \mathbf{D}_w^{1/2} \mathbf{X}\left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right)^{-2} \left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right) \mathbf{X}^T \mathbf{D}_w^{1/2} \\
&= \mathbf{D}_w^{1/2} \mathbf{X}\left(\mathbf{X}^T \mathbf{D}_w \mathbf{X}\right)^{-1} \mathbf{X}^T \mathbf{D}_w^{1/2}
\end{aligned}$$

que, evidentemente es simétrica.

3) $\mathbf{F}\mathbf{F}^{+}\mathbf{F} = \mathbf{F}$

$$\begin{aligned}\mathbf{F}\mathbf{F}^{+}\mathbf{F} &= \mathbf{D}_w^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}_w^{1/2}\mathbf{D}_w^{1/2}\mathbf{G}\mathbf{D}_w^{1/2} \\ &= \mathbf{D}_w^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)\mathbf{X}^T\mathbf{D}_w^{1/2} \\ &= \mathbf{D}_w^{1/2}\mathbf{X}\mathbf{X}^T\mathbf{D}_w^{1/2} = \mathbf{D}_w^{1/2}\mathbf{G}\mathbf{D}_w^{1/2} = \mathbf{F}\end{aligned}$$

4) $\mathbf{F}^{+}\mathbf{F}\mathbf{F}^{+} = \mathbf{F}^{+}$

$$\begin{aligned}\mathbf{F}^{+}\mathbf{F}\mathbf{F}^{+} &= \mathbf{D}_w^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-2}\mathbf{X}^T\mathbf{D}_w^{1/2}\mathbf{D}_w^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}_w^{1/2} \\ &= \mathbf{D}_w^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-2}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{D}_w^{1/2} \\ &= \mathbf{D}_w^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-2}\mathbf{X}^T\mathbf{D}_w^{1/2} = \mathbf{F}^{+}\end{aligned}$$

Para el cálculo numérico de $\mathbf{F}^{+} = \mathbf{D}_w^{1/2}\mathbf{X}\left(\mathbf{X}^T\mathbf{D}_w\mathbf{X}\right)^{-2}\mathbf{X}^T\mathbf{D}_w^{1/2}$ empleamos de nuevo la descomposición de Cholesky (4.30), tal y como queda reflejado en el anexo informático 4.2.

4.4. Selección de predictores

En Cuadras y Fortiana (1998) y Fortiana y Cuadras (1998) encontramos una generalización de la RBD a repuesta multivariante, donde la respuesta es introducida como una segunda matriz de distancias $\mathbf{\Delta}_\mathbf{Y}$. La RBD puede ser generalizada teniendo en cuenta que $\mathbf{\Delta}_\mathbf{Y}^{(2)} = \mathbf{1}_n\mathbf{g}_\mathbf{Y} + \mathbf{g}_\mathbf{Y}^T\mathbf{1}_n^T - 2\mathbf{G}_\mathbf{Y}$ donde $\mathbf{G}_\mathbf{Y} = \left(\mathbf{Y} - \mathbf{1}_n\overline{y}\right)\left(\mathbf{Y} - \mathbf{1}_n\overline{y}\right)^T$, se ajusta mediante $\hat{\mathbf{\Delta}}_\mathbf{Y}^{(2)} = \mathbf{1}_n\hat{\mathbf{g}}_\mathbf{Y} + \hat{\mathbf{g}}_\mathbf{Y}^T\mathbf{1}_n^T - 2\hat{\mathbf{G}}_\mathbf{Y}$, donde $\hat{\mathbf{G}}_\mathbf{Y} = \mathbf{P}_\mathbf{F}\mathbf{G}_\mathbf{Y}\mathbf{P}_\mathbf{F}$. En este trabajo tenemos en cuenta una respuesta cuantitativa $\mathbf{Y}$, y $\hat{\mathbf{G}}_\mathbf{Y} = \left(\hat{\mathbf{Y}} - \mathbf{1}_n\overline{y}\right)\left(\hat{\mathbf{Y}} - \mathbf{1}_n\overline{y}\right)^T$. Utilizamos por conveniencia alguna notación procedente de la formulación general, especialmente en lo que se refiere a las *variabilidades geométricas* de las dos matrices de distancias: $\mathbf{\Delta}_\mathbf{Y}^{(2)}$, $\mathbf{\Delta}_\mathbf{F}^{(2)}$ y sus productos escalares asociados $\mathbf{G}_\mathbf{Y}$ y $\mathbf{G}_\mathbf{F}$.

4.4.1. Medidas y tests estadísticos

La *variabilidad geométrica* de una matriz de distancias $\boldsymbol{\Delta}$ asociada con la matriz de productos escalares G se define como

$$V(\boldsymbol{\Delta}) = \frac{1}{2n^2}\mathbf{1}_n^T \boldsymbol{\Delta}^{(2)} \mathbf{1}_n = \frac{1}{n} traza(\mathbf{G}) . \tag{4.52}$$

Esta cantidad es la generalización natural, en los modelos basados en distancias, del concepto clásico de variabilidad total. En el caso de datos agregados su expresión es como sigue,

$$V\left(\boldsymbol{\Delta}(\upsilon)\right) = \frac{1}{2}\mathbf{w}^T \boldsymbol{\Delta}(\upsilon)^{(2)} \mathbf{w} . \tag{4.53}$$

Coeficiente de determinación

$$R^2_{\mathbf{Y},\mathbf{F}} = \frac{V(\hat{\boldsymbol{\Delta}}_{\mathbf{Y}})}{V(\boldsymbol{\Delta}_{\mathbf{Y}})} \tag{4.54}$$

Coeficiente de correlación parcial al cuadrado

$$r^2_{\mathbf{Y},\mathbf{F}_{k+1}|\mathbf{F}_1\mathbf{F}_2\cdots\mathbf{F}_k} = \frac{V^{k+1}(\hat{\boldsymbol{\Delta}}_{\mathbf{Y}}) - V^k(\hat{\boldsymbol{\Delta}}_{\mathbf{Y}})}{V(\boldsymbol{\Delta}_{\mathbf{Y}}) - V^k(\hat{\boldsymbol{\Delta}}_{\mathbf{Y}})} \tag{4.55}$$

donde

$V^k\left(\hat{\boldsymbol{\Delta}}_{\mathbf{Y}}\right)$ proviene de $\hat{\mathbf{Y}} = RBD\left(\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_k\right)$ y $V^{k+1}\left(\hat{\boldsymbol{\Delta}}_{\mathbf{Y}}\right)$ de $\hat{\mathbf{Y}} = RBD\left(\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_k, \mathbf{F}_{k+1}\right)$.

Test estadísticos

Supongamos que queremos comparar dos modelos,

$$\begin{aligned} M1: &\quad \hat{\mathbf{Y}} = RBD(\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_k) \\ M2: &\quad \hat{\mathbf{Y}} = RBD(\mathbf{F}_1, \mathbf{F}_2, \ldots, \mathbf{F}_k, \mathbf{F}_{k+1}) \end{aligned} \tag{4.56}$$

para comprobar si la variable añadida, $\mathbf{F}_{k+1}$, (fácilmente extensible al caso de un conjunto de predictores) es significativa en la explicación de la variabilidad de la respuesta, definimos el siguiente test estadístico, que en términos de variabilidades geométricas, es:

$$Q(\mathbf{F}_{k+1} \mid \mathbf{F}_1 \ldots \mathbf{F}_k) = \frac{V^{k+1}(\hat{\Delta}_{\mathbf{Y}}) - V^{k}(\hat{\Delta}_{\mathbf{Y}})}{V(\Delta_{\mathbf{Y}}) - V^{k+1}(\hat{\Delta}_{\mathbf{Y}})}. \tag{4.57}$$

Finalmente, cuando ya tenemos el modelo resultante con K variables seleccionadas, comprobamos si la variabilidad explicada por éstas sobre la no explicada es suficiente como para que la regresión tenga poder predictivo, a través del p-valor del estadístico:

$$Q = \frac{V^{K}(\hat{\Delta}_{\mathbf{Y}})}{V(\Delta_{\mathbf{Y}}) - V^{K}(\hat{\Delta}_{\mathbf{Y}})}. \tag{4.58}$$

Cabe notar, que los estadísticos sólo dependen de las interdistancias entre individuos.

4.4.2. Aspectos computacionales: estimación *bootstrap*

p-valores asociados a los estadísticos

Las distribuciones de los estadísticos (4.57) y (4.58) las estimaremos mediante simulación, haciendo uso de la metodología *bootstrap* [Diaconis y Efron (1983); Efron y Tibshirani (1993)]. Los modelos basados en distancias son especialmente adecuados para el empleo de *bootstrap* [Fortiana y Cuadras (1994)], pues el hecho que todas las interdistancias entre individuos de un remuestreo aparezcan ya en la matriz de distancias inicial, nos permite simular y vectorizar los remuestreos mediante matrices de multiplicidades, lo que es de gran economía computacional.

Un algoritmo para el cálculo de los *p*-valores

Partimos de la muestra original con n individuos con observación de la respuesta, $\mathbf{Y}$, y de una matriz de distancias euclídeas entre individuos, $\Delta_{\mathbf{F}}^{(2)}$, calculada en el espacio predictor a partir de $\mathbf{F}_1,\ldots,\mathbf{F}_P$. Un remuestreo vendrá determinado por una permutación con repetición del conjunto de individuos, y

será determinado por un vector de multiplicidades $\mathbf{N}=(N_1,N_2,\ldots,N_n)^T$, $N_j \geq 0$, $1 \leq j \leq n$ tal que $N_1+N_2+\cdots+N_n=n$. Generamos los vectores de multiplicidades siguiendo una distribución Multinomial de parámetros $\left[n;\left(\frac{1}{n},\frac{1}{n},\cdots,\frac{1}{n}\right)\right]$. Cada vector es el de las multiplicidades de n números aleatorios de distribución Uniforme $U[0,1]$ en n intervalos de igual amplitud entre 0 y 1. Este procedimiento lo repetimos B veces, para así obtener B vectores de multiplicidades: $M_{n\times B}=\left(N^1,N^2,\ldots,N^B\right)_{n\times B}$, lo que equivale a B muestras aleatorias de tamaño n a partir de la muestra original, para poder estimar la distribución de probabilidad, y en particular el deseado p-valor.

Todas las matrices de distancias requeridas, $\Delta_{\mathbf{F}}^{(2)}(b)$, $b=1,\ldots,B$, tendrán elementos de la matriz de distancias original, $\Delta_{\mathbf{F}}^{(2)}$. Así, sólo es necesario buscarlos entre los $\frac{n(n-1)}{2}$ elementos de la euclídea original. Por supuesto, todas las matrices construidas también serán euclídeas. Los elementos de una matriz específica $\Delta_{\mathbf{F}}^{(2)}(b)$ serán 0 para las posiciones correspondientes a individuos repetidos, y la correspondiente distancia para los no repetidos. Utilizamos la siguiente propiedad para la construcción de las B matrices requeridas:

Sea $\boldsymbol{\gamma}$ una matriz ($B\cdot n\times n$) que contiene, para cada uno de los B remuestreos, por ejemplo, para el b-ésimo, $\boldsymbol{\gamma}(b)$: N_1^b filas de $(1,0,\ldots,0)_{1\times n}$, N_2^b filas de $(0,1,\ldots,0)_{1\times n}$, ... , N_n^b filas de $(0,0,\ldots,1)_{1\times n}$. Entonces, para la b-ésima muestra, b = 1,2,...,B, la matriz de distancias al cuadrado, que será el input en el modelo basado en distancias, es:

$$\Delta_{\mathbf{F}}^{(2)}(b)=\boldsymbol{\gamma}(b)\Delta_{\mathbf{F}}^{(2)}\boldsymbol{\gamma}(b)^T. \tag{4.59}$$

Para cada muestra generada necesitamos obtener dos matrices de distancias, una que se corresponda con $k+1$ y otra con k predictores, derivadas de las dos matrices originales, la de $k+1$ y la k predictores. Por la misma razón, para cada estadístico necesitamos realizar dos inversas generalizadas.

Podemos utilizar también esta transformación para obtener la variable respuesta correspondiente a cada muestra a partir de la respuesta original mediante el producto:

$$\mathbf{Y}(b)=\gamma(b)\mathbf{Y}. \tag{4.60}$$

El *p*-valor asociado se estima calculando el percentil del estadístico Q de la muestra original, y aceptamos o rechazamos si la variable añadida explica suficiente variabilidad en el modelo, comparando este *p*-valor con un nivel de significación pre-establecido α^*.

Cabe notar que para el cálculo de los estadísticos asociados a cada remuestreo, hacemos uso del modelo ponderado, tal y como se refleja en el anexo 4.2, ya que cada muestra generada será de tamaño n, pero con individuos repetidos. Así, realizamos los cálculos con tamaños m_b para $b = 1,...,B$ iguales al número de individuos diferentes en cada caso, con $m_b \leq n$. Esto implica descomposiciones y cálculos de menor dimensión.

Si la muestra inicial procede de datos agregados, el cálculo de los *p*-valores es similar, tan sólo hay que tener en cuenta que el tamaño n proviene del sumatorio $n=\sum_{i=1}^{m} n_i$, por lo que a la hora de generar las muestras repartiremos las probabilidades del intervalo [0,1] de manera adecuada, es decir, en lugar de intervalos de igual amplitud $\frac{0}{n},\frac{1}{n},\frac{1+1}{n},\cdots,\frac{(n-1)+1}{n}$ crearemos intervalos siguiendo las amplitudes $\frac{0}{n},\frac{n_1}{n},\frac{n_1+n_2}{n},\cdots,\frac{\sum_{i=1}^{m-1} n_i + n_m}{n}$, realizaremos n tiradas y obtendremos el número de veces que aparece cada individuo de los m iniciales en cada intervalo. A su vez utilizando el modelo ponderado realizamos los cálculos descartando los individuos que no aparecen en los remuestreos, y obtendremos tamaños m_b' para $b=1,\ldots,B$ con $m_b' \leq m \leq n$.

4.4.3. Proceso de selección

Al igual que para el MLG, utilizamos un proceso de introducción paso a paso, que combina en cada etapa una fase introducción con una de eliminación.

Proceso de selección paso a paso

El proceso se inicia suponiendo que no disponemos de ninguna variable incorporada en el modelo, y combina en cada paso una fase de introducción, para ver qué variable es la siguiente a ser introducida, con una de eliminación para comprobar si alguna de las variables remanentes se ha vuelto no significativa. Fijemos un nivel de significación α^* y supongamos que disponemos en total de P factores potenciales de riesgo, $\mathbf{F}_1,\mathbf{F}_2,\ldots,\mathbf{F}_p$. Sea k el número de variables seleccionadas resultantes en el paso anterior, $\mathbf{F}(1),\mathbf{F}(2),\ldots,\mathbf{F}(k)$, con $k \leq P$. Definimos, de manera genérica, las fases de introducción y de eliminación para cada paso del proceso:

Fase de Introducción: Seleccionamos, al menos temporalmente, la variable que al ser introducida explique un mayor porcentaje de variabilidad mediante el menor p-valor estimado de entre las P - k variables aún no seleccionadas a partir de los correspondientes estadísticos $Q\left(\mathbf{F}_p \mid \mathbf{F}(1)\mathbf{F}(2)\cdots\mathbf{F}(k)\right)$. Entonces,

$$\mathbf{F}(k+1) = \left\{ \mathbf{F}_p \,/ \min_{p\in\{1,2,\ldots,P\}\setminus\{(1),(2),\ldots,(k)\}} p-valor\left(\mathbf{F}_p \mid \mathbf{F}(1)\mathbf{F}(2)\cdots\mathbf{F}(k)\right)\right\}. \tag{4.61}$$

Fase de Eliminación: Si alguna de las variables seleccionadas en las fases de introducción se vuelve no significativa en el modelo es eliminada. Para ello estimamos los p-valores, $p(i)$, de los correspondientes estadísticos $Q\left(\mathbf{F}(i)\mid\mathbf{F}(1),\ldots,\mathbf{F}(i-1),\mathbf{F}(i+1),\ldots,\mathbf{F}(k+1)\right)$ para $i=1,2,\ldots,k+1$, y elegimos el máximo p-valor al que nomenclamos $p(m)$,

$$p(m) = máx\left\{p(i)\right\}_{i=1,2,\ldots,k+1} \tag{4.62}$$

Entonces,

- Si $p(m) < \alpha^* \Rightarrow$ ninguna variable es eliminada
- Si $p(m) \geq \alpha^* \Rightarrow \mathbf{F}(m)$ es eliminada del modelo,
  - Si $(m) \neq (k+1)$ vamos al paso siguiente con el conjunto de variables $\left\{\mathbf{F}(1),\ldots,\mathbf{F}(m-1),\mathbf{F}(m+1),\ldots,\mathbf{F}(k+1)\right\}$.

- Si $(m) = (k+1)$ el proceso se detiene con el conjunto resultante de predictores $\{\mathbf{F}(1), \mathbf{F}(2), \ldots, \mathbf{F}(k)\}$.

4.4.4. Criterios de introducción de variables y validación en modelos lineales

Algunos de los criterios que encontramos en la bibliografía de selección de variables para el modelo de regresión lineal $\mathbf{Y}_{n\times 1} = \mathbf{X}_{n\times j}\boldsymbol{\beta}_{j\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, si suponemos que estamos analizando un modelo con j parámetros, para $j = 1, \ldots, m$ son [Peña (1990)]:

- *Coeficiente de determinación*: $R^2 = \dfrac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$

Es un mal criterio, pues aumenta al introducir variables sea cual sea su efecto, por lo que podríamos escoger modelos con variables innecesarias.

- *Coeficiente de determinación corregido*: $\overline{R}^2 = 1 - (1 - R^2)\dfrac{n-1}{n-j}$

Para corregir el defecto anterior se modifica de este modo, pero escoger mediante este criterio equivale a imponer una regla amplia de entrada de variables.

- *Error cuadrático medio*: $ECM = \dfrac{RSS}{n} = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n}$

El modelo con menor error cuadrático medio también es aquel que tiene un mayor coeficiente de correlación.

- *Varianza residual*: $S_R^2 = \dfrac{RSS}{n-j} = \dfrac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-j}$

El modelo con menor varianza residual es también el que tiene un mayor coeficiente de correlación corregido por grados de libertad.

Algunos otros métodos son los que minimizan los siguientes criterios sobre j [Rao y Wu (1989)]:

- $n\log\left(\frac{RSS(j)}{n}\right)+2j$ Akaike 1973

- $RSS(j)+2\frac{RSS(j)}{n-j}j$ Akaike 1970

- $RSS(j)+2\frac{RSS(m)}{n-m}j$ Mallows 1973

- $RSS(j)+\alpha\frac{RSS(m)}{n-m}j$ Shibata 1984

- $n\log\left(\frac{RSS(j)}{n}\right)+j\log n$ Schwartz 1978

- $n\log\left(\frac{RSS(j)}{n}\right)+jc\log\log n$ Hannan y Quinn 1979

- $n\log\left(\frac{RSS(j)}{n}\right)+jC_n$ Bai, Krishnaiah y Zhao 1986

donde c y α son constantes y C_n es tal que: $\frac{C_n}{n}\to 0$ y $\frac{C_n}{\log\log n}\to\infty$ cuando $n\to\infty$.

Métodos de validación cruzada

- Error cuadrático medio por validación *leave-one-out*: $ECMV=\frac{EC_V}{n}=\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_{(i)}\right)^2}{n}$

- La medida de robustez: $B^2=\frac{RSS}{EC_V}=\frac{\sum_{i=1}^{n}\left(y_i-\hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i-\hat{y}_{(i)}\right)^2}$

Esta medida, cuando las estimaciones son próximas a los valores tiende a 1, y cuando no lo son, tiende a cero.

- Métodos de validación cruzada *leave-*n_υ*-out* con $\frac{n_\upsilon}{n} \to 1$ [Shao (1993)].

Y alguna otra referencia: Kim y Hwang (2000), Miller (2000), y Peduzzi, Hardy y Holford (1980).

En el caso de la RBD, al igual que para el MLG, para la selección de variables, hay dos puntos en los que necesitamos criterio:

- cuando en una fase de introducción hemos de decidir cual será la siguiente variable a ser analizada,
- y cuando finalmente obtenemos un modelo y deseamos validarlo.

Tal y como proponemos, construimos los estadísticos (4.57) y (4.58) . Para el cálculo de los *p*-valores asociados necesitamos hacer uso de *bootstrap* para la estimación de las correspondientes distribuciones. En la práctica, estos criterios conllevan tiempo computacional, en especial para bases de datos relacionadas con carteras actuariales de responsabilidad civil de automóviles, en las que el número de individuos es elevado. Por ello analizamos alternativas que nos conduzcan a resultados similares con menores esfuerzos de cálculo.

Puesto que en la RBD, el input es la matriz de distancias entre individuos que contiene los perfiles de los individuos respecto a los predictores utilizados, es en esta matriz donde incorporamos la información de un predictor añadido. Los criterios diseñados para modelos lineales que minimizan sobre *j* alguna función no nos serán de utilidad, pues deseamos seleccionar predictores (observables), no las variables (latentes) que determinan las configuraciones euclídeas de la matriz de distancias.

Como ejemplo, utilizamos la aplicación 2 del capítulo 5, donde decidimos quedarnos con las variables: **Uso**, **Malus**, **Sexo** y **Vehicu**. Aprovechamos para observar el comportamiento de varios criterios y para la validación del modelo resultante. Realizamos cálculos para los 4 primeros pasos del proceso de selección.

En las columnas de las 4 tablas de resultados que detallamos a continuación aparecen los siguientes conceptos:

$$Geo0 = V(\Delta_Y),\ Geo1 = V^{k+1}(\hat{\Delta}_Y),\ Geo2 = V^{k}(\hat{\Delta}_Y),\ RSS = V(\Delta_Y) - V^{k+1}(\hat{\Delta}_Y),\ R^2 = \frac{V^{k+1}(\hat{\Delta}_Y)}{V(\Delta_Y)},$$

$$r^2 = \frac{V^{k+1}(\hat{\Delta}_Y) - V^k(\hat{\Delta}_Y)}{V(\Delta_Y) - V^k(\hat{\Delta}_Y)},\ Qparcial = \frac{V^{k+1}(\hat{\Delta}_Y) - V^k(\hat{\Delta}_Y)}{V(\Delta_Y) - V^{k+1}(\hat{\Delta}_Y)},\ Qglobal = \frac{V^{k+1}(\hat{\Delta}_Y)}{V(\Delta_Y) - V^{k+1}(\hat{\Delta}_Y)}.$$

Ge0 y *Ge1* no aportan información adicional, las ponemos a modo informativo. Puesto que en general *VT = VE + VNE*, en nuestro caso, en términos de variabilidades geométricas se cumplirá que *Geo0 = Geo2 + RSS*. También se cumplirá que a mayor R^2 mayor r^2, por lo que el primer bloque proporcinaría criterios equivalentes. Adjunto tenemos el estadístico parcial con el correspondiente *p*-valor estimado mediante *bootstrap*, concretamente con $B = 500$ remuestreos. Y análogamente el estadístico y *p*-valor global. Se utiliza también el error cuadrático de validación y la medida de robustez B^2:

Primer paso:

	Geo0	*Geo2*	*RSS*	R^2	*Q*	*p-valor*	EC_V	B^2
Potencia	3.8352E15	3.1533E14	3.5199E15	0.0822	0.0896	0.950	4.4348E15	0.7937
Antivehi	3.8352E15	1.4614E14	3.6891E15	0.0381	0.0396	0.948	4.0277E15	0.9159
Valorveh	3.8352E15	6.2518E14	3.2100E15	0.1630	0.1948	0.860	4.1735E15	0.7691
Edad	3.8352E15	2.8586E15	9.7665E14	0.7453	2.9271	0.900	7.2202E15	0.1353
Anticarn	3.8352E15	3.4883E15	3.4691E14	0.9096	10.055	0.814	6.0900E15	0.0569
Malus	3.8352E15	4.8640E13	3.7866E15	0.0127	0.0128	**0.502**	**3.8844E15**	**0.9748**
Sexo	3.8352E15	5.0182E12	3.8302E15	0.0013	0.0013	**0.516**	**3.8609E15**	**0.9920**
Zona	3.8352E15	3.6922E13	3.7983E15	0.0096	0.0097	0.948	3.9560E15	0.9602
Vehicu	3.8352E15	4.0732E12	3.8311E15	0.0011	0.0011	**0.814**	**3.8599E15**	**0.9925**
Uso	3.8352E15	7.2291E12	3.8280E15	0.0019	0.0019	**0.498**	**3.8588E15**	**0.9920**
Pago	3.8352E15	3.9494E12	3.8313E15	0.0010	0.0010	0.804	3.8719E15	0.9895

Segundo paso:

	Geo0	*Geo1*	*Geo2*	*RSS*	R^2	r^2	*Qparcial*	*p-v*	*Qglobal*	*p-v*	EC_V	B^2
Potencia	3.8352E15	7.2291E12	3.2042E14	3.5148E15	0.0835	0.0818	0.0891	0.946	0.0911	0.946	4.4478E15	0.7902
Antivehi	3.8352E15	7.2291E12	1.5382E14	3.6814E15	0.0401	0.0383	0.0398	0.942	0.0418	0.958	4.0344E15	0.9125
Valorveh	3.8352E15	7.2291E12	6.2589E14	3.2093E15	0.1632	0.1616	0.1928	0.856	0.1950	0.864	4.1886E15	0.7662
Edad	3.8352E15	7.2291E12	2.8736E15	9.6158E14	0.7493	0.7488	2.9809	0.900	2.9884	0.900	7.2504E15	0.1326
Anticarn	3.8352E15	7.2291E12	3.5023E15	3.3287E14	0.9132	0.9130	10.500	0.832	10.522	0.834	6.1065E15	0.0545
Malus	3.8352E15	7.2291E12	5.1302E13	3.7839E15	0.0133	0.0115	0.0116	**0.500**	0.0136	**0.536**	**3.8894E15**	**0.9729**
Sexo	3.8352E15	7.2291E12	1.2376E13	3.8228E15	0.0032	0.0013	0.0013	**0.518**	0.0032	**0.668**	**3.8675E15**	**0.9885**
Zona	3.8352E15	7.2291E12	4.4044E13	3.7912E15	0.0114	0.0096	0.0097	0.950	0.0116	0.960	3.9621E15	0.9569
Vehicu	3.8352E15	7.2291E12	1.1092E13	3.8241E15	0.0029	0.0010	0.0010	**0.816**	0.0029	**0.700**	**3.8671E15**	**0.9889**
Pago	3.8352E15	7.2291E12	1.0445E13	3.8248E15	0.0027	0.0008	0.0008	0.838	0.0027	0.796	3.8794E15	0.9859

Tercer paso:

	Geo0	*Geo1*	*Geo2*	*RSS*	R^2	r^2	*Qparcial*	*p-v*	*Qglobal*	*p-v*	EC_V	B^2
Potencia	3.8352E15	5.1302E13	3.3679E14	3.4984E15	0.0878	0.0754	0.0816	0.978	0.0963	0.966	4.7714E15	0.7332
Antivehi	3.8352E15	5.1302E13	1.9090E14	3.6443E15	0.0498	0.0369	0.0383	0.942	0.0524	0.940	4.0589E15	0.8979
Valorveh	3.8352E15	5.1302E13	6.6037E14	3.1748E15	0.1722	0.1609	0.1918	0.858	0.2080	0.872	4.2493E15	0.7471
Edad	3.8352E15	5.1302E13	2.8828E15	9.5239E14	0.7517	0.7483	2.9731	0.900	3.0269	0.904	7.2971E15	0.1305
Anticarn	3.8352E15	5.1302E13	3.5273E15	3.0788E14	0.9197	0.9186	11.290	0.836	11.457	0.840	6.2018E15	0.0496
Sexo	3.8352E15	5.1302E13	5.5910E13	3.7793E15	0.0146	0.0012	0.0012	**0.546**	0.0148	**0.604**	**3.8991E15**	**0.9693**
Zona	3.8352E15	5.1302E13	9.2812E13	3.7424E15	0.0242	0.0109	0.0111	0.948	0.0248	0.920	4.0095E15	0.9334
Vehicu	3.8352E15	5.1302E13	5.9419E13	3.7758E15	0.0155	0.0021	0.0021	**0.652**	0.0157	**0.568**	**3.8988E15**	**0.9684**
Pago	3.8352E15	5.1302E13	5.6124E13	3.7791E15	0.0146	0.0013	0.0013	0.806	0.0148	0.624	3.9123E15	0.9659

Cuarto paso:

	Geo0	*Geo1*	*Geo2*	*RSS*	R^2	r^2	*Qparcial*	*p-v*	*Qglobal*	*p-v*	EC_V	B^2
Potencia	3.8352E15	5.5910E13	3.3949E14	3.4957E15	0.0885	0.0750	0.0811	0.982	0.0971	0.972	4.7804E15	0.7313
Antivehi	3.8352E15	5.5910E13	1.9865E14	3.6366E15	0.0518	0.3778	0.0393	0.942	0.0546	0.930	4.0640E15	0.8948
Valorveh	3.8352E15	5.5910E13	6.6166E14	3.1735E15	0.1725	0.1603	0.1909	0.858	0.2085	0.886	4.2654E15	0.7440
Edad	3.8352E15	5.5910E13	2.8988E15	9.3642E14	0.7558	0.7522	3.0359	0.902	3.0956	0.900	7.4420E15	0.1258
Anticarn	3.8352E15	5.5910E13	3.5307E15	3.0450E14	0.9206	0.9194	11.411	0.842	11.595	0.854	6.2738E15	0.0485
Zona	3.8352E15	5.5910E13	9.6180E14	3.7390E15	0.0251	0.0107	0.0108	0.942	0.0257	0.938	4.0218E15	0.9297
Vehicu	3.8352E15	5.5910E13	6.4895E14	3.7703E15	0.0169	0.0024	0.0024	**0.624**	0.0172	**0.630**	**3.9079E15**	**0.9648**
Pago	3.8352E15	5.5910E13	6.1408E14	3..7738E15	0.0160	0.0015	0.0015	0.794	0.0163	0.682	3.9216E15	0.9623

Como era de esperar, el primer bloque no nos es de utilidad como criterio de introducción ni de validación. Los siguientes *p*-valores parciales y globales son los criterios que nosotros proponemos como "buenos". Observamos como los menores valores del EC_V y los mayores de B^2 son los asociados a los menores *p*-valores como cabía esperar. Por un lado, confirmamos que el criterio del *p*-valor calculado mediante *bootstrap* nos conduce a un resultado apropiado, pero por otro no hemos encontrado un algoritmo que nos reduzca el tiempo computacional del proceso, ya que el método *leave-one-out* consiste en eliminar uno por uno a los individuos, lo que nos lleva a estimar 455 modelos diferentes.

Adicionalmente realizamos la representación de la respuesta, **Y**, versus la estimación, **Yhat**, y un gráfico de residuos, **Y-Yhat**, versus la estimación, **Yhat**:

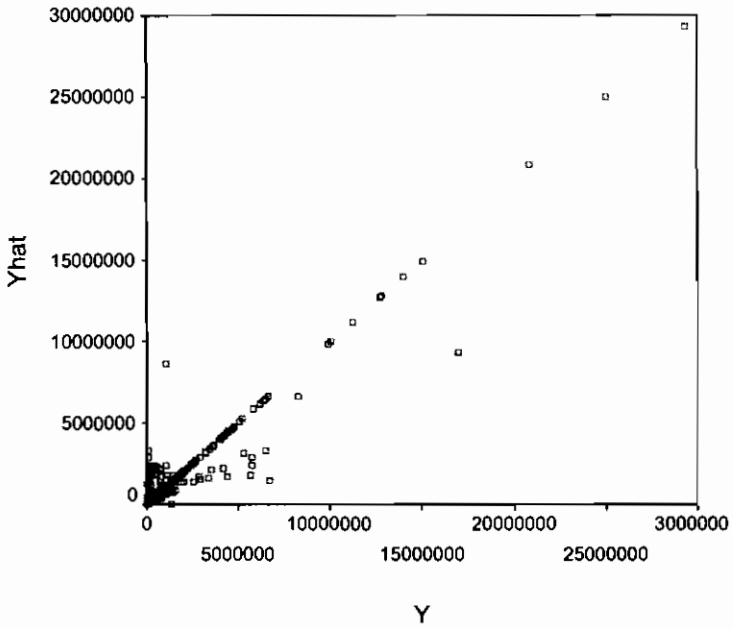


Figura 4.1. Gráfico de la respuesta *versus* la estimación.

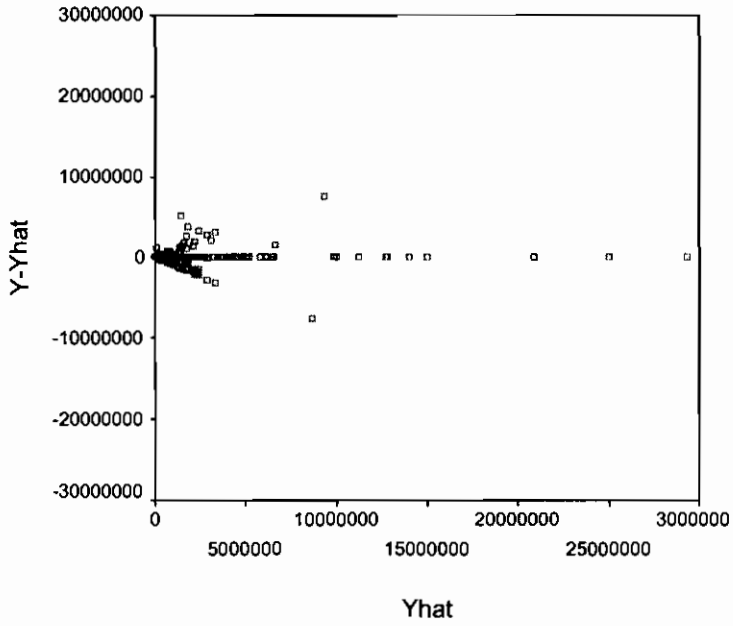


Figura 4.2. Gráfico de los residuos *versus* la estimación.

4.5. Implementación y software utilizado

Hemos utilizado el programa `octave` [http://www.octave.org/] para implementar los programas informáticos que nos permiten llevar a la práctica lo expuesto en este capítulo. Los códigos de los programas se encuentran en el anexo 4.2.

Con los programas del anexo 4.2 hemos realizado los cálculos de las aplicaciones del capítulo 5 en lo referente a este capítulo. Y adicionalmente hemos utilizado el lenguaje de programación para llevar a cabo labores que el programa SPSS no permite, al menos de una manera fácil. Por ejemplo, el programa cummodel.m nos agrega los datos a partir de un conjunto de predictores categóricos, y el programa poisson.m calcula el estadístico (3.62) del capítulo 3, referente a la validación según Albrecht del modelo Poisson. Este estadístico, Q, necesita simultáneamente operar con el número de siniestros individual y la estimación de la frecuencia de siniestralidad obtenida con datos agregados. Sin la ayuda de un lenguaje de programación hubiera sido imposible llevar a cabo el presente trabajo.

El programa `octave` no tiene, en principio, restricción explícita en el número de pólizas ni en el de factores de riesgo. Sin embargo, para una cartera real con un número de pólizas del orden de 10^6, por ejemplo, sería necesario:

⇒ Reemplazar los prototipos realizados en lenguaje interpretado (`octave`) por un programa ad-hoc en algún lenguaje compilado (fortran/C++) y

⇒ Realizar un muestreo, posiblemente estratificado, del conjunto de individuos a fin de evaluar el modelo y decidir los predictores significativos con un gasto computacional practicable.

ANEXO 4.1. Funciones de distancias entre individuos

Distancia		Métrica	Euclídea
(D1)	$d_{ij}=\left(\frac{1}{p}\sum_{k=1}^{p}\left(x_{ik}-x_{jk}\right)^2\right)^{1/2}$	Sí	Sí
(D2)	$d_{ij}=\left(\frac{1}{p}\sum_{k=1}^{q}\left(\frac{x_{ik}-x_{jk}}{r_k}\right)^2\right)^{1/2}$	Sí	Sí
(D3)	$d_{ij}=\left(\frac{1}{p}\sum_{k=1}^{p}\frac{\left\|x_{ik}-x_{jk}\right\|}{r_k}\right)^{1/2}$	Sí	No
(D4)	$d_{ij}=\left(\frac{1}{p}\sum_{k=1}^{p}\left(\frac{x_{ik}-x_{jk}}{r_k}\right)^q\right)^{1/q}$	Sí	No
(D5)	$d_{ij}=\left(\mathbf{x}_i-\mathbf{x}_j\right)^T\boldsymbol{\Sigma}^{-1}\left(\mathbf{x}_i-\mathbf{x}_j\right)$	Sí	Sí
(D6)	$d_{ij}=\left(\frac{1}{p}\sum_{k=1}^{p}\frac{\left(x_{ik}-x_{jk}\right)^2}{\left(x_{ik}+x_{jk}\right)^2}\right)^{1/2}$	Sí (No)	Sí (No)
(D7)	$d_{ij}=\frac{1}{p}\sum_{k=1}^{p}\frac{\left\|x_{ik}-x_{jk}\right\|}{\left\|x_{ik}+x_{jk}\right\|}$	Sí (No)	No
(D8)	$d_{ij}=\frac{1}{p}\sum_{k=1}^{p}\frac{\left\|x_{ik}-x_{jk}\right\|}{\left\|x_{ik}\right\|+\left\|x_{jk}\right\|}$	Sí	No
(D9)	$d_{ij}=\frac{\sum_{k=1}^{p}\left\|x_{ik}-x_{jk}\right\|}{\sum_{k=1}^{p}\left(x_{ik}+x_{jk}\right)}$	No	No
(D10)	$d_{ij}=\frac{\sum_{k=1}^{p}\left\|x_{ik}-x_{jk}\right\|}{\sum_{k=1}^{p}\mathrm{Max}\left(x_{ik},x_{jk}\right)}$	Sí (No)	No
(D11)	$d_{ij}=\frac{1}{p}\sum_{k=1}^{p}\left(1-\frac{\mathrm{Min}\left(x_{ik},x_{jk}\right)}{\mathrm{Max}\left(x_{ik},x_{jk}\right)}\right)$	Sí (No)	No

Tabla 4.1. Distancias para variables cuantitativas.

Similaridad	Rango	Métrica	Euclídea	Autor
(S1) $s_{ij} = \frac{a}{b+c}$	$0,\infty$			Kulczynsky
(S2) $s_{ij} = \frac{a}{a+b+c+d}$	0,1	Sí	Sí	Russell y Rao
(S3) $s_{ij} = \frac{a}{a+\frac{1}{2}(b+c)+d}$	0,1	Sí	Sí	Sorensen
(S4) $s_{ij} = \frac{a}{a+\theta(b+c)}$	0,1	Sí, si $\theta \geq \frac{1}{3}$	Sí, si $\theta \geq \frac{1}{2}$	$\theta = 1$ Jaccard $\theta = 2$ Sokal-Sneath
(S5) $s_{ij} = \frac{a+d}{a+\theta(b+c)+d}$	0,1	Sí, si $\theta \geq \frac{1}{3}$	Sí, si $\theta \geq 1$	$\theta = 1$ Sokal y Michener $\theta = 2$ Rogers y Tanimoto
(S6) $s_{ij} = \frac{a-(b+c)+d}{a+b+c+d}$	-1,1	Sí	Sí	Harman
(S7) $s_{ij} = \frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}\right)$	0,1	No	No	Kulczynsky
(S8) $s_{ij} = \frac{1}{2}\left(\frac{a}{a+b}+\frac{a}{a+c}+\frac{d}{c+d}+\frac{d}{b+d}\right)$	0,1	No	No	Anderberg
(S9) $s_{ij} = \frac{a}{\sqrt{(a+b)(a+c)}}$	0,1	Sí	Sí	Ochiai
(S10) $s_{ij} = \frac{ad}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	0,1	Sí	Sí	
(S11) $s_{ij} = \frac{ad-bc}{\sqrt{(a+b)(a+c)(b+d)(c+d)}}$	-1,1	Sí	Sí	Pearson
(S12) $s_{ij} = \frac{ad-bc}{ad+bc}$	-1,1	No	No	Yule
(S13) $s_{ij} = \frac{2a}{(a+b)(a+c)}$	$0,\frac{2}{p}$	Sí	Sí	Dice

Tabla 4.2. Similaridades para variables binarias.

ANEXO 4.2. Anexo informático

➢ bootw.m

```
# calcula las multiplicidades de B muestras de tamano n=sum(W)
# con m=rows(W) respuestas diferentes procedentes del modelo acumulado
# para el modelo con pesos, o m=n i W=ones(1:n)' para el no ponderado
# function [multiw]=bootw(B,W)
function [multiw]=bootw(B,W)
empty_elements_ok=0;
n=sum(W);
m=rows(W);
W=W./n;
multiw=zeros(m,B);
for j=1:B,
      for k=1:n,
            alea=rand(1,1);
            if ((alea>=0)&&(alea<=W(1))),
                  multiw(1,j)=multiw(1,j)+ 1;
            endif
            for i=2:m,
                  if ((alea>=(sum(W(1:i-1))))&&(alea<=(sum(W(1:i))))),
                        multiw(i,j)=multiw(i,j)+ 1;
                  endif
            endfor
      endfor
endfor
endfunction
```

➢ chol1.m

```
#  function [T,P,II,r]=chol1(A)
#
#  Modified Cholesky algorithm for positive semidefinte matrices.
#
#
#  Let A be a symmetric, positive semidefinite matrix, of rank r.
#  There exists a permutation matrix P such that P'*A*P has a
#  unique Cholesky decomposition, which takes the form
#
#            P'*A*P= T*T',     T = | T_11 0 |
#                                  | T_21 0 |,
#
#  where T_11 is an (r,r) lower triangular matrix with positive
```

```
# diagonal elements.
#
# Input:
#
# A,  a symmetric, positive semidefinite matrix
#
# Output:
#
# T,  lower triangular matrix (as described above,
#       after omitting the null columns)
# P,  permutation matrix such that P'*A*P=T*T'
# II, the permutation associated to P (P = E(:,II)), where E is the
#     identity matrix.
# r,  rank of A
#
# Note:
#
# The builtin constant epsilon is used to check for almost-zero pivot
# ------------------------------------------------------------------
function [T,P,II,r]=chol1(A)
  epsilon=1.0e-7;
  n=length(A);
  II=1:n;
  T=zeros(n,n);
# ------------------------------------------------------------------
# Search pivot: the largest diagonal element p of the remaining
# box.
# ------------------------------------------------------------------
  for k=1:n
    dk=diag(A(k:n,k:n));
    p=max(dk);
    if p<=epsilon,
      r=k-1;
# msg=sprintf("chol1: end algorithm, rank = %d",r);
# disp(msg);
      break
    else
      for i=k:n,
        if A(i,i)>=p,
          s=i;
          break
        endif
      endfor
      if s>k,
        JJ=1:n;JJ(k)=s;JJ(s)=k;
        t=II(k);II(k)=II(s);II(s)=t;
        A=A(JJ,JJ);
```

```
        T=T(JJ,:);
      endif
# ----------------------------------------------------------------
#  Nonnull entries in the k-th row of T equal those in the k-th row
#  of A, divided by sqrt(A(k,k)).
#  The product Tk'*Tk is subtracted from the box A(k+ 1:n,k+ 1:n)
# ----------------------------------------------------------------
      T(k,k)=sqrt(A(k,k));
      if k<n
        t=k+ 1;
        Tk=A(t:n,k)/T(k,k);
        T(t:n,k)=Tk;
        A(t:n,t:n)=A(t:n,t:n)-Tk*Tk';
      else
        r=n;
      endif
    endif
  endfor
  P=eye(n);
  P=P(:,II);
  T=T(:,1:r);
endfunction
```

➢ **cummodel.m**

```
# function [YCUM,XCUM,W]=cummodel(Y,X,N,op)
# función que prepara el modelo acumulado para op 1=db 2=glm
# inputs: Y=respuesta (indx1)
#         X=predictores categóricos (indxp) los valores ir n
#         para el predictor p-ésimo desde 1 hasta N(p)
#         N=vector con el n£mero de clases de cada predictor (px1)
#         op---> 1=db  2=glm
# outputs: YCUM=respuesta acumulada (cellfull,1)
#          XCUM=predictores correspondientes, si es db con los valores
#          originales de los predictores (cellfullxp)
#          y si es glm con las dummies (cellfullxsum(N))
#          W=vector de pesos que indica cuántos individuos hemos sumado
#          para cada ratingcell.
function [YCUM,XCUM,W]=cummodel(Y,X,N,op)
p=rows(N);
ind=rows(Y);
W=zeros(ind,1);
YCUM=zeros(ind,1);
XCUM=zeros(ind,p);
cont=1;
W(cont,1)=1;
```

```
YCUM(cont,1)=Y(cont,1);
XCUM(cont,:)=X(cont,:);
for i=2:ind,
      control=0;
      for j=1:i-1,
            f=X(i,:)';g=XCUM(j,:)';
            match=sum(f==g);
            if p==match,
                  YCUM(j,1)=YCUM(j,1)+ Y(i,1);
                  W(j,1)=W(j,1)+ 1;
                  control=1;
            endif
      endfor
      if control==0,
            cont=cont+ 1;
            XCUM(cont,:)=X(i,:);
            YCUM(cont,1)=Y(i,1);
            W(cont,1)=1;
      endif
endfor
YCUM=YCUM(1:cont,1);
XCUM=XCUM(1:cont,:);
W=W(1:cont,1);
if op>=2,
      dum=sum(N);
      XDUM=zeros(cont,dum);
      a=0;
      for i=1:p,
            for j=1:N(i),
                  a=a+ 1;
                  XDUM(:,a:a)=XCUM(:,i:i);
                  for h=1:cont,
                        ab=XDUM(h:h,a:a);
                        ac=j;
                        ad=sum(ab==ac);
                        if ad>=1,
                              XDUM(h:h,a:a)=1;
                        else
                              XDUM(h:h,a:a)=0;
                        endif
                  endfor
            endfor
      endfor
      XCUM=zeros(cont,dum);
      XCUM(:,:)=XDUM(:,:);
endif
endfunction
```

➢ estadg.m

```
# function [Qv]=estadg(multiw,d2,y,ele)
function [Qv]=estadg(multiw,d2,y,ele)
n=sum(multiw(:,1:1));
m=rows(y);
B=columns(multiw);
geomvar2=zeros(B,1);
geomvar0=zeros(B,1);
for z=1:B,
      f=multiw(:,z); g=zeros(m,1);
      siz=sum([f!=g]);
      M=zeros(siz,m);
      wm=zeros(siz,1);
      a=1;
      for j=1:m,
            if multiw(j,z)!=0,
                  M(a,j)=1;
                  wm(a,1)=multiw(j,z);
                  a=a+ 1;
            endif
      endfor
      dm2=zeros(siz,siz);
      ym=zeros(siz,1);
      ym=M*y;
      dm2=M*d2*M';
      wm=wm./n;
      geomvar0(z)=n*(sum(wm.*((ym-(wm'*ym)).*(ym-(wm'*ym)))));
      [geomvar2(z)]=rbdp(dm2,ym,ele,wm,n);
endfor
Qv=(geomvar2)./(geomvar0-geomvar2);
endfunction
```

➢ estadp.m

```
# function [Q]=estadp(multiw,d1,d2,y,ele,p)
function [Q]=estadp(multiw,d1,d2,y,ele,p)
n=sum(multiw(:,1:1));
m=rows(y);
B=columns(multiw);
geomvar1=zeros(B,1);
geomvar2=zeros(B,1);
geomvar0=zeros(B,1);
for z=1:B,
      f=multiw(:,z); g=zeros(m,1);
      siz=sum([f!=g]);
```

```
        M=zeros(siz,m);
        wm=zeros(siz,1);
        a=1;
        for j=1:m,
            if multiw(j,z)!=0,
                M(a,j)=1;
                wm(a,1)=multiw(j,z);
                a=a+ 1;
            endif
        endfor
        dm1=zeros(siz,siz);
        dm2=zeros(siz,siz);
        ym=zeros(siz,1);
        ym=M*y;
        dm1=M*d1*M';
        dm2=M*d2*M';
        wm=wm./n;
        geomvar0(z)=n*(sum(wm.*((ym-(wm'*ym)).*(ym-(wm'*ym)))));
        [geomvar2(z)]=rbdp(dm2,ym,ele,wm,n);
        if p>=2,
          [geomvar1(z)]=rbdp(dm1,ym,ele,wm,n);
        endif
    endfor
Q=(geomvar2-geomvar1)./(geomvar0-geomvar2);
endfunction
```

➢ **gow.m**

```
# function d=gow(a,b,pc,pq,pb,rc)
#
# Computes the squared distance associated with Gower's similarity
# coefficient between two vectors a, b.
# There are pc continuous variables, pq qualitative variables, and
# pb binary variables. We assume that they are ordered as (c,q,b)
#
# Input: Two vectors, a, b, of equal size p =pc+ pq+ pb
#       pc = number of continuous variables
#       pq = number of qualitative variables
#       pb = number of binary variables
#       rc = a vector (1,pc) containing the ranges=(max-min)
#           for the continuous variables.
#
# Output: The squared distance between a and b.
#
function d=gow(a,b,pc,pq,pb,rc)
  p=length(a);
```

```
  nmatch=0;
  nposmatch=0;
  sc=0;
  nnegmatch=0;
# ------------------------------------------------------------
# Quantitative variables
# ------------------------------------------------------------
  if pc>0,
    ac=a(1:pc);bc=b(1:pc);
    sc=sum(ones(1,pc)-abs(ac-bc)./rc);
  endif

# ------------------------------------------------------------
# Qualitative variables
# ------------------------------------------------------------
  if pq>0,
    aq=a(pc+ 1:pc+ pq);bq=b(pc+ 1:pc+ pq);
    nmatch=sum(aq==bq);
  endif
# ------------------------------------------------------------
# Binary variables
# ------------------------------------------------------------
  if pb>0,
     ab=a(pc+ pq+ 1:p);bb=b(pc+ pq+ 1:p);
     nposmatch=sum(ab.*bb);
     nnegmatch=(sum((ab-bb)==0)-sum(ab.*bb));
  else
    nposmatch=0;
    nnegmatch=0;
  endif
# ------------------------------------------------------------
  if nnegmatch==p,
      s=0;
  else
      s=(sc+ nmatch+ nposmatch)/(p-nnegmatch);
# Para la distancia (5.1)
#       s=(sc+ nmatch+ nposmatch)/(p);
  endif
  d=1-s;
endfunction
```

➢ **gower.m**

```
#   function dist=gower(x,ncont,ncat,nbin)
function dist=gower(x,ncont,ncat,nbin)
if ncont>0,
```

```
 x1=x(:,1:ncont);
 Gh=zeros(1:ncont);
 for j=1:ncont,
     Gh(j)=(max(x1(:,j:j))-min(x1(:,j:j)));
 endfor
else
 Gh=0;
endif
n=rows(x);
dist=zeros(n,n);
for i=1:n-1,
     for j=i+ 1:n,
          dist(i,j)=gow(x(i:i,:),x(j:j,:),ncont,ncat,nbin,Gh);
          dist(j,i)=dist(i,j);
     endfor
endfor
endfunction
```

- **pbdp.m**

```
#  function [Yhat]=pbdp(Dx,Y,ele,W)
function [Yhat]=pbdp(Dx,Y,ele,W)
  empty_list_elements_ok = 0;
  [m,m1]=size(Dx);
  Yhat=zeros(size(Y));
  Y0=Y-(W'*Y);
  K=eye(m)-ones(m,1)*W';
  Gx=-0.5*K*Dx*K';
  Dw=diag(W);
  gx=diag(Gx);
  Gx=(Dw.^ (1/2))*Gx*(Dw.^ (1/2));
  if ele==1,
  [Tx,Pix,a,b]=chol1(Gx);
  Kx=inv(Tx'*Tx);
  Gxplus=Pix*Tx*Kx*Kx*Tx'*Pix';
  else
  Gxplus=pinv(Gx);
  endif
  unom=ones(1:m)';
  Yhat=(W'*Y)+ 0.5*((unom*(gx'))-Dx)*(Dw.^ (.5))*Gxplus*(Dw.^ (.5))*Y0;
endfunction
```

- **pnuevo.m**

```
#  function yhat=pnuevo(Y,W,Dx,dxi,ele)
```

```
function yhat=pnuevo(Y,W,Dx,dxi,ele)
empty_list_elements_ok = 0;
[m,m1]=size(Dx);
Y0=Y-(W'*Y);
K=eye(m)-ones(m,1)*W';
Gx=-0.5*K*Dx*K';
Dw=diag(W);
gx=diag(Gx);
Gx=(Dw.^ (1/2))*Gx*(Dw.^ (1/2));
if ele==1,
  [Tx,Pix,a,b]=chol1(Gx);
  Kx=inv(Tx'*Tx);
  Gxplus=Pix*Tx*Kx*Kx*Tx'*Pix';
else
  Gxplus=pinv(Gx);
endif
yhat=(W'*Y)+ 0.5*(gx-dxi)'*(Dw.^ (.5))*Gxplus*(Dw.^ (.5))*Y0;
endfunction
```

- **poisson.m**

```
# function [Q,Qw,Qd]=poisson(yn,X,ymeanm,yhatm,Xcum)
# función que calcula los estadísticos del apartado 3.5.3.1.1 de Albretch (1983a).
function [Q,Qw,Qd]=poisson(yn,X,ymeanm,yhatm,Xcum)
p=columns(X);
n=rows(X);
m=rows(Xcum);
yhatn=zeros(n,1);
ymeann=zeros(n,1);
for u=1:m,
      for i=1:n,
            f=X(i,:)';g=Xcum(u,:)';
            match=sum(f==g);
            if p==match,
               yhatn(i,1)=yhatm(u,1);
               ymeann(i,1)=ymeanm(u,1);
            endif
      endfor
endfor
Q=sum(((yn-yhatn).*(yn-yhatn))./yhatn);
Qw=sum(((yn-ymeann).*(yn-ymeann))./yhatn);
Qd=sum(((ymeann-yhatn).*(ymeann-yhatn))./yhatn);

endfunction
```

- **pvalg.m**

```
# function [pvv,Qmv,R22,geo0,geo2]=
#            =pvalg(Ycum,W,Xcum,ncont,ncat,nbin,multiw,ele)
#
# Realiza el contraste (4.58)
# Inputs:
#  Ycum es la media o bien Y directamente
#  Xcum o bien X directamente
#  W son los pesos que no est n entre 0 y 1
#  multiw=matriz(mxB) con las multiplicidades de las B muestras
#  ele vale:
#          -> 0 si queremos pseudoinversa
#          -> 1 si queremos Cholesky modificado
# Outputs: pvv,Qmv,R22,geo0,geo2

function [pvv,Qmv,R22,geo0,geo2]=
=pvalg(Ycum,W,Xcum,ncont,ncat,nbin,multiw,ele)

empty_list_elements_ok=0;
m=rows(Ycum);
B=columns(multiw);
p=columns(Xcum);
n=sum(W);
W=W./n;
geo0=n*(sum(W.*((Ycum-(W'*Ycum)).*(Ycum-(W'*Ycum)))));
d2=zeros(m,m);
d2=gower(Xcum,ncont,ncat,nbin);
[geo2]=rbdp(d2,Ycum,ele,W,n);
[Qv]=estadg(multiw,d2,Ycum,ele);
Qmv=(geo2)/(geo0-geo2);
v=0;
for i=1:B,
      if Qv(i,1)<=Qmv,
            v=v+1;
      endif
endfor
pvv=1-(v/B);
R22=geo2/geo0;
endfunction
```

- **pvalp.m**

```
#    function [pv,Qm,R21,R22,r2,geo0,geo1,geo2]=
#                =pvalp(Ycum,W,Xcum,ncont,ncat,nbin,addbin,multiw,ele,lo)
#
```

```
# Realiza el contraste (4.57)
# Inputs:
#  Ycum es la media o bien Y directamente
#  Xcum o bien X directamente [cont¡nuas categ¢ricas binarias]
#  W son los pesos que no est n entre 0 y 1
#  multiw=matriz(mxB) con las multiplicidades de las B muestras
#  ele vale:
#         -> 0 si queremos pseudoinversa
#         -> 1 si queremos Cholesky modificado
#  lo vale:
#         -> 0 si la variable a¤adida es cont¡nua
#         -> 1 si es categórica
#         -> 2 si es (o son) binaria(s)
#
#  addbin: numero de binarias incluidas de una vez
#
# Outputs: pv,Qm,R21,R22,r2,geo0,geo1,geo2

function [pv,Qm,R21,R22,r2,geo0,geo1,geo2]=
=pvalp(Ycum,W,Xcum,ncont,ncat,nbin,addbin,multiw,ele,lo)

empty_list_elements_ok=0;
m=rows(Ycum);
B=columns(multiw);
p=columns(Xcum);
n=sum(W);
W=W./n;
geo0=n*(sum(W.*((Ycum-(W'*Ycum)).*(Ycum-(W'*Ycum)))));
geo1=0;
d2=zeros(m,m);
d1=zeros(m,m);
ncont1=ncont;
ncat1=ncat;
nbin1=nbin;
if p>1,
      if lo==0,
            if ncat+ nbin==0,
                  X1cum(:,1:ncont-1)=Xcum(:,1:ncont-1);
            elseif ncont==1,
                  X1cum(:,1:p-1)=Xcum(:,2:p);
            else
                  X1cum(:,1:ncont-1)=Xcum(:,1:ncont-1);
                  X1cum(:,ncont:p-1)=Xcum(:,ncont+ 1:p);
            endif
            ncont1=ncont1-1;
      elseif lo==1,
            if ncont+ nbin==0,
```

```
                X1cum(:,1:ncat-1)=Xcum(:,1:ncat-1);
            elseif (ncont==0)&(ncat==1),
                X1cum(:,1:nbin)=Xcum(:,2:p);
            elseif nbin==0,
                X1cum(:,1:p-1)=Xcum(:,1:p-1);
            else
                X1cum(:,1:ncont+ ncat-1)=Xcum(:,1:ncont+ ncat-1);
                X1cum(:,ncont+ ncat:p-1)=Xcum(:,ncont+ ncat+ 1:p);
            endif
            ncat1=ncat1-1;
        elseif lo==2,
            if ncont+ ncat+ nbin==addbin,
                nbin=1;
            else
                X1cum=Xcum(:,1:p-addbin);
                nbin1=nbin1-addbin;
            endif
        endif
    endif
    p=ncont+ ncat+ nbin;
    d2=gower(Xcum,ncont,ncat,nbin);
    [geo2]=rbdp(d2,Ycum,ele,W,n);
    if p>1,
        d1=gower(X1cum,ncont1,ncat1,nbin1);
        [geo1]=rbdp(d1,Ycum,ele,W,n);
    endif
    [Q]=estadp(multiw,d1,d2,Ycum,ele,p);
    Qm=(geo2-geo1)/(geo0-geo2);
    u=0;
    for i=1:B,
        if Q(i,1)<=Qm,
            u=u+ 1;
        endif
    endfor
    pv=1-(u/B);
    R21=geo1/geo0;
    R22=geo2/geo0;
    r2=(geo2-geo1)/(geo0-geo1);
    endfunction
```

- **rbdp.m**

```
    #   function [geomvar]=rbdp(Dx,Y,ele,W,n)
    function [geomvar]=rbdp(Dx,Y,ele,W,n)
      empty_list_elements_ok = 0;
      [m,m1]=size(Dx);
```

```
  Yhat=zeros(size(Y));
  Y0=Y-(W'*Y);
  K=eye(m)-ones(m,1)*W';
  Gx=-0.5*K*Dx*K';
  Dw=diag(W);
  gx=diag(Gx);
  Gx=(Dw.^ (1/2))*Gx*(Dw.^ (1/2));
  if ele==1,
  [Tx,Pix,a,b]=chol1(Gx);
  Kx=inv(Tx'*Tx);
  Gxplus=Pix*Tx*Kx*Kx*Tx'*Pix';
  else
  Gxplus=pinv(Gx);
  endif
  unom=ones(1:m)';
  Yhat=(W'*Y)+0.5*((unom*(gx'))-Dx)*(Dw.^ (.5))*Gxplus*(Dw.^ (.5))*Y0;
geomvar=n*(sum(W.*((Yhat-(W'*Yhat)).*(Yhat-(W'*Yhat)))));
endfunction
```

- **yhat.m**

```
# function [Yhat]=yhat(Ycum,W,Xcum,ncont,ncat,nbin,ele)
# Realiza la predicción de Y, (4.49) para Δ2(2) = Δ(2).
function [Yhat]=yhat(Ycum,W,Xcum,ncont,ncat,nbin,ele)
empty_list_elements_ok=0;
m=rows(W);
n=sum(W);
W=W./n;
d2=zeros(m,m);
d2=gower(Xcum,ncont,ncat,nbin);
[Yhat]=pbdp(d2,Ycum,ele,W);
endfunction
```

- **ynuevo.m**

```
# function yhat=ynuevo(Ycum,W,Xcum,xnou,ncont,ncat,nbin,ele)
# xnou vector fila
# Realiza la predicción para un nuevo individuo, (4.48).
function yhat=ynuevo(Ycum,W,Xcum,xnou,ncont,ncat,nbin,ele)
empty_list_elements_ok=0;
m=rows(W);
n=sum(W);
W=W./n;
d2=zeros(m+1,m+1);
```

```
Xcum=[xnou;Xcum];
d2=gower(Xcum,ncont,ncat,nbin);
dxi=d2(2:m+ 1,1:1);
d22=zeros(m,m);
d22=d2(2:m+ 1,2:m+ 1);
yhat=pnuevo(Ycum,W,d22,dxi,ele);
endfunction
```

- **valida.m**

```
# function [geo0,geo2,rss,ecv,B2]=valida(Ycum,W,Xcum,ncont,ncat,nbin,ele)
# Calcula la RSS, el ECV y el coeficiente B2.
function [geo0,geo2,rss,ecv,B2]=valida(Ycum,W,Xcum,ncont,ncat,nbin,ele)
empty_list_elemants_ok = 1;
Yhat=zeros(size(Ycum));
n=sum(W);
m=rows(Ycum);
d2=zeros(m,m);
d2=gower(Xcum,ncont,ncat,nbin);
W=W./n;
Yhat=pbdp(d2,Ycum,ele,W);
geo2=n*(sum(W.*((Yhat-(W'*Yhat)).*(Yhat-(W'*Yhat)))));
rss=n*(sum(W.*((Ycum-Yhat).*(Ycum-Yhat))));
geo0=n*(sum(W.*((Ycum-(W'*Ycum)).*(Ycum-(W'*Ycum)))));
Yhat=zeros(size(Ycum));
  if W(1:1,:)==1/n,
   I0=[2:m];
   Wi=W.*n./(n-1);
   Wi=Wi(I0,:);
  else
   I0=[1:m];
   Wi=W.*n./(n-1);
   Wi(1:1,:)=Wi(1:1,:)-(1/(n-1));
  endif
  Yi=Ycum(I0,:);
  Dxi=d2(I0,I0);
  ui=d2(:,1);
  dxi=ui(I0);
  Yhat(1,:)=pnuevo(Yi,Wi,Dxi,dxi,ele);
  if W(m:m,:)==1/n,
   I0=[1:m-1];
   Wi=W.*n./(n-1);
   Wi=Wi(I0,:);
  else
   I0=[1:m];
   Wi=W.*n./(n-1);
```

```
    Wi(m:m,:)=Wi(m:m,:)-(1/(n-1));
   endif
   Yi=Ycum(I0,:);
   Dxi=d2(I0,I0);
   ui=d2(:,m);
   dxi=ui(I0);
   Yhat(m,:)=pnuevo(Yi,Wi,Dxi,dxi,ele);
for i=2:m-1,
   if W(i:i,:)==1/n,
    I0=[1:i-1,i+1:m];
    Wi=W.*n./(n-1);
    Wi=Wi(I0,:);
   else
    I0=[1:m];
    Wi=W.*n./(n-1);
    Wi=Wi(I0,:);
    Wi(i:i,:)=Wi(i:i,:)-(1/(n-1));
   endif
   Yi=Ycum(I0,:);
   Dxi=d2(I0,I0);
   ui=d2(:,i);
   dxi=ui(I0);
   Yhat(i,:)=pnuevo(Yi,Wi,Dxi,dxi,ele);
endfor
ecv=n*(sum(W.*((Ycum-Yhat).*(Ycum-Yhat))));
B2=rss/ecv;
endfunction
```

Interacciones

- **pvalpi.m**

```
# function [pv,Qm,R21,R22,r2,geo0,geo1,geo2]=pvalpi(Ycum,W,Xa,Xb,multiw)
# Realiza el contraste (4.57), en referencia a la aplicación 5.1, haciendo
# uso de la fórmula (4.32)
# Con Xa = Antigüedad  y  Xb = Estado
function [pv,Qm,R21,R22,r2,geo0,geo1,geo2]=pvalpi(Ycum,W,Xa,Xb,multiw)
empty_list_elements_ok=0;
m=rows(Ycum);
B=columns(multiw);
n=sum(W);
W=W./n;
geo0=n*(sum(W.*((Ycum-(W'*Ycum)).*(Ycum-(W'*Ycum)))));
geo1=0;
da=zeros(m,m);
```

```
db=zeros(m,m);
da=gower(Xa,0,1,0);
db=gower(Xb,0,1,0);

# Para p-valor(A:E)
p=1;
# Para p-valor(A:E|A)
p=2;
# Para p-valor(A:E|A,E), p-valor(A|A:E,E) y p-valor(E|A:E,A)
p=3;

if p>1,
     [geo1]=rbdpi1(da,db,Ycum,0,W,n);
endif
[geo2]=rbdpi2(da,db,Ycum,0,W,n);
[Q]=estadpi(multiw,da,db,Ycum,0,p);
Qm=(geo2-geo1)/(geo0-geo2);
u=0;
for i=1:B,
      if Q(i,1)<=Qm,
            u=u+ 1;
      endif
endfor
pv=1-(u/B);
R21=geo1/geo0;
R22=geo2/geo0;
r2=(geo2-geo1)/(geo0-geo1);
endfunction
```

- **estadpi.m**

```
# function [Q]=estadpi(multiw,da,db,y,ele,p)
function [Q]=estadpi(multiw,da,db,y,ele,p)
n=sum(multiw(:,1:1));
m=rows(y);
B=columns(multiw);
geomvar1=zeros(B,1);
geomvar2=zeros(B,1);
geomvar0=zeros(B,1);
for z=1:B,
      f=multiw(:,z); g=zeros(m,1);
      siz=sum([f!=g]);
      M=zeros(siz,m);
      wm=zeros(siz,1);
      a=1;
      for j=1:m,
```

```
            if multiw(j,z)!=0,
                M(a,j)=1;
                wm(a,1)=multiw(j,z);
                a=a+ 1;
            endif
        endfor
        dma=zeros(siz,siz);
        dmb=zeros(siz,siz);
        ym=zeros(siz,1);
        ym=M*y;
        dma=M*da*M';
        dmb=M*db*M';
        wm=wm./n;
        geomvar0(z)=n*(sum(wm.*((ym-(wm'*ym)).*(ym-(wm'*ym)))));
        [geomvar2(z)]=rbdpi2(dma,dmb,ym,ele,wm,n);
        if p>=2,
          [geomvar1(z)]=rbdpi1(dma,dmb,ym,ele,wm,n);
        endif
 endfor
Q=(geomvar2-geomvar1)./(geomvar0-geomvar2);
endfunction
```

- **rbdpi1.m**

```
#  function [geomvar]=rbdpi1(Da,Db,Y,ele,W,n)
function [geomvar]=rbdpi1(Da,Db,Y,ele,W,n)
  empty_list_elements_ok = 0;
  [m,m1]=size(Da);
  Yhat=zeros(size(Y));
  Y0=Y-(W'*Y);
  K=eye(m)-ones(m,1)*W';
  unom=ones(1:m)';
  Ga=-0.5*K*Da*K';
  Gb=-0.5*K*Db*K';

  # Para p-valor(A:E|A)
    Gx=Ga;
  # Para p-valor(A:E|A,E)
    Gx=Ga+ Gb;
  # Para p-valor(A|A:E,E)
    Gx=Ga.*Gb+ Gb;
  # Para p-valor(E|A:E,A)
    Gx=Ga.*Gb+ Ga;

  Dw=diag(W);
  gx=diag(Gx);
```

```
  Dx=gx*unom'+ unom*gx'-2.*Gx;
  Gx=(Dw.^ (1/2))*Gx*(Dw.^ (1/2));
  if ele==1,
  [Tx,Pix,a,b]=chol1(Gx);
  Kx=inv(Tx'*Tx);
  Gxplus=Pix*Tx*Kx*Kx*Tx'*Pix';
  else
  Gxplus=pinv(Gx);
  endif
  Yhat=(W'*Y)+0.5*((unom*(gx'))-Dx)*(Dw.^ (.5))*Gxplus*(Dw.^ (.5))*Y0;
geomvar=n*(sum(W.*((Yhat-(W'*Yhat)).*(Yhat-(W'*Yhat)))));
endfunction
```

- **rbdpi2.m**

```
#   function [geomvar]=rbdpi2(Da,Db,Y,ele,W,n)
function [geomvar]=rbdpi2(Da,Db,Y,ele,W,n)
  empty_list_elements_ok = 0;
  [m,m1]=size(Da);
  Yhat=zeros(size(Y));
  Y0=Y-(W'*Y);
  K=eye(m)-ones(m,1)*W';
  unom=ones(1:m)';
  Ga=-0.5*K*Da*K';
  Gb=-0.5*K*Db*K';
  # Para p-valor(A:E)
    Gx=Ga.*Gb;
  # Para p-valor(A:E|A)
    Gx=Ga+Ga.*Gb;
  # Para p-valor(A:E|A,E), p-valor(A|A:E,E) y p-valor(E|A:E,A)
    Gx=Ga+Gb+Ga.*Gb;
  Dw=diag(W);
  gx=diag(Gx);
  Dx=gx*unom'+ unom*gx'-2.*Gx;
  Gx=(Dw.^ (1/2))*Gx*(Dw.^ (1/2));
  if ele==1,
  [Tx,Pix,a,b]=chol1(Gx);
  Kx=inv(Tx'*Tx);
  Gxplus=Pix*Tx*Kx*Kx*Tx'*Pix';
  else
  Gxplus=pinv(Gx);
  endif
  Yhat=(W'*Y)+0.5*((unom*(gx'))-Dx)*(Dw.^ (.5))*Gxplus*(Dw.^ (.5))*Y0;
geomvar=n*(sum(W.*((Yhat-(W'*Yhat)).*(Yhat-(W'*Yhat)))));
endfunction
```

Capítulo 5

Aplicaciones prácticas

En este capítulo ilustramos la aplicabilidad de las tres metodologías estadísticas presentadas con detalle en los capítulos 3 y 4 para la selección de variables de tarifa: el análisis de segmentación, el modelo lineal generalizado y la regresión basada en distancias. Para ello hacemos uso de los datos presentados en el apartado 2.3, en el cual se analizan sus características generales: tipo de información (agregada o desagregada), variable de siniestralidad a analizar (cuantía por siniestro o número de siniestros), y los predictores disponibles.

Dado que el tema del trabajo tiene una aplicación real importante, el de ser el primer paso para la obtención del precio del seguro, no podemos quedarnos en el plano teórico sino que es necesario comprobar las dificultades o peculiaridades de los cálculo reales con las diferentes técnicas. Para ello utilizamos cuatro conjuntos de datos distintos que se desarrollan en las cuatro aplicaciones en que se divide el capítulo.

La primera aplicación utiliza unos datos sobre las cuantías de los siniestros de los impagos de préstamos de una entidad financiera, cuya característica principal es su simplicidad. La tercera aplicación utiliza unos datos sobre las cuantías de siniestros para la cobertura de daños propios del seguro del automóvil, también simples pero agregados y por lo tanto ponderados. Estas dos aplicaciones se utilizan principalmente para comprobar el correcto funcionamiento en la práctica de la metodología de selección propuesta para la RBD.

Las aplicaciones segunda y cuarta utilizan datos reales de España sobre carteras del seguro del automóvil. La segunda se refiere a la cuantía por siniestro para daños personales y la cuarta al número de siniestros para daños materiales. Al ser datos reales, se trata de ficheros complicados que reflejan la realidad de los estudios de tarificación que deben realizar las compañías de seguros. El objetivo de analizar estos dos conjuntos de datos es el de aplicar e ilustrar las metodologías estudiadas en el trabajo a unos datos complejos, para lo que necesitamos tener en cuenta las características peculiares que en España tienen los seguros de automóviles que hemos detallado en el capítulo 2. El análisis de

los resultados, nos permite obtener unas conclusiones que, aunque restringidas a las carteras analizadas, son muy interesantes y pueden ser, con los debidos matices, extrapolables al seguro del automóvil en España.

Resumimos a continuación la estructura de las cuatro aplicaciones.

Aplicación 1. Impagos de préstamos: cuantía de un siniestro

Utilizamos los datos descritos en el apartado 2.3.4 (tabla 2.3) que hacen referencia a impagos de préstamos. Analizamos la cuantía por siniestro haciendo uso de información desagregada con sólo dos factores cualitativos. Estos datos son muy manejables, por lo que con ellos realizamos varios estudios. Inicialmente estudiamos las asociaciones entre pares de variables, al igual que hacemos en el resto de aplicaciones. Y posteriormente, aplicamos las metodologías de selección:

Respecto al AS: En el apartado 3.6.1 hemos analizado, con estos datos, diferentes maneras de discretización de la variable respuesta continua, cuantía de un siniestro, con el objetivo de aplicar posteriormente las técnicas de segmentación que utilizan repuesta categórica, tanto ordinal como nominal. En esta aplicación realizamos un estudio de la sensibilidad del AS respecto al tipo de algoritmo utilizado, al número de clases en que se ha discretizado a la respuesta (aprovechando las realizadas en el apartado 3.6.1) y a cambios en los parámetros implicados en el proceso jerárquico de cluster que representan las técnicas de segmentación. Una vez obtenemos el árbol que consideramos más correcto podemos estimar la cuantía de un siniestro con simplemente la media aritmética de las cuantías en los segmentos terminales. Otra alternativa es la de utilizar los segmentos obtenidos, los cuales son homogéneos respecto a la variable de siniestralidad estudiada, como los grupos de tarifa a incorporar directamente en un modelo de regresión, sin pasar entonces por un proceso de selección. En esta aplicación los utilizamos concretamente para configurar los grupos de partida del modelo de credibilidad de Bülhmann-Straub, el cual aplicamos a estos datos para la estimación de primas.

Respecto al MLG: En el apartado 3.5.2.3 hemos estudiado la manera de codificar a los predictores cualitativos mediante variables binarias para ser utilizados como *input* en el MLG. En esta aplicación, por ser la primera, detallamos el proceso de creación de variables binarias para el tratamiento de tales predictores así como sus interacciones. Aunque lo usual será realizar la selección de variables con un

modelo dado (fijada una distribución del error y una función de enlace), y posteriormente, con los predictores seleccionados, estudiar el ajuste de otros modelos, en esta aplicación hemos realizado ambas cosas: hemos realizado los procesos con diferentes modelos y posteriormente los hemos validado. Nos hemos extendido para ilustrar diferentes aspectos del MLG.

Respecto a la RBD: En este apartado ilustramos como, a diferencia del MLG, en la RBD tenemos una mayor gama de tratamientos para los predictores cualitativos. Ésta es la primera aplicación en la que estudiamos el funcionamiento práctico de la metodología de selección propuesta, por lo que utilizamos como función de distancias el coeficiente de coincidencias, ya que con éste y dando tratamiento categórico a los predictores, deberíamos obtener el mismo resultado que con regresión clásica utilizando tantas variables binarias como clases tengan los factores. En el apartado del MLG hemos realizado el proceso de selección con la combinación de error Normal y el enlace identidad, el cual desemboca también al mismo modelo clásico. Adicionalmente, ya que ésta es la única aplicación en la que incluimos los términos de interacción, realizamos un estudio del tratamiento de estos en la RBD.

Para finalizar, dedicamos un apartado para ilustrar, tanto con el MLG como con la RBD, la problemática que lleva asociada el disponer de datos censurados.

Aplicación 2. Cartera C1 de responsabilidad civil de automóviles: cuantía de un siniestro para daños personales

Utilizamos los datos descritos en el apartado 2.3.1, que hacen referencia a la cartera C1 de responsabilidad civil del automóvil. Analizamos la cuantía por siniestro para daños personales a partir de información desagregada. Los factores de riesgo son de tipo mixto.

Para la selección de variables de tarifa: respecto al MLG utilizamos el modelo de distribución del error Gamma junto con el enlace logarítmico, y respecto a la RBD utilizamos como función de distancia el índice de similaridad de Gower.

En el apartado 4.4.4 encontramos la validación exhaustiva del modelo resultante de la RBD.

En el apartado 3.6.2, a modo de ejemplo, habíamos procedido a la discretización de los predictores

continuos de estos datos. El resultado de tal discretización se hubiera podido utilizar alternativamente para los dos modelos de regresión estudiados, aunque en esta aplicación les hemos dado tratamiento continuo. Sin embargo, para aplicar el AS, sí necesitamos partir de una discretización inicial y amplia de los factores continuos. Esta discretización inicial será la misma que la inicial empleada en el apartado 3.6.2. Realizamos varios árboles de segmentación en función de diferentes parámetros, utilizando en cualquier caso el algoritmo ordinal de SPSS y una discretización de las cuantías realizada con Ward para 31categorías.

Aplicación 3. Datos de Baxter referentes al seguro del automóvil: cuantía de un siniestro para daños propios

Utilizamos los datos descritos en el apartado 2.3.3 (tabla 2.2) que hacen referencia al seguro del automóvil. Analizamos la cuantía por siniestro para la cobertura de daños propios. Los datos de partida son agregados, a partir de sólo tres factores cualitativos. Son datos que suelen utilizarse para ilustrar el uso del MLG en el campo actuarial, tal y como exponemos en la aplicación. Puesto que son bastante manejables y ésta es la primera aplicación en la que tratamos con datos agregados y por lo tanto ponderados, nos sirve, al igual que la aplicación 1, para estudiar el comportamiento de la metodología de selección propuesta para la RBDP, en este caso ponderado. Realizamos adicionalmente dos árboles de segmentación utilizando el algoritmo ordinal de SPSS y una discretización de la respuesta con Ward para 31 categorías, en los que la única diferencia es la definición de monótono o libre para los predictores.

Aplicación 4. Cartera C2 de responsabilidad civil de automóviles: número de siniestros para daños materiales

Utilizamos los datos descritos en el apartado 2.3.2, que hacen referencia a la cartera C2 de responsabilidad civil del automóvil. Analizamos el número de siniestros para daños materiales. Son datos desagregados, y los factores de riesgo están todos inicialmente discretizados. Ésta es la única aplicación en la que estudiamos el número de siniestros.

En un primer apartado realizamos una agrupación de zonas, ya que la agrupación inicial queda demasiado dispersa para estos datos, que tan sólo suponen un 10% de las pólizas de la cartera total. En un segundo apartado asignamos las marcas de clase para los factores cuantitativos para el cálculo de las medidas de asociación entre pares de variables.

Para poder realizar el estudio con la frecuencia de siniestralidad procedemos primero, de manera detallada, al proceso de agregación de los datos. Una vez agregados, aplicamos a la frecuencia de siniestralidad el MLG de estructura de error Poisson ponderado, y por tanto con un parámetro de dispersión $\phi = 1$, junto con el enlace logarítmico. Realizamos la selección de predictores y estudiamos el cumplimiento de las hipótesis de Poisson, de homogeneidad y de independencia para el modelo resultante, hecho que queda recogido en el parámetro de dispersión, tal y como explicamos en el capítulo 3. Realizamos el test de dispersión obteniendo que el modelo es disperso. Suponemos un parámetro constante y por lo tanto igual para cada celda, cuya estimación resulta $\hat{\phi} = 0.8259$, por lo que obtenemos una infra dispersión.

Respecto a la RBD utilizamos el índice de similaridad de Gower como función de distancias en el análisis de la frecuencia de siniestralidad, ya que hemos dado tratamiento cuantitativo a la potencia y al nivel de bonus, y por lo tanto estamos analizando predictores de tipo mixto.

Realizamos el AS, utilizando el algoritmo ordinal del SPSS, directamente sobre el número de siniestros desagregado, utilizando la discretización de partida de los factores de riesgo.

5.1. Aplicación 1. Impagos de préstamos: cuantía de un siniestro

Los datos relativos a esta aplicación son los descritos en el apartado 2.3.4 (tabla 2.3 del anexo 2.3), y hacen referencia a impagos de préstamos. Tal y como se comentado en tal apartado, tan sólo podemos analizar el comportamiento de la cuantía por siniestro. La información la podemos resumir, de manera agregada, mediante una tabla cruzada de $3 \times 3 = 9$ combinaciones, tabla 2.4 del anexo 2.3. Estos datos son bastante manejables y aunque los factores son categóricos los datos están desagregados, lo que nos permite tener cierto margen de maniobrabilidad.

5.1.1. Relaciones entre pares de variables

Asociación de las cuantías con cada predictor

Calculamos (3.4),

- Cuantías y Estado: $\eta = 0.1682$
- Cuantías y Antigüedad: $\eta = 0.2572$
- Cuantías y Antigüedad*Estado[25]: $\eta = 0.3186$

y obtenemos que las relaciones no son altas, y que la antigüedad está más asociada que el estado. Al calcular la asociación entre las cuantías y la variable "interacción", Antigüedad*Estado, obtenemos tan sólo $\eta = 0.3186$. Esto es así, ya que los factores contemplados no nos proporcionan una homogeneidad perfecta dentro de las celdas. A modo de complemento realizamos un análisis de la varianza para cada predictor:

Antigüedad

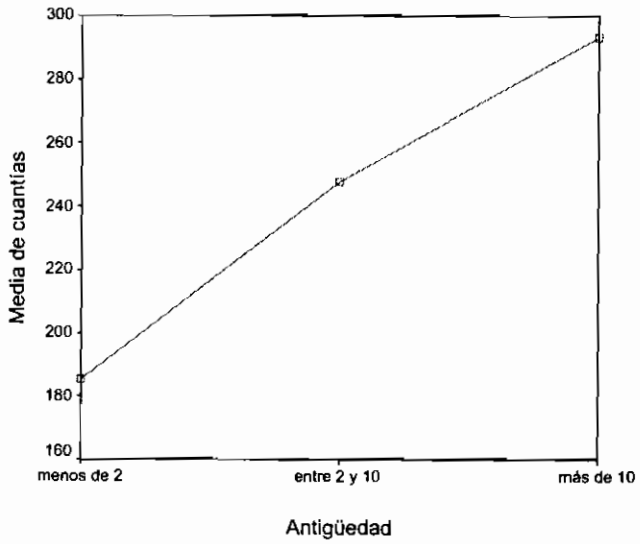
ANOVA

cuantías

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	781684.324	2	390842.162	14.102	.000
Intra-grupos	11030952.3	398	27715.960		
Total	11812636.6	400			

[25] Utilizamos el símbolo * para denotar a la variable categórica que representa a la interacción de ambas variables, es decir, la variable categórica con el código de las celdas resultantes de cruzar a las variables.

Gráfico de las medias



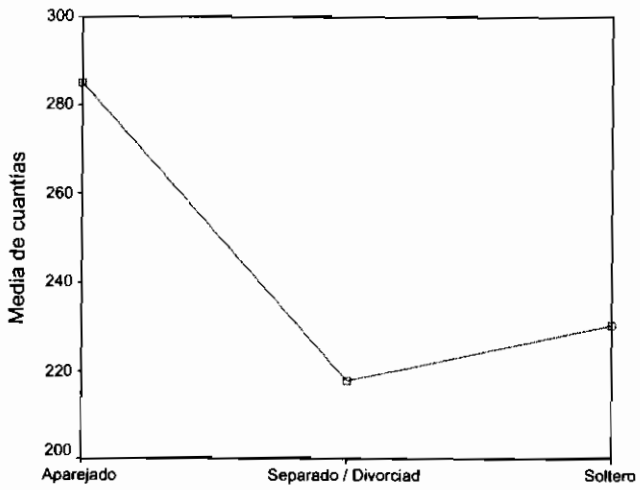
Estado

ANOVA

cuantías

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	334221.983	2	167110.991	5.794	.003
Intra-grupos	11478414.6	398	28840.238		
Total	11812636.6	400			

Gráfico de las medias



A partir de los gráficos de medias, observamos que a mayor antigüedad en el puesto laboral, mayor cuantía en el impago. Este hecho podría explicarse en que a mayor confianza en la estabilidad, se solicitan créditos de mayor importe. Del mismo modo, si el individuo es casado, los importes son mayores. Posiblemente, si estudiáramos la frecuencia de siniestralidad, ésta iría en dirección contraria.

Si fijamos un nivel de significación del 5% observamos que los p-valores resultantes de las anovas nos indican que hay diferencias en las medias de las clases, y que las diferencias en la antigüedad son algo mayores que en el estado.

Asociación predictor con predictor

Calculamos (3.6), (3.7), (3.8) y la correlación canónica r_1,

- Estado y Antigüedad: $C = 0.0358$, $T = 0.0358$, $P = 0.0474$, $r_1 = 0.0448$

y obtenemos que los predictores no están muy relacionados entre sí con cualquiera de los coeficientes calculados. Éste es un punto positivo para la correcta aplicación del AS, de este modo evitaremos la paradoja de Simpson. También es adecuado en los modelos de regresión para no introducir información redundante en el caso de incorporar ambos predictores.

El objetivo es la obtención de los grupos de tarifa, pero con la finalidad de acabar estimando el riesgo. Sin ningún tipo de análisis, una primera estimación es la media aritmética global, 242.17, realizada con los 401 casos, o bien las medias de la tabla cruzada de cada combinación, tabla 2.4.

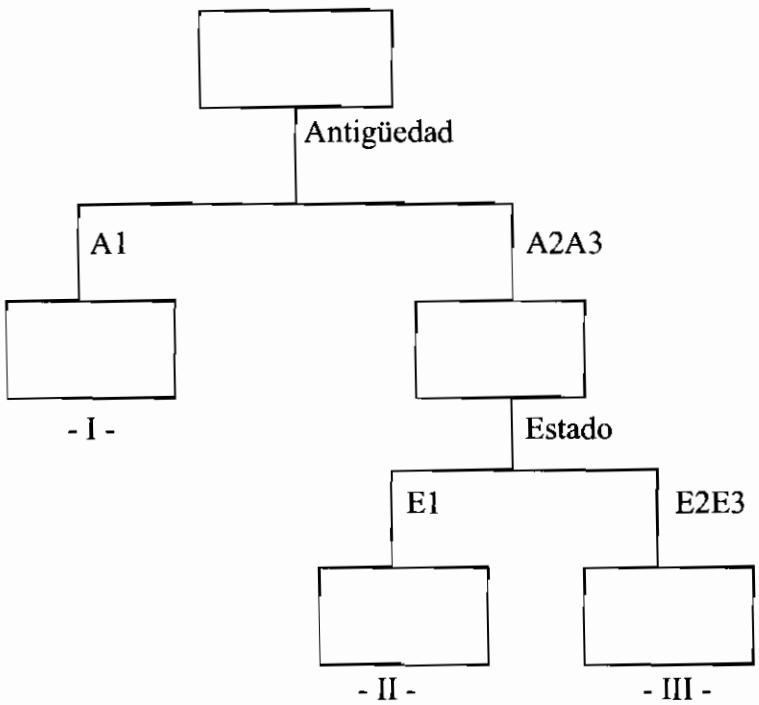
5.1.2. Análisis de segmentación

Puesto que la respuesta, cuantías de los siniestros, es continua, para la aplicación de CHAID hemos procedido a discretizarla, tal y como se describe en el apartado 3.6.1. Aquí, hacemos uso de las discretizaciones de Ward para 10 y 31 categorías, y de la de la variedad de *k-means* para 4 categorías. En todas las aplicaciones de AS que pasamos a detallar, la antigüedad ha sido definida como monótona (si era definida como libre para observar la agrupación obtenida, el resultado era el mismo) y el estado como libre.

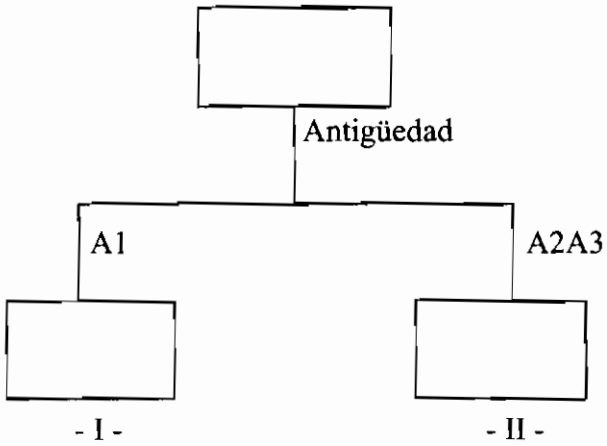
Como parámetros comunes hemos definido: el nivel de significación permitido para la fase de agrupación de categorías y de selección del mejor predictor del 5%; el tamaño mínimo para analizar un nodo de 50 individuos; el tamaño mínimo para formar un grupo de 25 individuos; como mucho 3 niveles en el árbol (dejando así llegar hasta el final); como mucho 20 grupos terminales (que en este caso era imposible); ajuste de Bonferroni en la fase de selección del "mejor" predictor para tener en cuenta la fase de agrupación de categorías, y en el caso de la FIRM el mínimo entre éste y la comparación múltiple; respecto a la constante A a añadir al denominador de la *ji*-cuadrado (3.13) en CATFIRM, A = 0 (*ji*-cuadrado estándar de Pearson) y A = 0.5, y en cualquier caso aproximación no-asintótica en el cálculo de los *p*-valores; y en SPSS siempre el *likelihood ratio* (3.11) ó (3.12).

Resultando los siguientes árboles de segmentación:

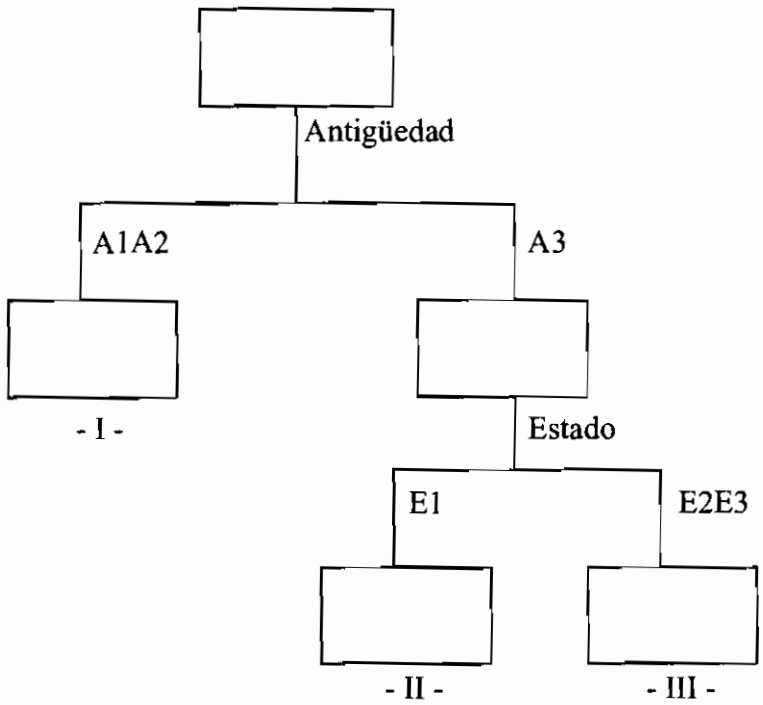
Árbol 1



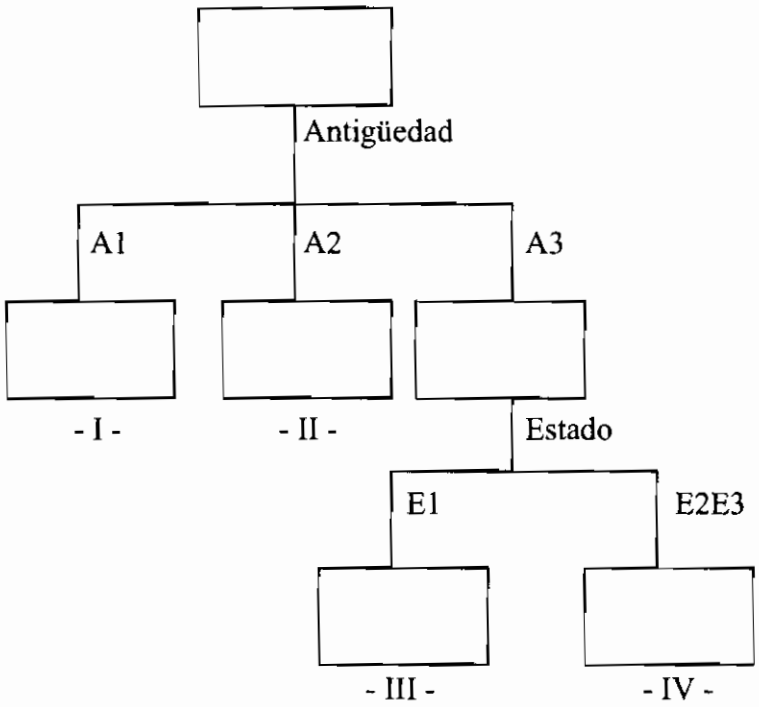
Árbol 2



Árbol 3



Árbol 4



El árbol 1 se corresponde con

- CATFIRM con k-means 4 categorías y A = 0 y 0.5
- CATFIRM con Ward 10 categorías y A = 0
- CHAID nominal con k-means 4 y Ward 10 categorías
- CHAID ordinal con k-means 4 categorías

El árbol 2 se corresponde con

- CATFIRM con Ward 10 categorías y A = 0.5

El árbol 3 se corresponde con

- CHAID nominal con Ward 31 categorías

El árbol 4 se corresponde con

- CONFIRM para variable respuesta continua
- CHAID ordinal con Ward 10 y 31 categorías

Comentarios y conclusiones

Todo este conjunto nos sirve para realizar un estudio comparativo del efecto de discretizar la respuesta en un número diferente de clases, de dar diferentes valores a los parámetros implicados en los algoritmos y de ver hasta qué punto es correcto considerar la respuesta continua como categórica:

⇒ Árbol1: Observamos como, tanto si reducimos en exceso la información (k = 4), como si utilizamos los algoritmos para respuesta nominal, los resultados son diferentes del resto. Si hemos de decidir entre información o dispersión de las tablas del proceso, debemos inclinarnos por la no pérdida de información al ser la respuesta de naturaleza continua, por lo que este árbol no lo consideraremos como válido.

⇒ Árbol 2: Observamos que en CATFIRM con 10 categorías y A = 0.5, el estado no tiene poder de discriminación ni en los de antigüedad de más de 2 años, ni en los de menos de 2 años, y por tanto han quedado sólo 2 segmentos terminales, los de antigüedad de menos de 2 años y los de más de 2 años. Esto ha sido así pues al introducir A = 0.5, hemos reducido la tendencia a formar grupos pequeños, tendencia que suele tener la *ji*-cuadrado estándar de Pearson, A = 0.

⇒ Árbol 3: En el caso del CHAID para 31 categorías nominal, obtenemos 3 segmentos terminales: los de antigüedad menor a 10 años; los de antigüedad mayor a 10 años que están aparejados; y los de antigüedad mayor a 10 años que estén separados/divorciados o solteros. En este caso, el algoritmo nominal ha decidido que el estado afecta a los de antigüedad mayor a 10 años de la misma forma que en el árbol 1, pero en un primer estadio ha decidido agrupar las categorías 1 y 2 de la antigüedad. Este es un caso aislado del resto de procesos nominales debido al excesivo número de categorías, 31, en que se ha dividido a las cuantías.

⇒ Árbol 4: En el CONFIRM y el CHAID para 31 y 10 categorías ordinal, en un primer estadio la antigüedad no ha agrupado sus categorías y el estado sólo ha afectado a los de antigüedad mayor a 10 años. Por lo tanto, aquí los segmentos terminales han sido 4: los de antigüedad menor de 2 años; los de antigüedad entre 2 y 10 años; los de antigüedad de más de 10 años que están aparejados; y los de antigüedad de más de 10 años que están separados/divorciados o solteros. Cabe destacar, que el CONFIRM es el algoritmo adecuado para la variable respuesta impagos de naturaleza continua. Como observamos, los resultados han sido los mismos utilizando el algoritmo para respuesta categórica ordinal, que recoge algo de información respecto a las puntuaciones (que en este caso eran las medias de las clases a partir de los datos originales) representativas de las clases formadas para k = 31 y 10 grupos.

Vemos por lo tanto, que con los mismos datos, los resultados obtenidos, según el algoritmo (XAID o CHAID) y según los parámetros establecidos para el proceso, no son los mismos. Nuestro interés es el de seleccionar el árbol más correcto para estos datos. Éste es el árbol 4, pues se corresponde al tratamiento de respuesta continua (CONFIRM). Hemos visto que obtenemos los mismos resultados con el tratamiento de categórica ordinal (CHAID ordinal) con el número máximo de categorías que permite el programa, 31, incluso con 10. El tratamiento de datos en el trabajo se realiza con SPSS. Puesto que el CHAID ordinal es un módulo de SPSS fácil de manejar, el resto de aplicaciones referentes a AS, serán calculadas con éste. Remarcando, que en las siguientes tres aplicaciones disponemos siempre de respuesta cuantitativa, por lo que para su aplicación procederemos a discretizarla con Ward 31, y los resultados presentados en los nodos serán las medias de dicha discretización junto con el número de casos. Notando que si la intención es la estimación del riesgo, las medias se deberían corresponder con las de la respuesta original, fácilmente calculables. A pesar de todo, puesto que las puntuaciones utilizadas son las medias en cada cluster de los datos originales, la media global de la discretización también se corresponde con la media global de la respuesta original.

Detallamos en la siguiente tabla los resultados visualizados en los nodos terminales para la aplicación con CONFIRM de XAID:

	A1	A2	A3
E1	**I** N = 133 Media = 185.33 Desv. Estándar = 137.92 Mín. = 5.05 Máx. = 665.58	**II** N = 135 Media = 247.80 Desv. Estándar = 173.22 Mín. = 12.11 Máx. = 748.68	**III** N = 44 Media = 366.61 Desv. Estándar = 218.99 Mín. = 48.56 Máx. = 814.95
E2			**IV** N = 89 Media = 257.07 Desv. Estándar = 153.90 Mín. = 10.29 Máx. = 563.36
E3			

Total
N = 401
Media = 242.17
Desv. Estándar = 171.85
Mín. = 5.05
Máx. = 814.95

Una vez tenemos los grupos exhaustivos y homogéneos respecto al riesgo, la primera posibilidad en la estimación de la siniestralidad, son las medias aritméticas de cada segmento terminal, las cuales están plasmadas en la tabla de resultados que acabamos de detallar.

Modelo de credibilidad

Como ya hemos comentado, el AS nos puede servir como paso previo para otras técnicas. Nosotros lo utilizamos para configurar los grupos de partida del modelo de credibilidad de *Bülhmann-Straub*. El modelo aplica un promedio ponderado de la media de la experiencia individual y de la experiencia de todo el colectivo:

$$z_j \cdot \bar{x}_j + (1 - z_j) \cdot \bar{x} \text{ para } j = 1,2,3,4$$

donde z_j es el factor de credibilidad, que es una variable que recoge la ponderación de la media de cada grupo. Éste depende del tamaño del grupo, de la varianza interna del grupo y de la varianza entre grupos: $0 \leq z_j \leq 1$. Valores altos de z_j implican una significación alta del grupo por separado. Un buen resumen de las características de este modelo de credibilidad, junto con abundantes citas, puede encontrarse en Goovaerts y Hoogstad (1987) y Pons (1995).

Los resultados de la aplicación para cada segmento son:

Segmento	Factor de credibilidad
I:	$z_1 = 0.9478470046$
II:	$z_2 = 0.9485799125$
III:	$z_3 = 0.8573988926$
IV:	$z_4 = 0.9240224682$

Observamos que todos los factores de credibilidad son altos, cercanos a 1. Esto indica que los 4 grupos de tarifa son homogéneos internamente y diferentes entre sí. Las diferencias entre los factores de credibilidad de cada grupo son debidas a los distintos tamaños.

Empíricamente hemos comprobado que aplicando el modelo de Bülhmann-Straub a cualquier otra agrupación posible de las 9 combinaciones iniciales da, en su conjunto, factores de credibilidad inferiores a los aquí obtenidos. La aplicación de otros modelos de credibilidad, como el *two-way* o el jerárquico de *Jewell* [Bermúdez y Pons (1997)], proporcionan una tabla cruzada cuyos factores de credibilidad son, en su conjunto, también inferiores a los aquí obtenidos.

La estimación con este modelo para cada grupo de tarifa es:

	A1	A2	A3
E1	**I**	**II**	**III** 348.86
E2	185.30	247.51	**IV** 255.93
E3			

5.1.3. Modelo lineal generalizado

La codificación utilizada para los predictores es la siguiente:

⇒ Para el Estado, utilizamos dos variables binarias, al igual que en (3.47), dejando la clase E3 para el efecto global,

$$X_{E1} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in E1 \end{cases} \qquad X_{E2} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in E2 \end{cases}$$

⇒ Para la Antigüedad, al igual que en (3.48), utilizamos dos variables binarias, dejando la clase A3 para el efecto global,

$$X_{A1} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in A1 \end{cases} \qquad X_{A2} = \begin{cases} 0 & \text{en otro caso} \\ 1 & \text{si } i \in A2 \end{cases}$$

⇒ Y para la interacción, las cuatro binarias asociadas,

$$X_{A1E1}=\begin{cases}0 & \text{en otro caso}\\ 1 \text{ si } & i\in A1\cap E1\end{cases} \qquad X_{A1E2}=\begin{cases}0 & \text{en otro caso}\\ 1 \text{ si } & i\in A1\cap E2\end{cases}$$

$$X_{A2E1}=\begin{cases}0 & \text{en otro caso}\\ 1 \text{ si } & i\in A2\cap E1\end{cases} \qquad X_{A2E2}=\begin{cases}0 & \text{en otro caso}\\ 1 \text{ si } & i\in A2\cap E2\end{cases}$$

En la siguiente tabla detallamos las diferentes combinaciones de distribución del error y función de enlace que utilizamos con estos datos referentes a cuantías de siniestros:

Error:	**Enlace:**	**Canónico:**
Gaussiana, $V(\mu_i)=1$	Identidad, $g(\mu_i)=\mu_i$	Sí
Gaussiana, $V(\mu_i)=1$	Logarítmico, $g(\mu_i)=\log(\mu_i)$	No
Gamma, $V(\mu_i)=\mu_i^2$	Identidad, $g(\mu_i)=\mu_i$	No
Gamma, $V(\mu_i)=\mu_i^2$	Logarítmico, $g(\mu_i)=\log(\mu_i)$	Sí
Gaussiana inversa, $V(\mu_i)=\mu_i^3$	Identidad, $g(\mu_i)=\mu_i$	No
Gaussiana inversa, $V(\mu_i)=\mu_i^3$	Logarítmico, $g(\mu_i)=\log(\mu_i)$	No
Gaussiana inversa, $V(\mu_i)=\mu_i^3$	$g(\mu_i)=1/\mu_i^2$	Sí

Tabla 5.1. Combinaciones de error y enlace.

Una vez realizamos los procesos de selección con las diferentes combinaciones de error y de enlace de la tabla 5.1, obtenemos en todas ellas que el conjunto de predictores resultante es el mismo. Posteriormente, dados los predictores elegidos, estudiamos el poder predictivo de todas las combinaciones (Tabla 5.2), resultando "la mejor" la Gaussiana | Logarítmica. Por ello, a modo de ejemplo, detallamos a continuación los resultados con el citado modelo. En los anexos 5.1, 5.2 y 5.4 se encuentran con detalle los resultados del resto de modelos.

Modelo Gaussiana | Logarítmico

Detallamos por separado los procesos de introducción y de eliminación, puesto que el paso a paso, en este caso, se corresponde con ambos. Para el cálculo de los *p*-valores hacemos uso de la distribución asintótica *F* de Fisher, (3.41):

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.003307	**0.001841**	--------
A	**0.000001**	--------	--------
A:E	0.004719	0.689690	**0.615348**
	F(1) = A	F(2) = E	F(3) = A:E

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	0.001587	**0.001841**	--------
A	0.062327	0.000001	**0.000001**
A:E	**0.615348**	--------	--------
	F[1] = A:E	F[2] = E	F[3] = A

Si fijamos por ejemplo $\alpha^* = 0.05$ como nivel de significación permitido, nos quedaremos con los efectos principales de ambos predictores. Las estimaciones de los coeficientes resultantes son:

Coeficiente:	***t*-valor:**
$\hat{\beta}_0 = 5.6233$	77.8052
$\hat{\beta}_{A1} = -0.4589$	–5.1128
$\hat{\beta}_{A2} = -0.1707$	–2.2999
$\hat{\beta}_{E1} = 0.2208$	2.7257
$\hat{\beta}_{E2} = -0.0433$	–0.4999

Si nos fijamos en los *t*-valores, vemos como el de menor importancia es el de la clase 2 del estado.

Al utilizar el enlace logarítmico estamos ante un modelo de efecto multiplicativo. Su interpretación es la siguiente:

$$\log(\hat{\mu}_i) = \hat{\beta}_0 + \hat{\beta}_{A1} \cdot X_{A1} + \hat{\beta}_{A2} \cdot X_{A2} + \hat{\beta}_{E1} \cdot X_{E1} + \hat{\beta}_{E2} \cdot X_{E2}$$

$$\log(\hat{\mu}_i) = 5.6233 - 0.4589 \cdot X_{A1} - 0.1707 \cdot X_{A2} + 0.2208 \cdot X_{E1} - 0.0433 \cdot X_{E2}$$

$$\hat{\mu}_i = e^{(5.6233 - 0.4589 \cdot X_{A1} - 0.1707 \cdot X_{A2} + 0.2208 \cdot X_{E1} - 0.0433 \cdot X_{E2})}$$

$$\hat{\mu}_i = e^{5.6233} \cdot e^{-0.4589 \cdot X_{A1}} \cdot e^{-0.1707 \cdot X_{A2}} \cdot e^{0.2208 \cdot X_{E1}} \cdot e^{-0.0433 \cdot X_{E2}}$$

$$\hat{\mu}_i = 276.81 \times 0.6320^{X_{A1}} \times 0.8431^{X_{A2}} \times 1.2471^{X_{E1}} \times 0.9576^{X_{E2}}$$

Puesto que las variables son binarias, sólo cuando su valor es 1 debemos multiplicar al efecto base por el correspondiente parámetro:

	A1	A2	A3
E1	$276.81 \times 1.2471 \times 0.6320$	$276.81 \times 1.2471 \times 0.8431$	276.81×1.2471
E2	$276.81 \times 0.9576 \times 0.6320$	$276.81 \times 0.9576 \times 0.8431$	276.81×0.9576
E3	276.81×0.6320	276.81×0.8431	276.81

Así, las estimaciones de cada clase son:

	A1	A2	A3
E1	218.17	291.04	345.22
E2	167.52	223.47	265.07
E3	174.93	233.37	276.81

En el anexo 5.3 detallamos las estimaciones equivalentes pero en el enlace en lugar de la respuesta, tanto para este modelo como para el resto.

Ahora analizamos el "poder predictivo" de las combinaciones de la tabla 5.1 dados los predictores, antigüedad y estado. Utilizamos para ello el *p*-valor desprendido del estadístico de ajuste global (3.45) y adicionalmente notamos el *pseudo* R^2 (3.46):

Modelo	Poder predictivo:	Enlace Canónico:
Gaussiana, Identidad**	$R^2 = 0.0916$, p-valor = 0	Sí
Gaussiana, Logarítmico*	$R^2 = 0.0954$, **p-valor = 0**	**No**
Gamma, Identidad	$R^2 = 0.0697$, p-valor = 0.000009	No
Gamma, Logarítmico	$R^2 = 0.0717$, p-valor = 0.000549	Sí
Gaussiana inversa, Identidad	$R^2 = 0.0338$, p-valor = 0.008463	No
Gaussiana inversa, Logarítmico	$R^2 = 0.0075$, p-valor = 0.545475	No
Gaussiana inversa, $g(\mu_i) = 1/\mu_i^2$	$R^2 = 0.0080$, p-valor = 0.508749	Sí

Tabla 5.2. Poder predictivo MLG.

* $F_{4,396} = 10.4415$, ** $F_{4,396} = 9.9838$

La combinación de error y enlace que nos ofrece un mejor resultado es la Gaussiana | Logarítmica.

Podemos realizar los siguientes comentarios:

Estamos en el caso de cuantías por siniestro, por lo que en general es recomendada la utilización de distribuciones con el rango de **Y** positivo y con asimetría positiva, como es el caso de la Gamma y de la Gaussiana inversa, a ser posible con un efecto multiplicativo, enlace logarítmico, que ayuda a no obtener estimaciones negativas para éstas y para cualquier otra distribución. En este caso la Gaussiana no nos ha proporcionado estimaciones negativas y ha sido la de mayor poder predictivo, por lo que en general no nos limitaremos a aceptar la utilización de la Gamma o de la Gaussiana inversa.

Comprobamos como para una distribución del error y unas variables dadas, no siempre el enlace canónico es el que mejor ajusta los datos, por ejemplo para la Gaussiana el mejor es el Logarítmico, y para la Gamma y la Gaussiana inversa el mejor es la Identidad, y en ningún caso es el canónico. Estos enlaces proporcionan propiedades simplificadoras en la formulación genérica de la familia exponencial, cosa que no implica que sean la combinación más adecuada para unos datos determinados.

Respecto a todos los modelos con el enlace logarítmico (de efecto multiplicactivo), en referencia a la interacción A:E, observamos como, fijada una distribución, el enlace logarítmico es el que más rechaza a la interacción [Brockman y Wright (1992)], al menos empíricamente. Para comprobarlo nos podemos fijar en la primera columna de los procesos tanto de introducción como de eliminación de los modelos analizados (anexo 5.1). Por ejemplo, para la distribución Gamma y de introducción, para el enlace logarítmico el *p*-valor es 0.017475, y para el enlace identidad 0.009194; para el de eliminación, para el logarítmico es 0.831886 y para el identidad 0.678287.

5.1.4. Regresión basada en distancias

En este apartado, a parte de seleccionar los predictores que mejor explican las cuantías de los impagos y comprobar empíricamente el buen funcionamiento de la metodología de selección propuesta en el capítulo 4 haciendo uso de *bootstrap*, ilustramos como con la RBD, a diferencia del MLG, cuando se trata con predictores cualitativos las posibilidades de tratamiento son amplias.

Tratamiento de predictores

Con el MLG, tal y como hemos visto en el apartado 3.5.2.3, debemos codificar variables binarias para las clases y no importa si se consideran binarias disjuntas o con códigos ordinales, la predicción es la misma. Sin embargo, con la RBD no ocurre así. Para predictores tanto nominales como ordinales se puede utilizar o bien una variable categórica o bien el conjunto de binarias disjuntas asociado al conjunto de clases. Y para predictores ordinales se puede utilizar, adicionalmente, el conjunto de binarias ordinales con tal de incluir la ordinalidad de las clases. La predicción no tiene porqué coincidir en ninguno de los casos. Así,

a) Si utilizamos el índice de similaridad de Gower, (4.8), y disponemos de p_2 variables binarias y p_3 variables categóricas únicamente, tenemos que $s_{ij} = \frac{\alpha_{ij} + a}{p_2 + p_3 - d}$. En el caso de disponer de sólo p_3 variables binarias, tenemos el coeficiente de Jaccard (tabla 4.2), $s_{ij} = \frac{a}{p_3 - d}$, y en el caso de sólo p_2 variables categóricas, tenemos el coeficiente de coincidencias (4.7), $s_{ij} = \frac{\alpha_{ij}}{p_2}$.

Si tratamos al estado y a la antigüedad como categóricas nominales, codificando una sola variable por predictor y regresamos, obtenemos la misma estimación que en el MLG Gaussiana | Identidad (anexo 5.4). Es decir, la misma estimación que con regresión clásica codificando a los predictores con tantas binarias como clases. Para cualquier otra combinación de tratamiento para la antigüedad y el estado obtenemos, con esta distancia, las medias de la tabla 2.4.

b) Puesto que con la distancia de Gower, no obtenemos una estimación diferente para cada combinación de tratamiento de predictores, vamos a construir una función de distancias que sí los proporcione:

Por ejemplo, utilizamos S2 (tabla 4.2), $s_{ij} = \frac{a}{p_3}$, para las variables p_3 binarias, conjuntamente con el coeficiente de coincidencias (4.7), $s_{ij} = \frac{\alpha_{ij}}{p_2}$, para las p_2 variables categóricas, y construimos la similaridad:

$$s_{ij} = \frac{\alpha_{ij} + a}{p_2 + p_3} \tag{5.1}$$

que por ser suma de euclídeas también lo es.

Con esta nueva distancia, dependiendo del tipo de información introducida, obtenemos diferentes predicciones. Si codificamos:

1. Dos variables categóricas, una para el estado y una para la antigüedad
2. Para el estado las 3 binarias disjuntas (3.47), y para la antigüedad las 3 binarias disjuntas (3.48)
3. Para el estado las 3 binarias disjuntas (3.47), y para la antigüedad las 2 binarias con codificación ordinal (3.49)
4. Para el estado 1 variable categórica y para la antigüedad las 2 binarias con codificación ordinal (3.49)
5. Para el estado 1 variable categórica y para la antigüedad las 3 binarias disjuntas (3.47)

Obtenemos las siguientes estimaciones:

	A1	A2	A3
E1	1. 227.09	1. 289.65	1. 332.61
	2. 227.52	2. 288.26	2. 330.03
	3. 234.98	3. 295.37	3. 337.73
	4. 263.89	4. 325.62	4. 366.61
	5. 227.31	5. 288.94	5. 331.30
E2	1. 163.84	1. 226.40	1. 269.36
	2. 166.01	2. 226.75	2. 268.52
	3. 163.22	3. 223.61	3. 265.97
	4. 150.50	4. 212.22	4. 253.21
	5. 164.94	5. 226.58	5. 268.93
E3	1. 173.64	1. 236.21	1. 279.16
	2. 175.73	2. 236.46	2. 278.23
	3. 172.81	3. 233.20	3. 275.56
	4. 158.86	4. 220.58	4. 261.57
	5. 174.70	5. 236.33	5. 278.69

Observamos un resultado diferente para cada una de las cinco combinaciones de tratamiento de los predictores implicados.

Proceso de selección paso a paso

Utilizamos el coeficiente de coincidencias, $s_{ij} = \frac{\alpha_{ij}}{p_2}$, en el caso de disponer de p_2 variables categóricas. Utilizamos esta similaridad, pues sabemos que el resultado obtenido en el proceso de selección debe coincidir con el resultado obtenido con el modelo clásico tal y como se detalla en el apartado 4.2.3 de casos particulares de la RBD (caso 3). Notamos que el caso particular se refiere tan sólo a los valores de probabilidad que no incluyen al término de interacción.

Realizamos el proceso de dos maneras:

a) Haciendo uso de la fórmula (4.32) para la inclusión de la interacción A:E.

b) Incluyendo una nueva variable categórica, a modo de predictor, con el código de la tabla cruzada en representación del término interacción, A*E.

Para la estimación de los *p*-valores de los contrastes, se han generado dos bloques de muestras, cada una con 500 muestras: $B = B1 + B2 = 500 + 500 = 1\,000$ simulaciones. En el anexo 5.5 se detallan los

p-valores por separado, para *B1* y para *B2*, para observar la convergencia en la estimación, la cual consideramos suficiente (aproximadamente en dos decimales). Observaremos en lo que sigue que los *p*-valores obtenidos son intuitivamente elevados, pero por ello no descartaremos variables. Las distribuciones exactas estimadas con la simulación son algo más uniformes que las distribuciones asintóticas, por lo que las colas son más amplias. Principalmente nos fijaremos en el patrón cuantitativo del conjunto de valores de probabilidad.

a) Haciendo uso de la fórmula (4.32) para la interacción A:E

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.246	**0.258**	--------
A	**0.233**	--------	--------
A:E	0.424	0.433	**0.456**
	F(1) = A	F(2) = E	F(3) = A:E

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	0.276	**0.258**	--------
A	0.251	0.244	**0.233**
A:E	**0.456**	--------	--------
	F[1] = A:E	F[2] = E	F[3] = A

El resultado obtenido ha sido el esperado: F(1) = A, F(2) = E y F(3) = A:E.

b) Incorporando a la interacción como un nuevo predictor categórico A*E

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.246	**0.258**	--------
A	**0.232**	--------	--------
A*E	0.387	0.414	**0.448**
	F(1) = A	F(2) = E	F(3) = A*E

Fases de eliminación:

Variable	*p*-valores
F(1)	**0.233**
F(2) | F(1)	**0.258**
F(1) | F(2)	0.244
F(3) | F(1)F(2)	0.448
F(1) | F(2)F(3)	0.364
F(2) | F(1)F(3)	**0.526**

Observamos que F[1] = E, diferente de F(3) = A*E. Al estar en el último paso, aprovechamos para realizar el proceso de eliminación progresiva completo:

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	**0.526**	--------	--------
A	0.364	0.003	**0.233**
A*E	0.448	**0.414**	--------
	F[1] = E	F[2] = A*E	F[3] = A

Se obtiene un resultado coherente: en el proceso de introducción progresiva, al igual que en caso anterior, tenemos que F(1) = A, F(2) = E y F(3) = A*E. Si consideramos las fases de eliminación para construir el paso a paso, al incluir A*E en el paso 3 del proceso, resulta que el E es la primera

variable a ser eliminada, su efecto queda incluido en la interacción A*E de una manera más significativa. Sin embargo observamos como en el paso 2 del proceso de eliminación, la A resulta mejor predictor que la interacción, y con diferencia.

Observaremos como los resultados son diferentes debido a que la inclusión de una nueva variable categórica no se corresponde exactamente con la idea de interacción clásica entre predictores para la RBD.

5.1.5. Estimación no científica de las marcas de clase

Cuando en el conjunto de predictores disponemos de variables continuas discretizadas en intervalos, y no disponemos de los datos originales cuantitativos, los datos se denominan *datos censurados*. En tal situación, nos es imposible realizar una estimación científica de las marcas de clase correspondientes a cada intervalo. Si los intervalos no son muy amplios, una posibilidad puede ser escoger los puntos medios, o valores intermedios que nos parezcan lógicos, pero tendremos problemas en los intervalos extremos no acotados (o libres en alguno de sus extremos) si no disponemos de información adicional.

Lo usual por lógica, sería discretizar en intervalos que nos resultaran homogéneos respecto a la estructura de la respuesta, lo más finos posible y teniendo en cuenta que en cada intervalo nos cayera un número mínimo de individuos. Con esta lógica, las marcas de clase casi nunca coincidirían con los puntos medios de los intervalos. La mayoría de programas informáticos cuando se trata de discretizar variables continuas, crean intervalos o de igual amplitud o con el mismo número de casos en cada nueva clase, criterios que tampoco permiten estimar correctamente puntuaciones para las marcas de clase *a posteriori*. Para deshacer la reducción de escala, se hace imprescindible disponer de los datos originales.

Cuando se lleva una dinámica, en la que inicialmente se discretiza haciendo uso de algún criterio, o simplemente porque aquellas clases son "las que se suelen contemplar" sin saber muy bien por qué, los datos se codifican en los ficheros por la simple pertenencia a los intervalos. Para tratar cuantitativamente a tales discretizaciones, necesitamos información externa o la opinión de expertos que nos oriente en la asignación de puntuaciones.

Concretamente, en el seguro del automóvil, es muy difícil extrapolar marcas de clase de algún tipo de estadísticas generales, y adicionalmente las carteras variarán de una compañía a otra, pues estarán compuestas por unas tipologías de pólizas diferentes aunque cubran el mismo riesgo. Pongamos el ejemplo de la edad del conductor principal: si tenemos que el último intervalo de edad es mayor de 65 años, podría ocurrir que la marca de clase fuese 66, que fuese 70 o que fuese 75. Si pretendemos estudiar con detalle las edades iniciales, y en los datos tenemos que el primer intervalo es menor a 25 años, del mismo modo, la marca podría coincidir aproximadamente con el punto medio, 21 años, estar por debajo, 19 años, o estar por encima en una media de 24 años.

Si lo que se pretende es tratar al predictor inicialmente de naturaleza cuantitativa como tal, los métodos de regresión no serán robustos respecto a las diferentes puntuaciones. Vamos a ilustrar tal efecto con los datos de esta aplicación, haciendo uso tanto del MLG como del MBD. Concretamente daremos tratamiento de continua a la antigüedad en el puesto de laboral. Recordemos que los intervalos venían definidos por los puntos de corte 2 y 10. El primer intervalo era de 0 a 2, el segundo de 2 a 10 y el tercero más de 10 años. Supongamos que un individuo puede empezar a trabajar a los 18 y que se jubilará a los 65, lo que implica que como cota superior para la antigüedad laboral tendremos 47 años. No es imposible que el impago se produzca a una edad avanzada, ya que no disponemos de información sobre las características del préstamo (ni inicio, ni temporalidad e incluso cuantía del capital a amortizar ni del amortizado hasta el instante del impago).

Si escogiésemos los puntos medios, las puntuaciones serían 1 para el primero, 6 para el segundo y 28 para el tercero. Pero estos intervalos no tienen por qué tener estas marcas de clase, así pues para el estudio comparativo de resultados, analizaremos 5 combinaciones diferentes: a) 1-6-28, b) 1-6-44, c) 1-6-11, d) 0.5-3-11 y e) 1.5-9-44. Utilizaremos el MLG Gaussiana | Logarítmico, tratando al estado con las tres binarias disjuntas (3.47) y a la antigüedad como continua, y el MBD utilizando la distancia de Gower, tratando al estado como categórico nominal y a la antigüedad como continua.

Realizamos la estimación para cada combinación de la a) a la e), para cada clase del estado y utilizando como puntuación la marca de clase de la antigüedad, tal y como se presenta en las dos tablas siguientes:

MLG	Antigüedad = 1.5	Antigüedad = 6	Antigüedad = 9	Antigüedad = 20	Antigüedad = 30
E1	a) 246.34 b) 249.73 c) 229.97 d) 250.19 e) 245.22	a) 261.44 b) 258.66 c) 280.45 d) 293.54 e) 254.55	a) 272.02 b) 264.78 c) 320.12 d) 326.52 e) 260.97	a) 314.59 b) 288.50 c) 519.99 d) 482.51 e) 285.92	a) 359.06 b) 311.91 c) 808.19 d) 688.15 e) 310.66
E2	a) 188.73 b) 191.25 c) 176.51 d) 191.74 e) 187.85	a) 200.30 b) 198.09 c) 215.25 d) 224.95 e) 195.00	a) 208.40 b) 202.78 c) 245.70 d) 250.23 e) 199.92	a) 241.02 b) 220.94 c) 399.09 d) 369.78 e) 219.03	a) 275.09 b) 238.87 c) 620.29 d) 527.37 e) 237.98
E3	a) 197.37 b) 200.16 c) 184.32 d) 200.47 e) 196.48	a) 209.48 b) 207.27 c) 224.78 d) 235.19 e) 203.96	a) 217.95 b) 212.17 c) 256.57 d) 261.62 e) 209.10	a) 252.07 b) 231.18 c) 416.76 d) 386.60 e) 229.09	a) 287.69 b) 249.93 c) 647.75 d) 551.37 e) 248.91

MBD	Antigüedad = 1.5	Antigüedad = 6	Antigüedad = 9	Antigüedad = 20	Antigüedad = 30
E1	a) 233.34 b) 233.34 c) 233.34 d) 252.11 e) 227.09	a) 289.65 b) 289.65 c) 289.65 d) 305.76 e) 264.63	a) 295.51 b) 293.04 c) 315.43 d) 321.87 e) 289.65	a) 316.99 b) 305.48 c) 332.61 d) 332.61 e) 303.15	a) 332.61 b) 316.78 c) 332.61 d) 332.61 e) 315.43
E2	a) 170.10 b) 170.10 c) 170.10 d) 188.87 e) 163.84	a) 226.40 b) 226.40 c) 226.40 d) 242.51 e) 201.38	a) 232.26 b) 229.80 c) 252.18 d) 258.62 e) 226.40	a) 253.74 b) 242.23 c) 269.36 d) 269.36 e) 239.90	a) 269.36 b) 253.53 c) 269.36 d) 269.36 e) 252.18
E3	a) 179.90 b) 179.90 c) 179.90 d) 198.67 e) 173.64	a) 236.21 b) 236.21 c) 236.21 d) 252.31 e) 211.18	a) 242.06 b) 239.60 c) 261.98 d) 268.42 e) 236.21	a) 263.54 b) 252.03 c) 279.16 d) 279.16 e) 249.71	a) 279.16 b) 263.34 c) 279.16 d) 279.16 e) 261.98

Observamos como para una misma combinación, según las puntuaciones de antigüedad empleadas en el modelo, las estimaciones son diferentes, tanto con el MLG como con el MBD. No vemos ni una sola celda en la que los valores de la a) a la e) sean iguales. Sin embargo, las estimaciones con el MBD son más estables o parecidas en cada celda, en consecuencia un modelo más robusto respecto a cambios en los datos.

Algunas puntuaciones tienen mayor asociación lineal con la respuesta,

	ρ
▪ Cuantías y Antigüedad 1-6-28:	0.233
▪ Cuantías y Antigüedad 1-6-44:	0.225
▪ Cuantías y Antigüedad 1-6-11:	0.256
▪ Cuantías y Antigüedad 0.5-3-11:	0.238
▪ Cuantías y Antigüedad 1.5-9-44:	0.232

aunque este hecho no implica que sean las más reales. Cabe notar, que si las puntuaciones empleadas no se aproximan a la realidad, las significaciones obtenidas en los procesos de selección no serán correctas.

ANEXO 5.1. Procesos de selección para los MLG

***Gaussiana | Identidad*:**

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.003307	**0.004220**	--------
A	**0.000001**	--------	--------
A:E	0.004719	0.689690	**0.365444**
	F(1) = A	F(2) = E	F(3) = A:E

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	0.001587	**0.004220**	--------
A	0.062327	0.000002	**0.000001**
A:E	**0.365444**	--------	--------
	F[1] = A:E	F[2] = E	F[3] = A

***Gamma | Identidad*:**

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.014157	**0.041501**	--------
A	**0.000014**	--------	--------
A:E	0.009194	0.685425	**0.678287**
	F(1) = A	F(2) = E	F(3) = A:E

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	0.041787	**0.041501**	--------
A	0.077710	0.000042	**0.000014**
A:E	**0.678287**	--------	--------
	F[1] = A:E	F[2] = E	F[3] = A

***Gamma | Logarítmico*:**

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.014157	**0.027136**	--------
A	**0.000014**	--------	--------
A:E	0.017475	0.685425	**0.831886**
	F(1) = A	F(2) = E	F(3) = A:E

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	0.041787	**0.027136**	--------
A	0.077710	0.000027	**0.000014**
A:E	**0.831886**	--------	--------
	F[1] = A:E	F[2] = E	F[3] = A

***Gaussiana inversa | Identidad*:**

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.137995	**0.254503**	--------
A	**0.004203**	--------	--------
A:E	0.140639	0.871811	**0.927537**
	F(1) = A	F(2) = E	F(3) = A:E

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	0.307652	0.254503	--------
A	0.272684	0.007888	0.004203
A:E	**0.927537**	--------	--------
	F[1] = A:E	F[2] = E	F[3] = A

***Gaussiana inversa | Logarítmico*:**

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.137995	**0.216624**	--------
A	**0.004203**	--------	--------
A:E	0.140639	0.871811	**0.967481**
	F(1) = A	F(2) = E	F(3) = A:E

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	0.307652	**0.216624**	--------
A	0.272684	0.006714	**0.004203**
A:E	**0.967481**	--------	--------
	F[1] = A:E	F[2] = E	F[3] = A

Gaussiana inversa | $g(\mu_i) = 1/\mu_i^2$:

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.137995	**0.193447**	--------
A	**0.004203**	--------	--------
A:E	0.140639	0.871811	**0.307652**
	F(1) = A	F(2) = E	F(3) = A:E

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	0.307652	**0.193447**	--------
A	0.272684	0.005996	**0.004203**
A:E	**0.987468**	--------	--------
	F[1] = A:E	F[2] = E	F[3] = A

ANEXO 5.2. Coeficientes para los MLG

***Gaussiana | Identidad*:**

Coeficiente:	*t*-valor:
$\hat{\beta}_0 = 279.1614$	14.8076
$\hat{\beta}_{A1} = -105.5201$	–5.2230
$\hat{\beta}_{A2} = -42.9560$	–2.1343
$\hat{\beta}_{E1} = 53.4468$	2.5451
$\hat{\beta}_{E2} = -9.8011$	–0.4940

***Gamma | Identidad*:**

Coeficiente:	*t*-valor:
$\hat{\beta}_0 = 277.7199$	13.5651
$\hat{\beta}_{A1} = -100.0557$	–4.9144
$\hat{\beta}_{A2} = -37.8243$	–1.6802
$\hat{\beta}_{E1} = 45.1015$	2.0879
$\hat{\beta}_{E2} = -9.5737$	–0.5361

***Gamma | Logarítmico*:**

Coeficiente:	*t*-valor:
$\hat{\beta}_0 = 5.6203$	71.5563
$\hat{\beta}_{A1} = -0.4402$	–5.2304
$\hat{\beta}_{A2} = -0.1471$	–1.7555
$\hat{\beta}_{E1} = 0.1960$	2.2406
$\hat{\beta}_{E2} = -0.0455$	–0.5515

***Gaussiana inversa* | *Identidad*:**

Coeficiente:	*t*-valor:
$\hat{\beta}_0 = 227.0101$	12.5434
$\hat{\beta}_{A1} = -98.2996$	–4.5408
$\hat{\beta}_{A2} = -36.0711$	–1.4683
$\hat{\beta}_{E1} = 41.6197$	1.8798
$\hat{\beta}_{E2} = -9.4198$	–0.5479

***Gaussiana inversa* | *Logarítmico*:**

Coeficiente:	*t*-valor:
$\hat{\beta}_0 = 5.6178$	66.4791
$\hat{\beta}_{A1} = -0.4328$	–5.0037
$\hat{\beta}_{A2} = -0.1379$	–1.5033
$\hat{\beta}_{E1} = 0.1854$	2.0080
$\hat{\beta}_{E2} = -0.0463$	–0.5658

***Gaussiana inversa* | $g(\mu_i) = 1/\mu_i^2$:**

Coeficiente:	*t*-valor:
$\hat{\beta}_0 = 1.3656 \times 10^{-5}$	5.4183
$\hat{\beta}_0 = 1.6724 \times 10^{-5}$	4.8379
$\hat{\beta}_0 = 4.1271 \times 10^{-6}$	1.6564
$\hat{\beta}_0 = -5.7742 \times 10^{-6}$	–2.0858
$\hat{\beta}_0 = 1.7152 \times 10^{-6}$	0.5464

ANEXO 5.3. Predicciones del enlace para los MLG

***Gaussiana | Logarítmico*:**

	A1	A2	A3
E1	5.39	5.67	5.84
E2	5.12	5.41	5.58
E3	5.16	5.45	5.62

***Gamma | Logarítmico*:**

	A1	A2	A3
E1	5.38	5.67	5.82
E2	5.13	5.43	5.57
E3	5.18	5.47	5.62

***Gaussiana inversa | Logarítmico*:**

	A1	A2	A3
E1	5.37	5.67	5.80
E2	5.14	5.43	5.57
E3	5.18	5.48	5.62

***Gaussiana inversa* | $g(\mu_i) = 1/\mu_i^2$:**

	A1	A2	A3
E1	2.4606×10^{-5}	1.2009×10^{-5}	7.8821×10^{-6}
E2	3.2095×10^{-5}	1.9498×10^{-5}	1.5371×10^{-5}
E3	3.0380×10^{-5}	1.7783×10^{-5}	1.3656×10^{-5}

ANEXO 5.4. Predicciones de la respuesta para los MLG

***Gaussiana | Identidad*:**

	A1	A2	A3
E1	227.09	289.65	332.61
E2	163.84	226.40	269.36
E3	173.64	236.21	279.16

***Gamma | Identidad*:**

	A1	A2	A3
E1	222.77	285.00	322.82
E2	168.09	230.32	268.15
E3	177.66	239.90	277.72

***Gamma | Logarítmico*:**

	A1	A2	A3
E1	216.17	289.78	335.74
E2	169.78	227.58	263.68
E3	177.69	238.20	275.97

***Gaussiana inversa | Identidad*:**

	A1	A2	A3
E1	220.33	282.56	318.63
E2	169.29	231.52	267.59
E3	178.71	240.94	277.01

***Gaussiana inversa | Logarítmico*:**

	A1	A2	A3
E1	214.96	288.68	331.39
E2	170.48	228.94	262.81
E3	178.57	239.81	275.29

***Gaussiana inversa* | $g(\mu_i) = 1/\mu_i^2$:**

	A1	A2	A3
E1	201.59	288.56	356.19
E2	176.51	226.46	255.06
E3	181.43	237.13	270.60

ANEXO 5.5. Procesos de selección para la RBD

Utilizando la fórmula (4.32) para el tratamiento de la interacción A:E, para *B1* y *B2* por separado:

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.246 y 0.246	**0.258 y 0.258**	--------
A	**0.220 y 0.245**	--------	--------
A:E	0.430 y 0.418	0.418 y 0.448	**0.454 y 0.458**
	F(1) = A	F(2) = E	F(3) = A:E

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	0.286 y 0.266	**0.258 y 0.258**	--------
A	0.242 y 0.260	0.230 y 0.258	**0.220 y 0.245**
A:E	**0.454 y 0.458**	--------	--------
	F[1] = A:E	F[2] = E	F[3] = A

Utilizando una variable categórica adicional en representación de la interacción A*E, para *B1* y *B2* por separado:

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
E	0.246 y 0.246	**0.258 y 0.258**	--------
A	**0.220 y 0.245**	--------	--------
A*E	0.384 y 0.390	0.408 y 0.420	**0.438 y 0.458**
	F(1) = A	F(2) = E	F(3) = A*E

Fases de eliminación:

Variable	*p*-valores
F(1)	**0.220 y 0.245**
F(2) | F(1)	**0.258 y 0.258**
F(1) | F(2)	0.230 y 0.258
F(3) | F(1)F(2)	0.438 y 0.458
F(1) | F(2)F(3)	0.372 y 0.356
F(2) | F(1)F(3)	**0.522 y 0.530**

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
E	0.522 y 0.530	--------	--------
A	0.372 y 0.356	0.006 y 0	**0.220 y 0.245**
A*E	**0.438 y 0.458**	**0.408 y 0.420**	--------
	F[1] = E	F[2] = A*E	F[3] = A

5.2. Aplicación 2. Cartera C1 de responsabilidad civil de automóviles: cuantía de un siniestro para daños personales

Utilizamos los datos descritos en el apartado 2.3.1, que hacen referencia a la cartera C1 de responsabilidad civil del automóvil. Analizamos la cuantía por siniestro para daños personales a partir de información desagregada. Los factores de riesgo son de tipo mixto.

De las 169 618 pólizas de la cartera hemos seleccionado del siguiente modo, con tal de obtener una porción de datos estándares con los que ilustrar las metodologías expuestas en el trabajo:

- Las que estaban en activo a 1 de enero de 1997 mediante la variable estado=1=Activa, obteniendo 111 231 pólizas
- De ellas seleccionamos las de clase turismos, obteniendo 90 179 pólizas
- De ellas seleccionamos las que no tienen franquicia de ningún tipo, obteniendo 81 106 pólizas
- De ellas seleccionamos las de grupo de vehículo de primera categoría, siendo igualmente 81 106
- De ellas seleccionamos las de expuesto igual a 1, mirando la fecha de inicio de la póliza y la fecha de anulación, así nos quedamos con las que la fecha de inicio efecto de la póliza es 31/12/95 o anterior y que la fecha de anulación es 1/1/97 o posterior, obteniendo 60 575 pólizas
- Escogemos las de tipo de producto estándar, obteniendo 44 832 pólizas.

En las 44 832 pólizas tenemos 451 con un siniestro y 3 con dos siniestros, por lo que en total analizaremos 451 + 3 + 3 = 457 cuantías. Observando el rango de las cuantías nos damos cuenta de que disponemos de 2 puntos extremos, uno de 60 241 698 pesetas y otro de 1 peseta. El rango de variación del resto está entre 13 080 y 29 324 960 pesetas. Decidimos eliminar ambas cuantías del estudio, así que finalmente analizamos 457 – 2 = 455 cuantías de siniestros para daños personales.

5.2.1. Relaciones entre pares de variables

Estudiamos en dos apartados diferentes las asociaciones entre las cuantías y cada factor individual, y las asociaciones entre factores dos a dos. Distinguimos el tratamiento cuantitativo o cualitativo de los factores.

5.2.1.1. Cuantías con factores

Asociación de las cuantías con cada factor cuantitativo

En esta aplicación damos tratamiento cuantitativo a los predictores de naturaleza cuantitativa. Estos también podrían incluirse en la regresión con tratamiento cualitativo pasando previamente por un proceso de discretización, el cual hemos descrito detalladamente en el apartado 3.6.2 del trabajo.

Calculamos (3.2),

Variables	ρ
Y – Potencia	0.0103
Y – Antivehi	0.0286
Y – Valorvehi	–0.0048
Y – Edad	0.0159
Y – Anticarn	0.0106
Y – Bonus-malus	–0.0059
Y – Bonus	–0.0014
Y – Malus	0.0174

Las correlaciones no son muy elevadas. La mayor es para la antigüedad del vehículo, que nos indica que a mayor antigüedad del vehículo mayor cuantía en el siniestro, seguida de la variable malus, que más abajo comentaremos.

La correlación entre las cuantías y la potencia sale positiva, lo que nos indica que en términos generales, ya que es una correlación demasiado pequeña para afirmar, tendremos que a mayor potencia mayor cuantía del siniestro. Del mismo modo, a menor valor del vehículo mayor cuantía.

Observamos que las correlaciones tanto de la edad como de la antigüedad del carnet (en principio del primer conductor) con las cuantías salen pequeñas y positivas. Cabría esperar que el signo fuera negativo, ya que "a mayor experiencia" las cuantías deberían ser menores. Sin embargo, si analizamos más afondo esta cartera, observamos que las formas de pago fraccionadas están asociadas a cuantías menores, y a su vez también se corresponden con edades menores, de lo que interpretamos que los jóvenes están asociados a cuantías menores en general.

Recordemos que en esta aplicación la variable bonus contiene cero para los individuos que están en la escala de malus, y la variable malus toma valores positivos que son mayores a mayor escala de malus y ceros para las pólizas de escala bonus. La variable bonus-malus la construimos como bonus menos malus, lo que nos proporciona la escala completa. Si observamos las correlaciones tenemos que: la variable bonus por si sola tiene una correlación negativa de 0.0014 con las cuantías, la variable malus tiene una correlación positiva con las cuantías de 0.0174 (ya que la escala la tenemos en signo contrario), y la variable bonus-malus una negativa de 0.0059. Observamos que la mayor correlación es para la variable malus por si sola, seguida de la variable bonus-malus y finalmente de la variable bonus. La variable bonus es la menos correlacionada con las cuantías, ya que estamos estudiando las pólizas que han sufrido al menos un siniestro en el período. En general, dichas pólizas estarán en la escala de malus, por lo que decidimos incluir exclusivamente como factor potencial al malus con signo positivo.

Respecto a la relación de la variable Bonus con el resto de factores sólo destacar su correlación con la antigüedad del carnet:

Bonus – Anticarn: $\rho = 0.114$

Es lógica una correlación positiva con la antigüedad del carnet del primer conductor, pues la variable bonus es una variable histórica acumulativa al igual que la antigüedad del carnet, y se espera que a mayor antigüedad mejor nivel.

Asociación de las cuantías con cada factor cualitativo

Estudiamos la asociación de las cuantías con cada predictor cualitativo haciendo uso de (3.4), y obtenemos:

Variables	η
Y – Sexo	0.0361
Y – Zona	0.0980
Y – Vehicu	0.0332
Y – Uso	0.0436
Y – Pago	0.0321

La más asociada con las cuantías es la zona. Los datos analizados son de toda España, por lo que es lógico que el importe de las cuantías es diferente dependiendo de la zona en que nos encontremos. También cabe notar que la zona es la variable cualitativa con más categorías con diferencia, 10 versus 4 como mucho, y por lo tanto es la variable categórica que más hace diferenciar las medias de las cuantías, las cuales toman un rango bastante amplio. Si en lugar de la medida de asociación nos fijamos en los *p*-valores resultantes de las anovas que encontramos en el anexo 5.6, obtenemos el siguiente resultado:

Variables	*p*-valor
Y – Sexo	0.441
Y – Zona	0.888
Y – Vehicu	0.923
Y – Uso	0.355
Y – Pago	0.792

Aunque la zona era la de mayor asociación según (3.4), observamos como su *p*-valor del anova no es el menor. Los dos menores son para las variables Uso y Sexo.

Ninguna de las variables tiene un *p*-valor pequeño, pero de los gráficos de medias del anexo 5.6, podemos realizar los siguientes comentarios:

- Sexo: La cuantía media de los hombres es menor que la de las mujeres, y aunque no se considera una diferencia significativa, junto con la variable Uso, el Sexo es la segunda de mayor significación.
- Zona: Las cuantías medias más elevadas son para las zonas 4 y 9, y las menores para las zonas 1, 2, 6 y 10, siendo mínimas las diferencias.
- Tipo de vehículo: Observamos que las clases están muy desproporcionadas en el número de siniestros, pero hemos decidido no agrupar las clases, pues aunque no hay diferencias significativas en las medias, se observa que para los 2 siniestros de los balilla duros, las cuantías son más bajas que para el resto.
- Uso del vehículo: La media de cuantía para el uso particular es menor que para el uso de empresa. Aunque las diferencias no llegan a ser significativas, ésta es la variable, junto con el sexo, que ofrece mayor diferencia.
- Forma de pago: Los pagos fraccionados se asocian a cuantías menores que los anuales, a diferencia de lo que ocurre en la aplicación 4 que hace referencia al número de siniestros para materiales. El

factor Pago puede ser entendido como una variable que resume otras características implícitas del asegurado. Si se decide incluir como posible variable de tarifa en una tarificación a priori, la forma de pago ésta es una variable que es elegida voluntariamente por el tomador, por lo que lo lógico financieramente sería incluir un recargo por el retardo de los pagos de las primas.

5.2.1.2. Factor con factor

Las relaciones entre factores que presentamos a continuación son las relaciones entre los factores de las pólizas que han sufrido siniestro, lo que no significa que sean las relaciones entre factores genéricas de todas las pólizas de la cartera.

Cualitativo con cualitativo

Calculamos (3.6), (3.7), (3.8) y la correlación canónica r_1,

		Sexo	Zona	Vehicu	Uso	Pago
Sexo	C =	1	**0.1615**	0.1164	0.0066	0.0831
	T =	1	0.0933	0.0884	0.0066	0.0693
	P =	1	0.1594	0.1156	0.0162	0.0825
	r_1 =	1	**0.1614**	0.1164	0.0105	0.0828
Zona	C =		1	**0.1533**	0.1166	**0.1664**
	T =		1	0.1054	0.0671	0.1140
	P =		1	0.2119	0.1160	0.2289
	r_1 =		**1**	**0.5383**	0.1168	**0.1714**
Vehicu	C =			**1**	0.1212	**0.0147**
	T =			1	0.1095	0.0120
	P =			1	0.1691	0.0286
	r_1 =			1	0.0301	**0.1631**
Uso	C =				1	0.0742
	T =				1	0.0624
	P =				1	0.0735
	r_1 =				1	0.0548
Pago	C =					1
	T =					**1**
	P =					**1**
	r_1 =					1

No resaltamos ninguna de las asociaciones en especial, tan solo comentamos las más relevantes comparativamente: La zona está asociada mínimamente con el sexo, con el tipo de vehículo y con la forma de pago, aunque remitimos de nuevo a que la zona es la variable con más categorías en comparación con el resto de cualitativas. Finalmente, la forma de pago tiene frecuencias diferentes según el tipo de vehículo, sin olvidar la frecuencia de dos del tipo de vehículo balilla duro.

Cuantitativo con cuantitativo

Calculamos (3.2),

$\rho =$	Potencia	Antivehi	Valorvehi	Edad	Anticarn	Malus
Potencia	1	**– 0.2445**	**0.8374**	0.0161	0.0722	0.0660
Antivehi		1	**– 0.3225**	**0.1722**	**0.0986**	0.0300
Valorvehi			1	0.0041	0.0742	**0.0971**
Edad				1	**0.7446**	0.0151
Anticarn					1	0.0391
Malus						1

Resaltamos las siguientes correlaciones :

- Entre la potencia y el valor del vehículo, a mayor potencia mayor valor del vehículo
- Entre la antigüedad del carnet y la edad del conductor principal, a mayor antigüedad mayor edad
- Entre la antigüedad y el valor del vehículo, a mayor antigüedad menor valor del vehículo
- Entre la antigüedad del vehículo y la potencia, a mayor antigüedad menor potencia
- Entre la edad del conductor y la antigüedad del vehículo, a mayor edad mayor antigüedad

Y secundariamente :

- A mayor antigüedad del carnet mayor antigüedad del vehículo
- A mayor valor del vehículo peor escala de malus

Las dos correlaciones más altas nos pueden llevar a decidir si en un proceso de selección, utilizamos la potencia y descartamos el valor del vehículo (o al revés), y si utilizamos la antigüedad del carnet o alternativamente la edad del conductor principal. Aunque al no ser, en ninguno de los dos casos, una

asociación máxima de 1, podemos optar por introducir ambas para ver cual de las alternativas es la que mejor ajusta teniendo en cuenta el resto de factores.

Cuantitativo con cualitativo

Utilizando la medida (3.4) obtenemos los siguientes resultados:

η	Potencia	Antivehi	Valorvehi	Edad	Anticarn	Malus
Sexo	**0.1311**	0.0354	**0.1663**	**0.1971**	**0.3334**	0.0171
Zona	0.1441	0.1712	**0.2091**	**0.1953**	0.1784	**0.2487**
Vehicu	**0.2891**	**0.1670**	**0.2478**	**0.1678**	**0.2375**	0.1300
Uso	0.0374	0.0514	0.0392	0.0348	0.0265	**0.1302**
Pago	0.0461	0.0442	0.0401	**0.2111**	**0.2286**	0.0510

Detallamos adicionalmente los *p*-valores resultantes de las anovas entre estos predictores:

p-valor	Potencia	Antivehi	Valorvehi	Edad	Anticarn	Malus
Sexo	**0.005**	0.451	**0.000**	**0.000**	**0.000**	0.725
Zona	0.404	0.147	**0.018**	**0.044**	0.105	**0.001**
Vehicu	**0.000**	**0.005**	**0.000**	**0.005**	**0.000**	0.053
Uso	0.426	0.274	0.404	0.459	0.573	**0.000**
Pago	0.614	0.651	0.696	**0.000**	**0.000**	0.549

Según (3.4) observamos que:

- La variable más asociada con el sexo es la antigüedad del carnet, seguida de la edad, el valor del vehículo y la potencia.
- La zona tiene una asociación elevada con casi todos los factores cuantitativos, y de nuevo recordamos que es la de mayor número de categorías. Los p-valores asociados a la zona demuestran que aunque es la que más diferencia las medias, estas diferencias no son las más significativas.
- El tipo de vehículo está bastante asociado con todos los factores cuantitativos debido a sus clases desproporcionadas, las cuales hemos decidido no agrupar en el análisis. Observando los gráficos de medias nos haremos una idea del sentido de tales diferencias.
- El uso del vehículo está asociado con el nivel de malus.

- La forma de pago está asociada con la antigüedad del carnet y la edad del conductor.

En el anexo 5.7 encontramos las anovas de los *p*-valores en negrilla, además de la del malus con el sexo. A partir de los correspondientes gráficos de medias realizamos los siguientes comentarios:

- Potencia y sexo: Las media de potencia es mayor en los hombres.
- Potencia y tipo de vehículo: Los balilla duros son los de mayor potencia, seguidos por los balilla blandos, los todo terreno y el resto de vehículos.
- Antigüedad y tipo de vehículo: Los todo terrenos son los más antiguos seguidos por los balilla duros, los balilla blandos y el resto de vehículos.
- Valor del vehículo y sexo: Los hombres conducen vehículos de mayor valor, al igual que de mayor potencia. Ya hemos visto como la potencia y el valor del vehículo están estrechamente relacionadas.
- Valor del vehículo y zona: Según la zona observamos diferentes medias en el valor del vehículo.
- Valor y tipo de vehículo: Observamos un gráfico bastante similar al de la potencia con el tipo de vehículo.
- Edad y sexo: La media de edad en los hombres es mayor que la de las mujeres.
- Edad y zona: Observamos diferentes medias de edad en las diferentes zonas, pero globalmente las diferencias no son muy significativas.
- Edad y tipo de vehículo: Observamos que las medias de edad menores son las de los balilla. Los balilla duros que hemos visto eran los de mayor valor y potencia y ahora observamos que van asociados a menores edades.
- Edad y forma de pago: Este gráfico nos ayuda a entender que los asegurados con pagos fraccionados son los jóvenes, que en general tendrán más siniestros pero de cuantías menores.
- Antigüedad del carnet y sexo: Esta visión de medias es simétrica a la de la edad y el sexo. Puesto que el colectivo de mujeres es más joven, también será el de menor antigüedad del carnet.
- Antigüedad del carnet y tipo de vehículo: Observamos un gráfico bastante similar al de la edad y el tipo de vehículo.
- Antigüedad del carnet y forma de pago: Observamos un gráfico bastante similar al de la edad del conductor principal y la forma de pago.
- Nivel de malus y sexo: Recordemos que al inicio plasmamos el gráfico de medias para la variable Bonus respecto al sexo y observamos como las mujeres tenían una media superior, aunque no se observaba una diferencia significativa. Ahora observamos de nuevo en el gráfico como las mujeres

también tienen un mejor nivel de malus.

- Nivel de malus y zona: Observamos diferentes niveles de Malus en las zonas. El malus era la variable cuantitativa más asociada con la zona. Es un hecho esperado, pues las zonas se forman inicialmente a partir de la información de siniestralidad de períodos anteriores, y la variable malus es una variable histórica que se supone que recoge tal información.
- Nivel de malus y uso del vehículo: El uso particular está asociado a un mejor nivel de malus que el uso empresarial como era de esperar, especialmente para las pólizas que han sufrido siniestro.

5.2.2. Modelo lineal generalizado

Hacemos uso de la combinación: distribución del error Gamma y enlace Logarítmico. Utilizamos una variable continua para cada predictor cuantitativo: potencia, antigüedad y valor del vehículo, edad del primer conductor y antigüedad del carnet, y malus, en total seis. Utilizamos una binaria para el sexo (que tiene dos categorías), nueve para la zona de circulación (que tiene diez categorías), tres para el tipo de vehículo (que tiene cuatro), una para el uso del vehículo (que tiene dos) y dos para la forma de pago (que tiene tres).

Proceso de selección paso a paso

Hacemos uso de la distribución asintótica F de Fisher (3.41) en la contrastación de (3.38). A continuación detallamos los resultados del proceso de selección paso a paso:

Fases de introducción:

Variable	*p*-valor	*p*-valor|F(1)	*p*-valor |F(1)F(2)	*p*-valor|F(1)...F(3)
Potencia	0.740	0.724	0.622	0.469
Antivehi	0.3833	0.387	**0.368**	--------
Valorvehi	0.873	0.821	0.937	0.777
Edad	0.643	0.641	0.445	0.430
Anticarn	0.747	0.714	0.461	0.401
Malus	0.564	0.839	0.889	0.974
Sexo	0.295	**0.291**	--------	--------
Zona	0.451	0.463	0.373	**0.303**
Vehicu	0.541	0.553	0.542	0.559
Uso	**0.255**	--------	--------	--------
Pago	0.595	0.648	0.582	0.589
	F(1) = **Uso**	F(2) = **Sexo**	F(3) = **Antivehi**	F(4) = **Zona**

Variable	*p*-valor|F(1)...F(4)	*p*-valor|F(1)...F(5)	*p*-valor|F(1)...F(6)	*p*-valor|F(1)...F(7)
Potencia	0.409	**0.442**	--------	--------
Antivehi	--------	--------	--------	--------
Valorvehi	0.451	0.511	0.928	1
Edad	0.491	0.981	0.983	**0.965**
Anticarn	**0.319**	--------	--------	--------
Malus	0.983	0.938	0.956	1
Sexo	--------	--------	--------	--------
Zona	--------	--------	--------	--------
Vehicu	0.590	0.647	**0.541**	--------
Uso	--------	--------	--------	--------
Pago	0.689	0.803	0.714	0.700
	F(5) = **Anticarn**	F(6) = **Potencia**	F(7) = **Vehicu**	F(8) = **Pago**

Fases de eliminación:

Variable	*p*-valores
F(1)	**0.255**
F(2) | F(1)	**0.291**
F(1) | F(2)	0.252
F(3) | F(1)F(2)	**0.368**
F(1) | F(2)F(3)	0.255
F(2) | F(1)F(3)	0.278
F(4) | F(1)F(2)F(3)	**0.303**
F(1) | F(2)F(3)F(4)	0.212
F(2) | F(1)F(3)F(4)	0.129
F(3) | F(1)F(2)F(4)	0.194
F(5) | F(1)F(2)F(3)F(4)	**0.319**
F(1) | F(2)F(3)F(4)F(5)	0.204
F(2) | F(1)F(3)F(4)F(5)	0.078
F(3) | F(1)F(2)F(4)F(5)	0.165
F(4) | F(1)F(2)F(3)F(5)	0.283
F(6) | F(1)F(2)F(3)F(4)F(5)	**0.442**
F(1) | F(2)F(3)F(4)F(5)F(6)	0.193
F(2) | F(1)F(3)F(4)F(5)F(6)	0.063
F(3) | F(1)F(2)F(4)F(5)F(6)	0.121
F(4) | F(1)F(2)F(3)F(5)F(6)	0.272
F(5) | F(1)F(2)F(3)F(4)F(6)	0.343

La primera variable a entrar en el modelo es F(1) = Uso, al igual que en la RBD que presentamos en el siguiente apartado.

La segunda es F(2) = Sexo, a diferencia que en la RBD, en la que es el Malus. Observamos aquí, que en la primera fase el Malus tiene asociado un *p*-valor de 0.564, que pasa a 0.836 en la segunda. Con el MLG, al introducir el Uso, es decir, sus dos binarias asociadas, el Malus es menos significativo en la

explicación de la respuesta. Esto es debido a que se trata a las dos binarias asociadas al uso como predictores continuos, los cuales toman valores de cero y uno. Si calculamos la correlación entre ambas binarias y el Malus obtenemos lo siguiente:

Uso particular y Malus: $\rho = -0.197$

Uso empresa y Malus: $\rho = 0.197$

La variable Malus contiene ceros para las pólizas de nivel bonus y valores relativamente pequeños (entre cero y uno) para las pólizas de escala malus. La variable Uso particular contiene unos para los vehículos de uso particular y ceros para el resto. Así, al obtener una correlación negativa, los vehículos de uso particular, en general, estarán en escala de bonus, y al revés para los de empresa, que en general estarán en escala de malus. Son variables que al ser tratadas como continuas proporcionan un perfil similar, cosa que no ocurre con la RBD, donde el malus es tratado como cuantitativo y el uso como categórico.

Siguiendo con el proceso, vamos obteniendo en los siguientes pasos que F(3) = Antivehi, F(4) = Zona, F(5) = Anticarn, F(6) = Potencia, F(7) = Vehicu y F(8) = Pago, o bien F(8) = Edad si decidimos no tener en cuenta la forma de pago.

Si restringimos el nivel de significación permitido, de hecho no seleccionaríamos ninguna variable. Pero dejamos margen para ver cuales de los factores potenciales son los más significativos. Dejando un buen margen y guiados por el patrón de aumento de los valores de probabilidad, podemos detener el proceso en el paso sexto por ejemplo. Igualmente hemos realizado dos fases más de introducción.

Nos quedamos con las cinco primeras variables como mucho:

Uso, **Sexo**, **Antivehi**, **Zona** y **Anticarn**

Los coeficientes resultantes para el predictor lineal son:

Término	Coeficiente	*t*-valor
Constante[26]	14.0783	16.9089
Uso particular	–0.5213	–0.8877
Hombre	–0.3729	–1.2777
Antigüedad del vehículo	0.0237	1.0358
Zona 1	0.2297	0.3488
Zona 2	0.1909	0.3062
Zona 3	0.5008	0.9918
Zona 4	0.9246	1.6847
Zona 5	0.4741	0.8352
Zona 6	0.1128	0.2148
Zona 7	0.4609	0.9251
Zona 8	0.2329	0.2951
Zona 9	0.6277	0.8582
Antigüedad del carnet	0.0103	0.7831

Respecto a la validación del modelo, vemos primeramente la significación conjunta, o lo que es lo mismo, el poder predictivo de la regresión, calculando el cociente de verosimilitudes (3.46) y el *p*-valor del estadístico (3.45) asociado al contraste (3.44):

$$R^2 = \frac{(1075.927 - 1040.798)}{1075.925} = 0.03265$$

$$F_{13,441} = \frac{(1075.927 - 1040.798)/13}{1040.798/441} = 1.14497 \qquad p\text{-valor} = 0.1333$$

Como era de esperar el *p*-valor obtenido no es muy pequeño, pues los predictores incluidos no son demasiado significatvos.

Plasmamos los gráficos de residuos que realiza el programa S-Plus:

[26] Uso empresarial, Mujer y Zona 10.

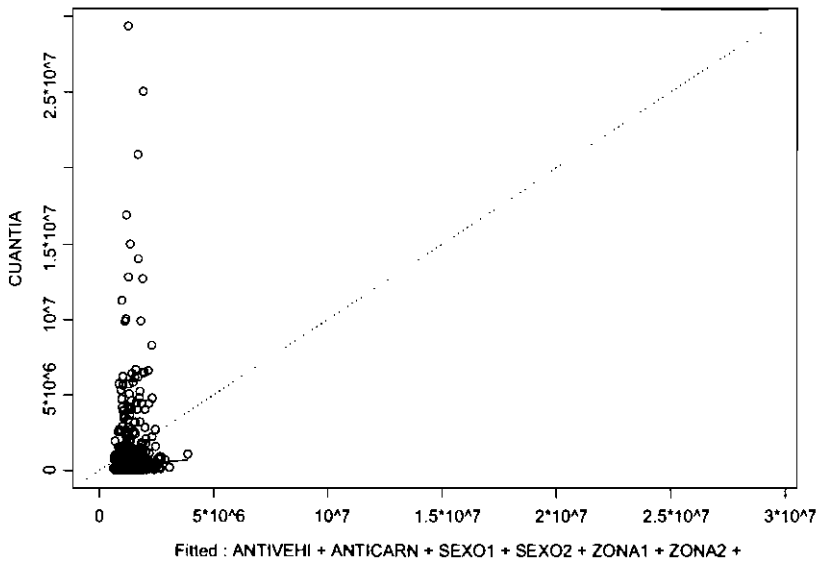


Figura 5.1. Gráfico de la respuesta *versus* la estimación.

Observamos que aún e incorporando un número elevado de parámetros, no conseguimos muy buen ajuste. Si comparamos este gráfico con el gráfico de la figura 4.1 (apartado 4.4.4), el cual es un gráfico paralelo realizado para la RBD teniendo en cuenta el Uso, Malus, Sexo y Vehicu podemos comprobar visualmente el mejor ajuste de este último.

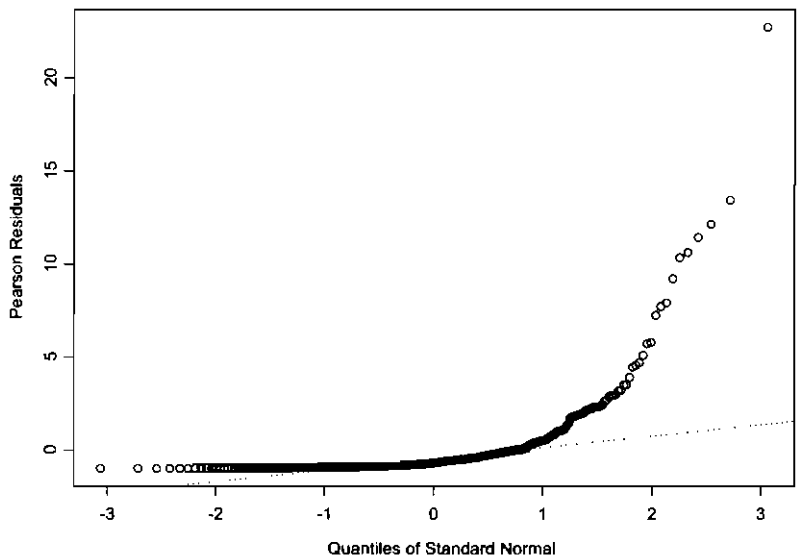


Figura 5.2. Gráfico de los residuos de Pearson.

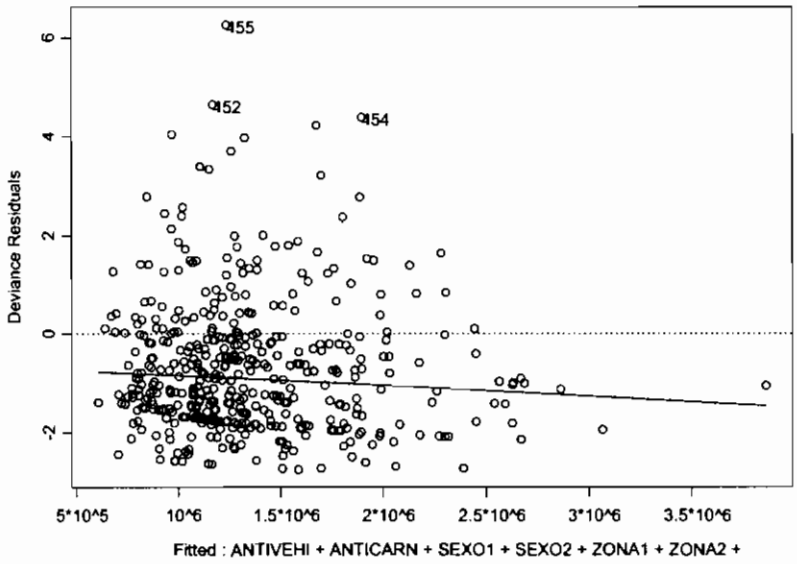


Figura 5.3. Gráfico de los residuos de desvianza.

Aunque el programa S-Plus nos marca tres puntos aislados, no olvidemos que ya hemos retirado del conjunto las dos cuantías extremas. Hemos comprobado que si se eliminan estos tres puntos, vuelven a salir otros dos diferentes a eliminar, debido al amplio rango de las cuantías.

5.2.3. Regresión basada en distancias

Hacemos uso del índice de similaridad de Gower, (4.8). Damos tratamiento cuantitativo a la potencia, la antigüedad y el valor del vehículo, a la edad del primer conductor y su antigüedad en el carnet, y a la variable malus. Damos tratamiento cualitativo al sexo del primer conductor, a la zona de circulación, al tipo y uso del vehículo, y a la forma de pago. Utilizamos $B = 500$ muestras para realizar los cálculos relacionados con la metodología *bootstrap* en la estimación de los valores de probabilidad del proceso de selección.

Proceso de selección paso a paso

Fases de introducción:

Variable	*p*-valor	*p*-valor|F(1)	*p*-valor |F(1)F(2)	*p*-valor|F(1)...F(3)	*p*-valor|F(1)...F(4)	*p*-valor|F(1)...F(5)
Potencia	0.950	0.946	0.978	0.982	0.974	0.900
Antivehi	0.948	0.942	0.942	0.942	0.942	0.990
Valorvehi	0.860	0.856	0.858	0.858	0.864	0.826
Edad	0.900	0.900	0.900	0.902	0.900	**0.492**
Anticarn	0.814	0.832	0.836	0.842	**0.846**	--------
Malus	0.502	**0.500**	--------	--------	--------	--------
Sexo	0.516	0.518	**0.546**	--------	--------	--------
Zona	0.948	0.950	0.948	0.942	0.932	0.934
Vehicu	0.814	0.816	0.652	**0.624**	--------	--------
Uso	**0.498**	--------	--------	--------	--------	--------
Pago	0.804	0.838	0.806	0.794	**0.818**	0.358
	F(1) = **Uso**	F(2) = **Malus**	F(3) = **Sexo**	F(4) = **Vehicu**	F(5) = **Anticarn**	F(6) = **Edad**

Fases de eliminación:

Variable	*p*-valores
F(1)	**0.498**
F(2) | F(1)	0.500
F(1) | F(2)	**0.544**
F(3) | F(1)F(2)	**0.546**
F(1) | F(2)F(3)	0.534
F(2) | F(1)F(3)	0.496
F(4) | F(1)F(2)F(3)	**0.624**
F(1) | F(2)F(3)F(4)	0.548
F(2) | F(1)F(3)F(4)	0.496
F(3) | F(1)F(2)F(4)	0.516
F(5) | F(1)F(2)F(3)F(4)	**0.846**
F(1) | F(2)F(3)F(4)F(5)	0.490
F(2) | F(1)F(3)F(4)F(5)	0.530
F(3) | F(1)F(2)F(4)F(5)	0.754
F(4) | F(1)F(2)F(3)F(5)	0.430

Observamos que en la segunda fase de eliminación el *p*-valor asociado al Uso es mayor que el del Malus, por lo que deberíamos proceder a su eliminación: F[1] = Uso. Si aceptáramos que ambos *p*-valores no superan el pre-establecido, deberíamos seguir el proceso con las dos variables incorporadas. Por seguridad, comprobamos si dado el Malus la siguiente a ser introducida sería el Uso:

Variable	*p*-valor|Malus
Potencia	0.976
Antivehi	0.946
Valorvehi	0.858
Edad	0.900
Anticarn	0.838
Sexo	0.558
Zona	0.948
Vehicu	0.644
Uso	**0.544**
Pago	0.786
	F(2) = **Uso**

Y obtenemos que la siguiente es el Uso, por lo que continuamos introduciendo dados el Uso y el Malus.

En el paso quinto hemos continuado, considerando que la variable Pago no iba a ser introducida (ya que recordemos que obteníamos menores cuantías a mayor fraccionamiento), por lo que F(5) = Anticarn.

En el paso sexto, en que F(6) = Edad, detenemos el proceso, pues obtenemos los siguientes resultados en la fase de introducción:

	Potencia	Antivehi	Valorvehi	Edad	Zona
p-valor	0.900	0.990	0.826	**0.492**	0.934
Qparcial	2.508	0.373	5.253	**6.085E5**	0.164
Geo0	3.8352E15	3.8352E15	3.8352E15	3.8352E15	3.8352E15
Geo1	3.5335E15	3.5335E15	3.5335E15	3.5335E15	3.5335E15
Geo2	3.7492E15	3.6155E15	3.7870E15	**3.8352E15**	3.5761E15

Al introducir la variable que correspondería, la Edad, ya explicamos toda la variabilidad de los datos, por lo que introducir otra variable no mejoraría el ajuste. Por lo que detenemos el proceso.

En este proceso ya hemos indicado que en el paso quinto se nos presentaba la opción de incluir o no la variable pago. Los resultados anteriores corresponden a no incluirla por las razones ya indicadas. Sin embargo podemos analizar también lo que ocurre si la incluimos:

Fases de introducción:

Variable	*p*-valor	*p*-valor|F(1)	*p*-valor |F(1)F(2)	*p*-valor|F(1)...F(3)	*p*-valor|F(1)...F(4)	*p*-valor|F(1)...F(5)
Potencia	0.950	0.946	0.978	0.982	0.974	0.970
Antivehi	0.948	0.942	0.942	0.942	0.942	0.940
Valorvehi	0.860	0.856	0.858	0.858	0.864	**0.870**
Edad	0.900	0.900	0.900	0.902	0.900	0.900
Anticarn	0.814	0.832	0.836	0.842	0.846	0.912
Malus	0.502	**0.500**	--------	--------	--------	--------
Sexo	0.516	0.518	**0.546**	--------	--------	--------
Zona	0.948	0.950	0.948	0.942	0.932	0.926
Vehicu	0.814	0.816	0.652	**0.624**	--------	--------
Uso	**0.498**	--------	--------	--------	--------	--------
Pago	0.804	0.838	0.806	0.794	**0.818**	--------
	F(1) = **Uso**	F(2) = **Malus**	F(3) = **Sexo**	F(4) = **Vehicu**	F(5) = **Pago**	F(6) = **Valorveh**

Fases de eliminación:

Variable	*p*-valores
F(5) | F(1)F(2)F(3)F(4)	**0.818**
F(1) | F(2)F(3)F(4)F(5)	0.580
F(2) | F(1)F(3)F(4)F(5)	0.514
F(3) | F(1)F(2)F(4)F(5)	0.520
F(4) | F(1)F(2)F(3)F(5)	0.674
F(6) | F(1)F(2)F(3)F(4)F(5)	**0.870**
F(1) | F(2)F(3)F(4)F(5)F(6)	0.758
F(2) | F(1)F(3)F(4)F(5)F(6)	0.518
F(3) | F(1)F(2)F(4)F(5)F(6)	0.646
F(4) | F(1)F(2)F(3)F(5)F(6)	0.576
F(5) | F(1)F(2)F(3)F(4)F(6)	0.844

El Pago entraría en el paso quinto en lugar de la Anticarn. Observamos que en el paso sexto F(6) = Valorveh en lugar de Anticarn. Este hecho podría tener explicación teniendo en cuenta la alta asociación entre la forma de pago y la antigüedad del carnet.

Como ya se ha comentado en el trabajo, es usual recoger la información desprendida de los procesos de selección no sólo del conjunto resultante de variables de tarifa, sino también la del resto de predictores que finalmente no han sido incluidos. En este caso, al igual que nos ha ocurrido con el MLG si restringimos el nivel de significación no acabaríamos seleccionando ningún predictor. Por lo que en lugar de restringir el nivel de significación nos fijamos en el patrón de aumento de los *p*-valores en la selección final del conjunto.

En el apartado 4.4.4 hemos realizado la validación exhaustiva del modelo resultante que utiliza finalmente como variables de tarifa las siguientes:

Uso, Malus, Sexo, y Vehicu

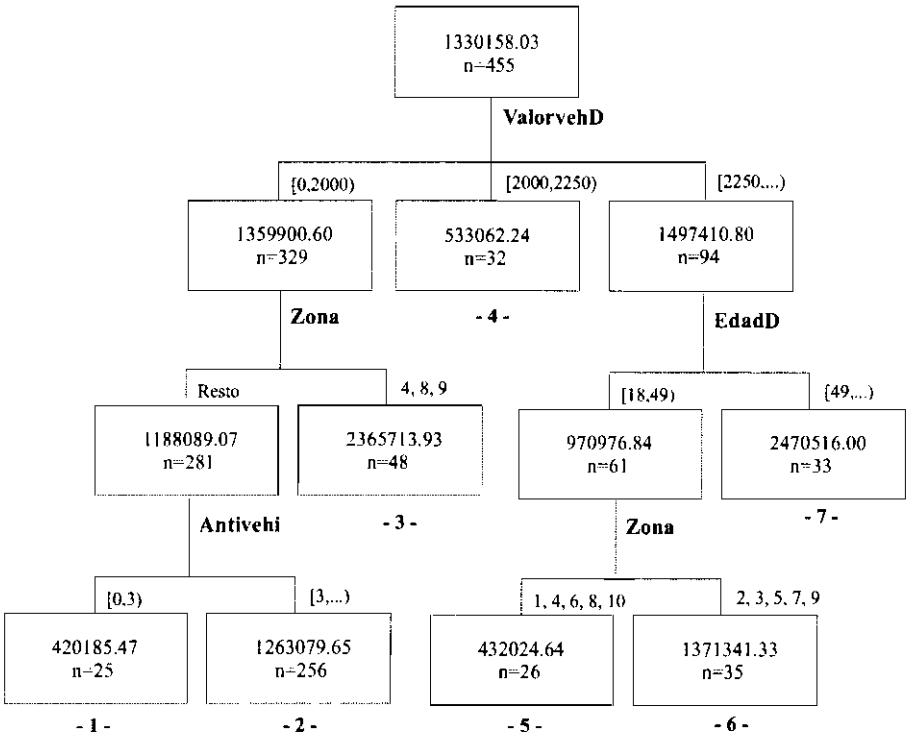
5.2.4. Análisis de segmentación

Utilizamos el programa SPSS CHAID ordinal. Para ello necesitamos discretizar la respuesta con Ward 31, y los predictores inicialmente continuos o discretos con un número elevado de clases. Nos referimos al apartado 3.6.2 del trabajo para el detalle de ambas discretizaciones. Respecto a la respuesta utilizamos, como ya hemos indicado, la discretización de Ward, y para los predictores potencia, valor del vehículo, antigüedad del carnet y edad del conductor habitual, los intervalos iniciales construidos en dicho apartado.

Hemos utilizado como predictores libres al sexo, la zona, el uso y el tipo de vehículo, y la forma de pago. Hemos utilizado como predictores monótonos a la potencia, la edad, la antigüedad del carnet, el valor del vehículo, la antigüedad del vehículo y el nivel de malus. Cabe notar que también se hubiera podido considerar a todas las variables como predictores libres, obteniendo en tal caso resultados diferentes. Respecto a las variables cuantitativas potencia, edad, antigüedad del carnet y valor del vehículo, se han utilizado los intervalos iniciales ya mencionados, y respecto al malus y la antigüedad del vehículo, las clases discretas ordinales originales. Se ha utilizado, en todos los árboles que presentamos a continuación, un nivel de significación para la selección del mejor predictor de 0.7, muy elevado, pero sino el proceso no consideraba que ningún predictor distinguiera suficientemente las medias de cuantía.

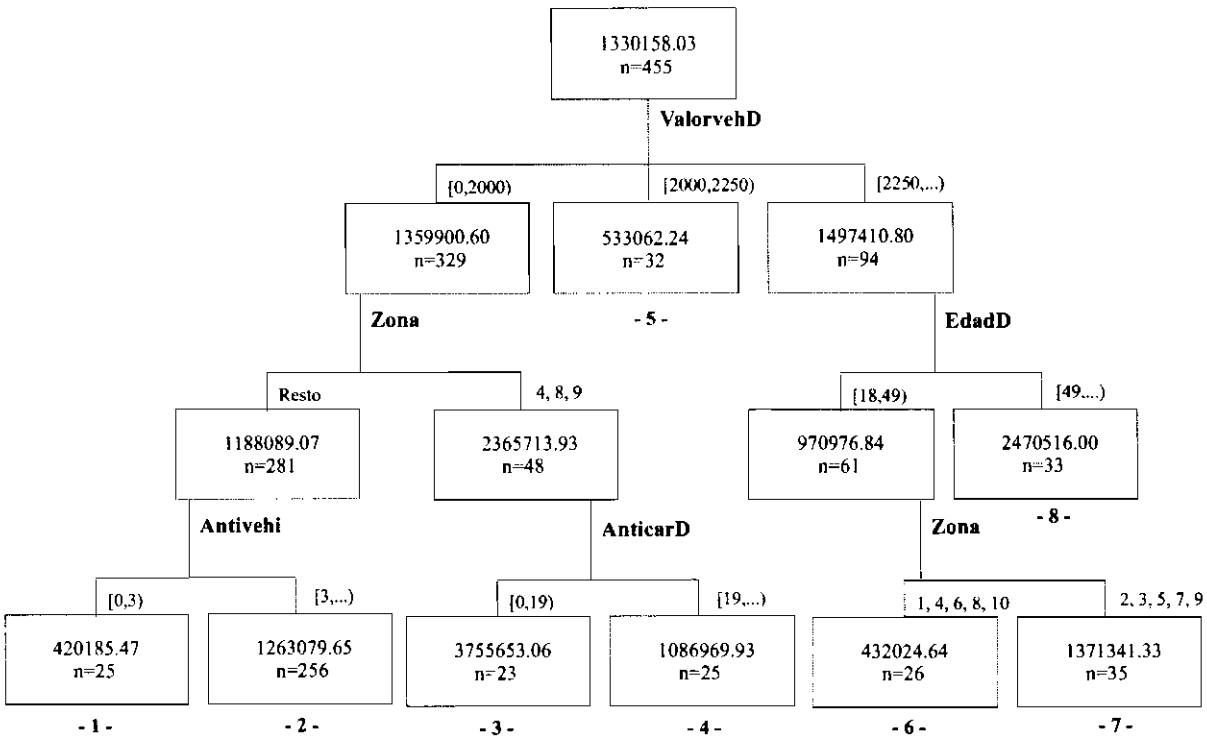
En el árbol 1, el nivel para la agrupación de categorías de un predictor es de 0.05, el tamaño mínimo para analizar un segmento de 50 y el mínimo para formarlo de 25:

Árbol 1



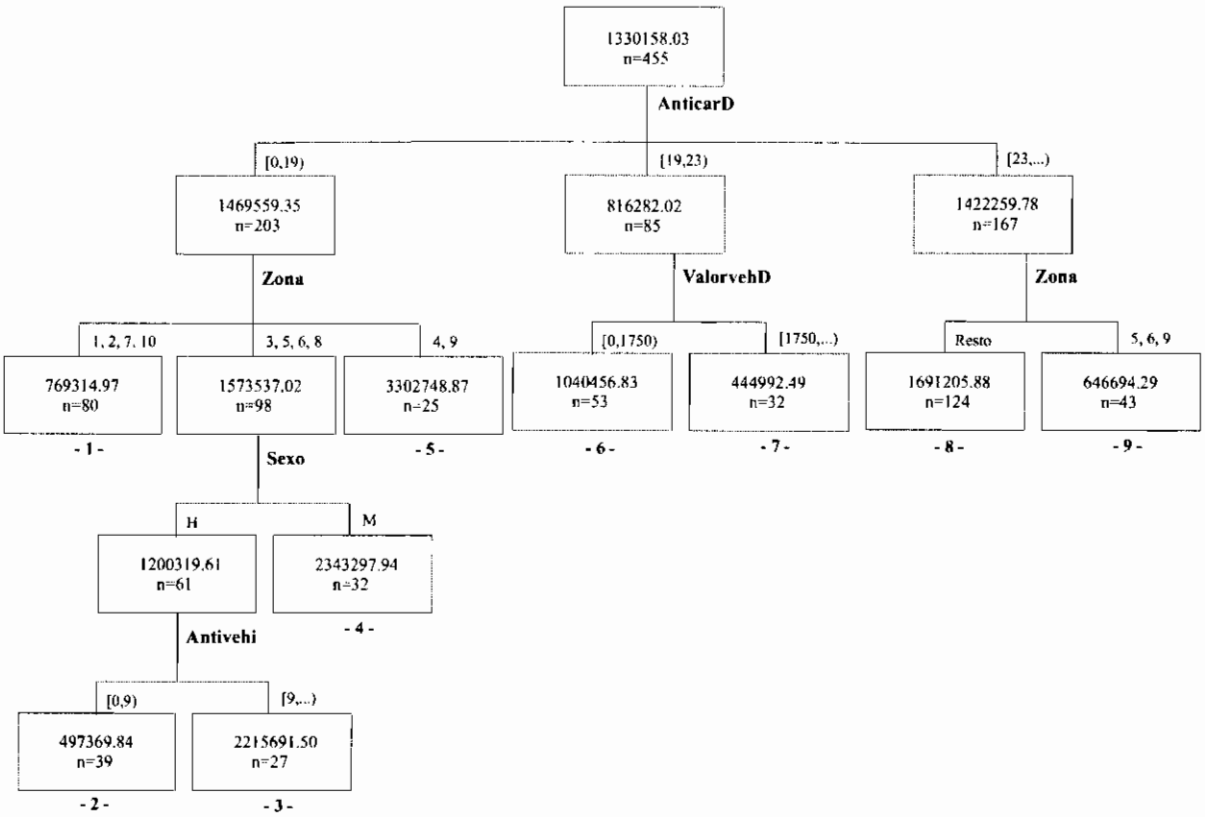
En el árbol 2 variamos los tamaños mínimos, para analizar de 25 y para formar de 14, obteniendo dos nodos terminales adicionales:

Árbol 2



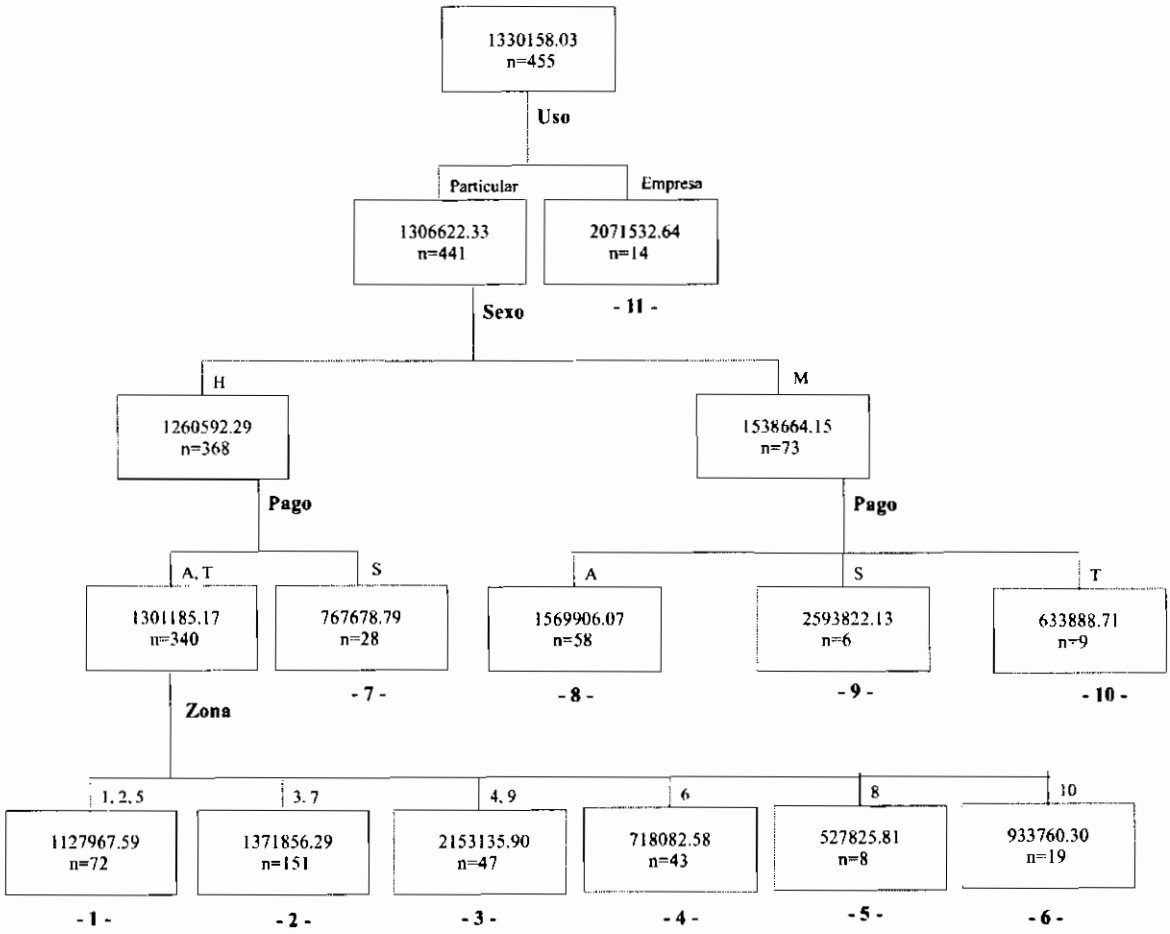
En el árbol 3, el nivel para la agrupación de categorías es de 0.1, el tamaño mínimo para analizar un segmento de 50 y el mínimo para formarlo de 25:

Árbol 3

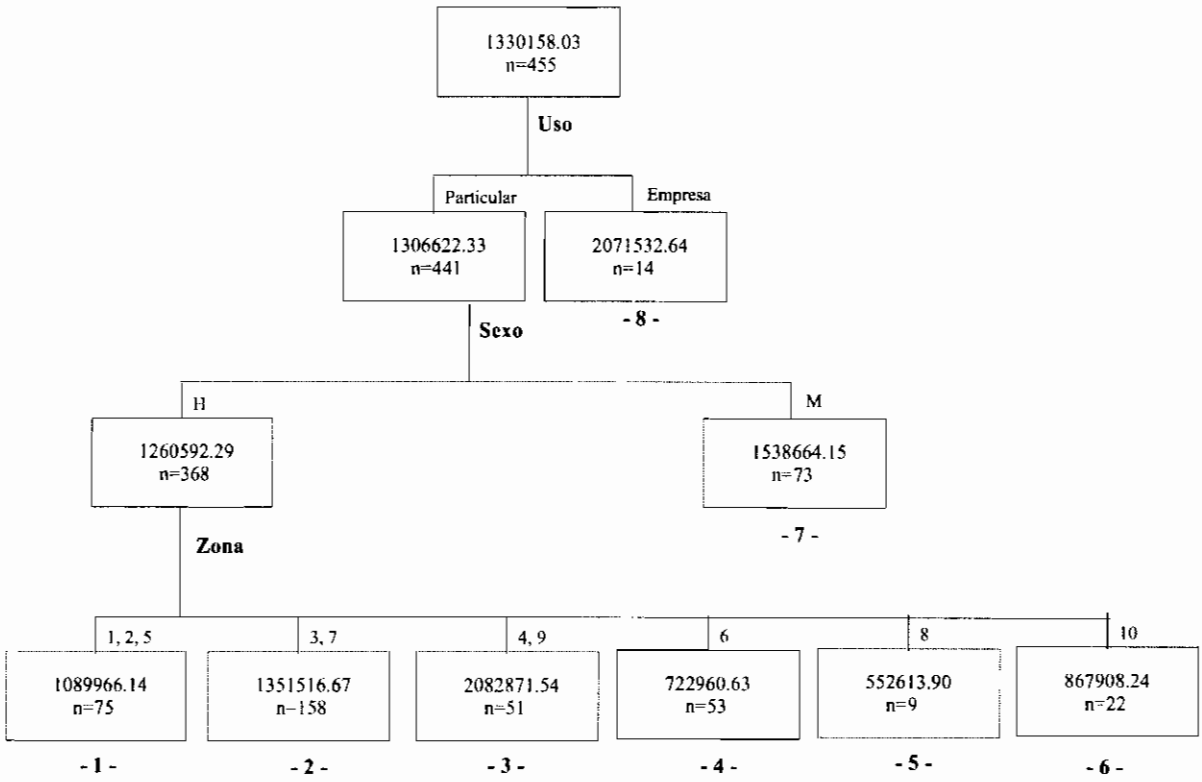


En los árboles 4 y 5 el nivel para la agrupación de categorías ha sido de 0.8, para que casi no agrupe las clases de los predictores, exigiendo tan solo un tamaño mínimo para analizar un segmento de 12 y uno mínimo para formarlo de 6. La diferencia estriba en que en el árbol 5 no tenemos en cuenta la forma de pago:

Árbol 4



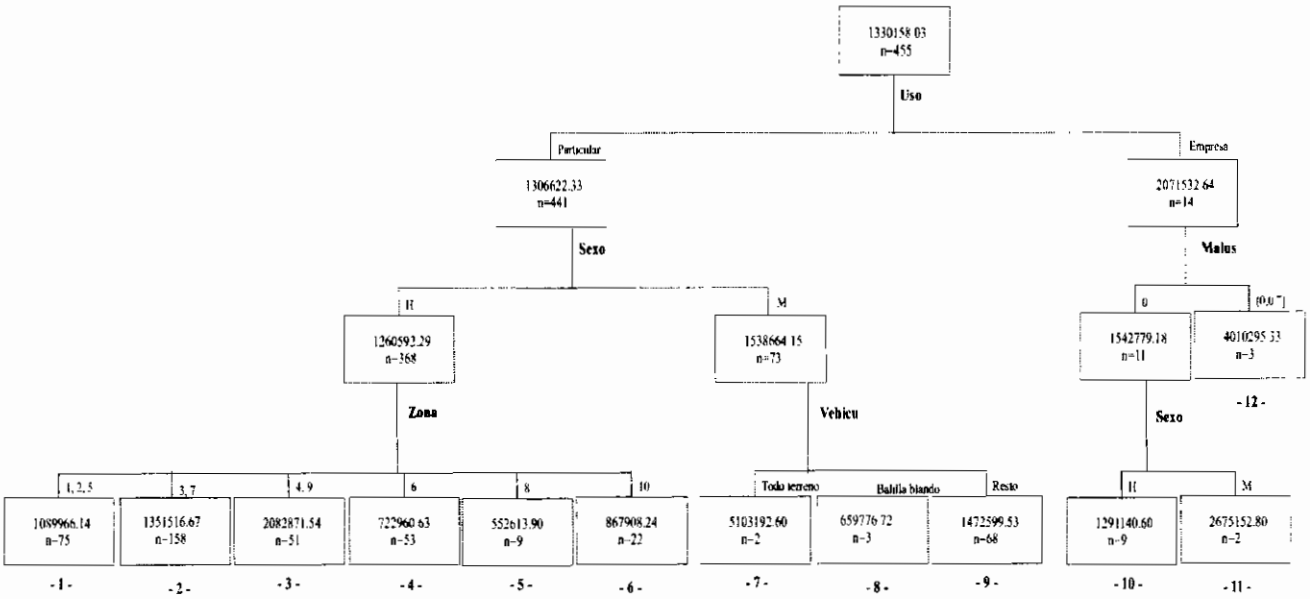
Árbol 5



Observamos que según variamos los parámetros los resultados son diferentes. Los de los árboles 4 y 5 se parecen más a los obtenidos en los dos apartados anteriores con los modelos de regresión.

Finalmente, en el árbol 6, variamos respecto del árbol 5 los tamaños mínimos, para ver qué ocurre con el tipo de vehículo, utilizando como tamaños mínimos ambos de 2 (ponemos 2 por la frecuencia de 2 del tipo de vehículo):

Árbol 6



Observamos en el árbol 6 una mayor concordancia con los resultados del MLG y de la RBD.

Puesto que es usual que la interacción de la edad y el sexo del primer conductor sea significativa en la siniestralidad, y no observamos que aparezca en ninguno de los árboles anteriores, hemos realizado un estudio teniendo en cuenta sólo ambas variables. Para poder observar las diferencias de medias en dicha interacción, nos vemos forzados a fijar el máximo valor de 1 tanto para la fase de agrupación (para que no agrupe las clases, ya que de otro modo no encuentra diferencias), como para la fase de selección del mejor predictor (para que los incorpore en el árbol). En tal caso obtenemos que en el primer nivel aparece el sexo y en el segundo la edad tan sólo para los hombres, separando, ya que así lo hemos establecido, cada una de las clases iniciales de edad. De lo que concluimos que en con estos datos tal interacción no es necesaria.

Anexo 5.6. Anovas de las cuantías con los factores cualitativos

Sexo

Descriptivos

CUANTIA

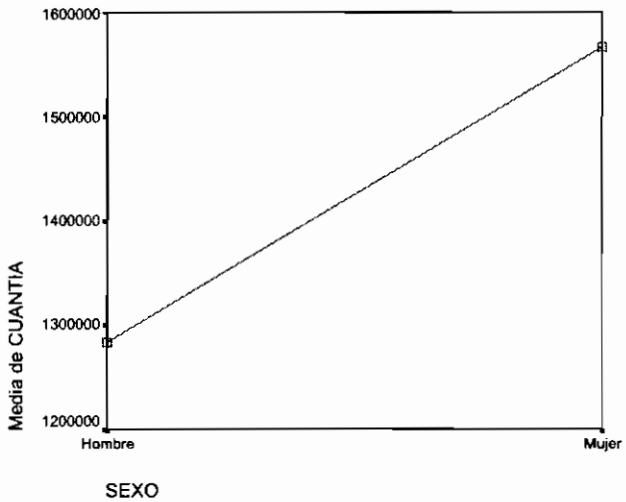
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Hombre	380	1283502	2973580.60	152541.47	983568.52441	1583436	13080.000	2.9E+07
Mujer	75	1566548	2544228.36	293782.19	981174.16433	2151922	13080.000	1.3E+07
Total	455	1330158	2906476.99	136257.74	1062383.896	1597932	13080.000	2.9E+07

ANOVA

CUANTIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	5.018E+12	1	5.02E+12	.594	.441
Intra-grupos	3.830E+15	453	8.46E+12		
Total	3.835E+15	454			

Gráfico de las medias



Zona

Descriptivos

CUANTIA

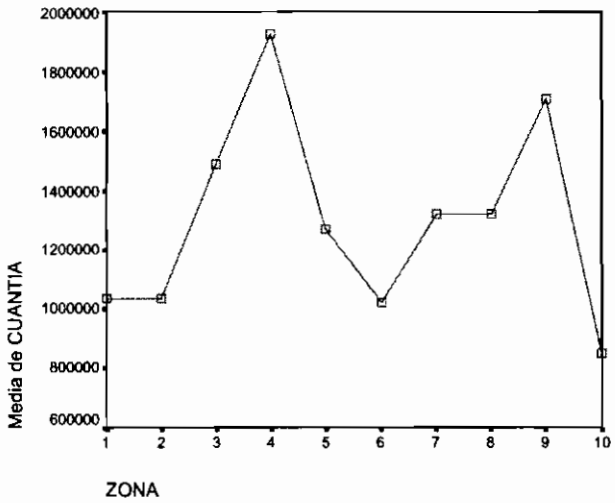
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	20	1035274	2231788.96	499043.18	-9235.78392	2079783	13080.000	9876217
2	25	1033793	1389632.63	277926.53	460180.52075	1607405	13080.000	5309530
3	97	1489939	2590056.93	262980.44	967927.43270	2011952	20000.000	1.5E+07
4	47	1923858	4772846.31	696191.19	522497.84580	3325218	21000.000	2.5E+07
5	40	1274300	2847218.42	450184.76	363714.89755	2184884	84000.000	1.7E+07
6	69	1024103	2157853.00	259774.97	505730.08033	1542476	13080.000	1.1E+07
7	109	1321821	3186123.91	305175.32	716910.69513	1926732	13080.000	2.9E+07
8	11	1324048	1950061.44	587965.65	13978.72028	2634117	20000.000	6173632
9	14	1709382	3613098.13	965641.09	-376758.537	3795523	45000.000	1.4E+07
10	23	848551.6	1139093.23	237517.36	355970.75938	1341132	100000.0	5730520
Total	455	1330158	2906476.99	136257.74	1062383.896	1597932	13080.000	2.9E+07

ANOVA

CUANTIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	3.692E+13	9	4.10E+12	.481	.888
Intra-grupos	3.798E+15	445	8.54E+12		
Total	3.835E+15	454			

Gráfico de las medias



Tipo de vehículo

Descriptivos

CUANTIA

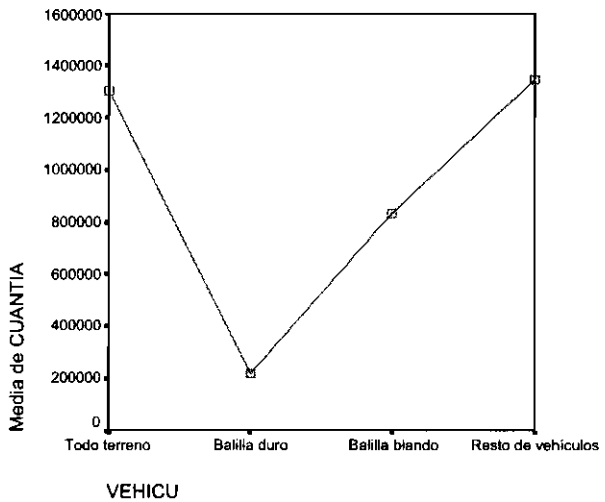
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Todo terreno	22	1300720	2191706.82	467273.46	328971.49868	2272468	62218.000	9887069
Balilla duro	2	215000.0	261629.509	185000.00	-2135647.88	2565648	30000.000	400000.0
Balilla blando	6	832483.0	701614.542	286432.94	96183.69403	1568782	73818.000	1937440
Resto de vehículos	425	1343956	2965102.11	143828.58	1061249.863	1626661	13080.000	2.9E+07
Total	455	1330158	2906476.99	136257.74	1062383.896	1597932	13080.000	2.9E+07

ANOVA

CUANTIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	4.073E+12	3	1.36E+12	.160	.923
Intra-grupos	3.831E+15	451	8.49E+12		
Total	3.835E+15	454			

Gráfico de las medias



Uso del vehículo

Descriptivos

CUANTIA

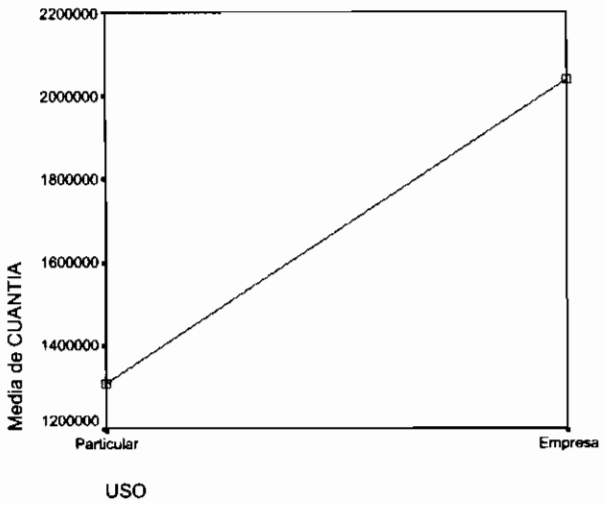
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Particular	441	1307699	2916040.11	138859.05	1034790.002	1580609	13080.000	2.9E+07
Empresa	14	2037603	2580064.77	689551.31	547918.31156	3527288	13080.000	8281658
Total	455	1330158	2906476.99	136257.74	1062383.896	1597932	13080.000	2.9E+07

ANOVA

CUANTIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	7.229E+12	1	7.23E+12	.855	.355
Intra-grupos	3.828E+15	453	8.45E+12		
Total	3.835E+15	454			

Gráfico de las medias



Forma de pago

Descriptivos

CUANTIA

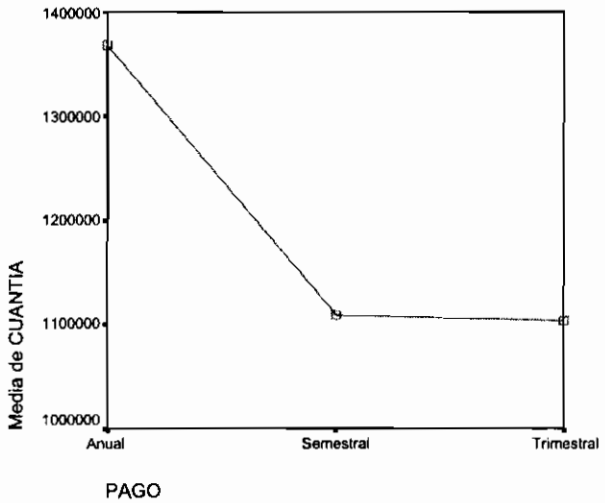
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	388	1368871	2995058.41	152051.05	1069921.154	1667820	13080.000	2.9E+07
Semestral	34	1108673	1583692.73	271601.06	556096.01140	1661249	13080.000	6211970
Trimestral	33	1103188	2941954.46	512128.54	60016.51869	2146360	22238.000	1.7E+07
Total	455	1330158	2906476.99	136257.74	1062383.896	1597932	13080.000	2.9E+07

ANOVA

CUANTIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	3.949E+12	2	1.97E+12	.233	.792
Intra-grupos	3.831E+15	452	8.48E+12		
Total	3.835E+15	454			

Gráfico de las medias



ANEXO 5.7. Anovas de los factores cuantitativos con los factores cualitativos

Potencia y sexo

Descriptivos

POTENCIA

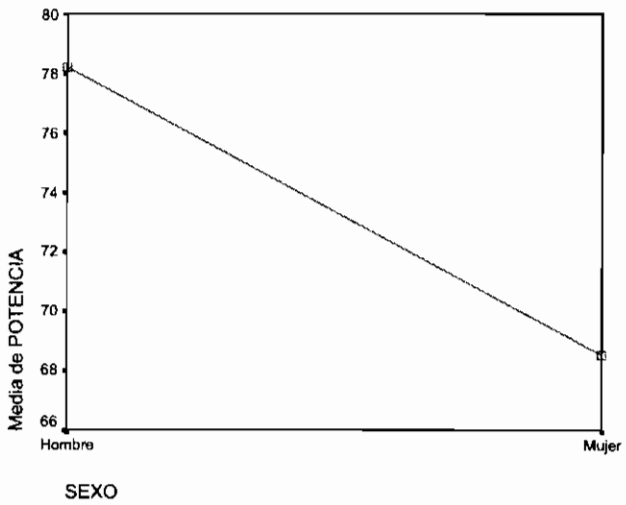
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Hombre	380	78.18	28.20	1.45	75.33	81.02	32	286
Mujer	75	68.53	20.64	2.38	63.78	73.28	37	128
Total	455	76.59	27.32	1.28	74.07	79.10	32	286

ANOVA

POTENCIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	5824.467	1	5824.467	7.923	.005
Intra-grupos	333007.854	453	735.117		
Total	338832.321	454			

Gráfico de las medias



Potencia y tipo de vehículo

Descriptivos

POTENCIA

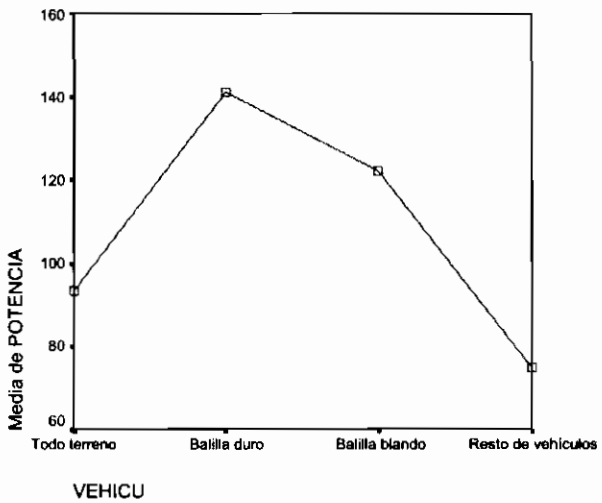
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Todo terreno	22	93.32	17.32	3.69	85.64	101.00	63	124
Balilla duro	2	141.00	12.73	9.00	26.64	255.36	132	150
Balilla blando	6	122.17	17.13	6.99	104.19	140.14	105	150
Resto de vehículos	425	74.77	26.71	1.30	72.23	77.32	32	286
Total	455	76.59	27.32	1.28	74.07	79.10	32	286

ANOVA

POTENCIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	28318.400	3	9439.467	13.710	.000
Intra-grupos	310513.921	451	688.501		
Total	338832.321	454			

Gráfico de las medias



Antigüedad del vehículo y tipo de vehículo

Descriptivos

ANTIVEHI

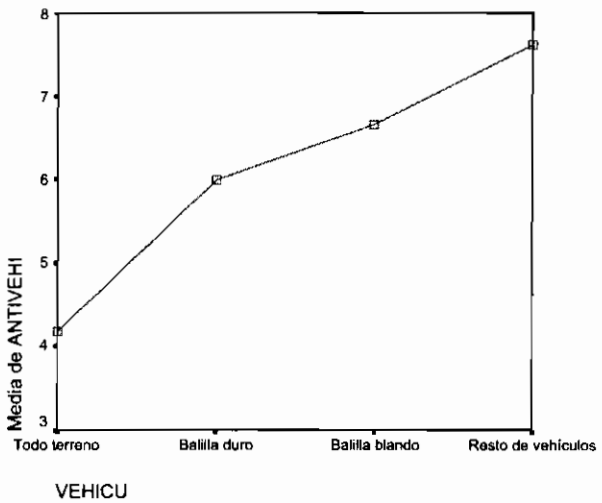
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Todo terreno	22	4.18	2.34	.50	3.14	5.22	1	9
Balilla duro	2	6.00	2.83	2.00	-19.41	31.41	4	8
Balilla blando	6	6.67	3.01	1.23	3.51	9.83	3	10
Resto de vehículos	425	7.62	4.54	.22	7.19	8.06	0	23
Total	455	7.44	4.49	.21	7.02	7.85	0	23

ANOVA

ANTIVEHI

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	255.594	3	85.198	4.315	.005
Intra-grupos	8904.371	451	19.744		
Total	9159.965	454			

Gráfico de las medias



Valor del vehículo y sexo

Descriptivos

VALORVEH

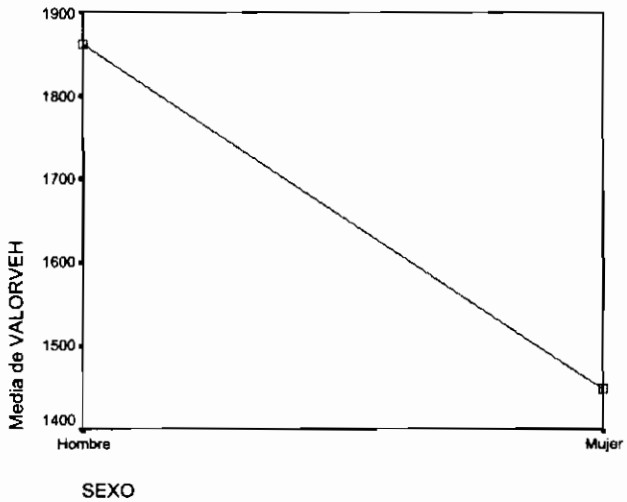
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Hombre	380	1861.67	980.62	50.30	1762.76	1960.58	600	10650
Mujer	75	1448.00	406.57	46.95	1354.46	1541.54	800	2650
Total	455	1793.48	923.75	43.31	1708.38	1878.59	600	10650

ANOVA

VALORVEH

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	10718603	1	10718603	12.890	.000
Intra-grupos	376685818	453	831536.023		
Total	387404422	454			

Gráfico de las medias



Valor del vehículo y zona

Descriptivos

VALORVEH

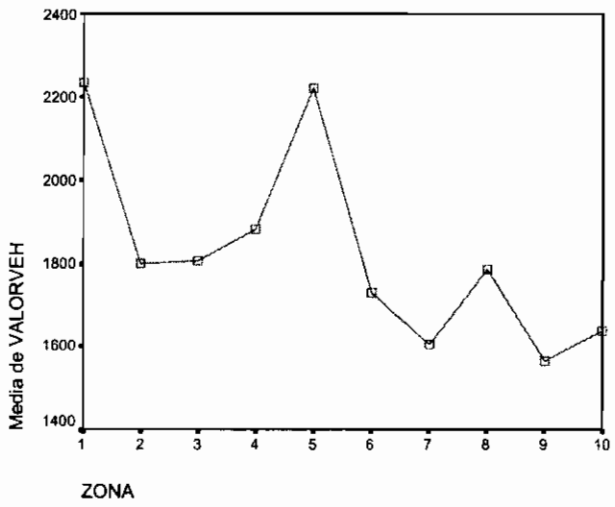
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	20	2235.00	1582.15	353.78	1494.53	2975.47	800	7200
2	25	1802.00	711.44	142.29	1508.33	2095.67	750	3900
3	97	1807.22	822.60	83.52	1641.43	1973.01	800	4350
4	47	1881.91	887.25	129.42	1621.41	2142.42	750	4250
5	40	2220.00	1685.87	266.56	1680.83	2759.17	1000	10650
6	69	1731.16	699.06	84.16	1563.23	1899.09	850	4300
7	109	1605.50	584.26	55.96	1494.58	1716.43	600	4750
8	11	1786.36	764.88	230.62	1272.51	2300.22	850	3200
9	14	1566.71	426.24	113.92	1320.61	1812.82	1000	2534
10	23	1639.13	963.42	200.89	1222.52	2055.74	800	5000
Total	455	1793.48	923.75	43.31	1708.38	1878.59	600	10650

ANOVA

VALORVEH

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	16951159	9	1883462.1	2.262	.018
Intra-grupos	370453263	445	832479.242		
Total	387404422	454			

Gráfico de las medias



Valor del vehículo y tipo de vehículo

Descriptivos

VALORVEH

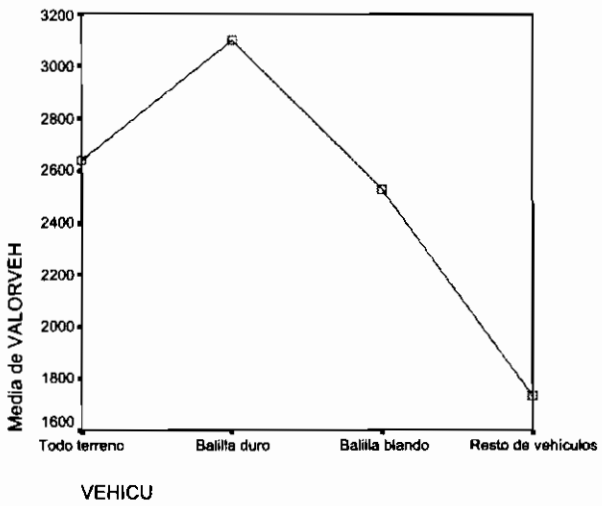
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Todo terreno	22	2636.36	830.69	177.10	2268.06	3004.67	1100	3900
Balilla duro	2	3100.00	1414.21	1000.00	-9606.20	15806.20	2100	4100
Balilla blando	6	2525.00	731.27	298.54	1757.58	3292.42	1950	3800
Resto de vehículos	425	1733.37	901.32	43.72	1647.44	1819.31	600	10650
Total	455	1793.48	923.75	43.31	1708.38	1878.59	600	10650

ANOVA

VALORVEH

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	23790085	3	7930028.3	9.836	.000
Intra-grupos	363614337	451	806240.214		
Total	387404422	454			

Gráfico de las medias



Edad y sexo

Descriptivos

Edad del cond habi

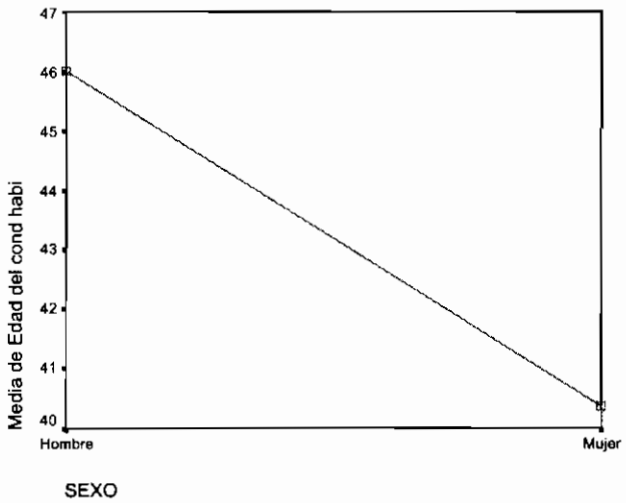
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Hombre	380	46.0031	10.6783	.5478	44.9260	47.0802	23.36	76.97
Mujer	75	40.3787	8.8688	1.0241	38.3381	42.4192	22.27	71.92
Total	455	45.0760	10.6007	.4970	44.0994	46.0526	22.27	76.97

ANOVA

Edad del cond habi

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1981.496	1	1981.496	18.305	.000
Intra-grupos	49036.368	453	108.248		
Total	51017.864	454			

Gráfico de las medias



Edad y zona

Descriptivos

Edad del cond habi

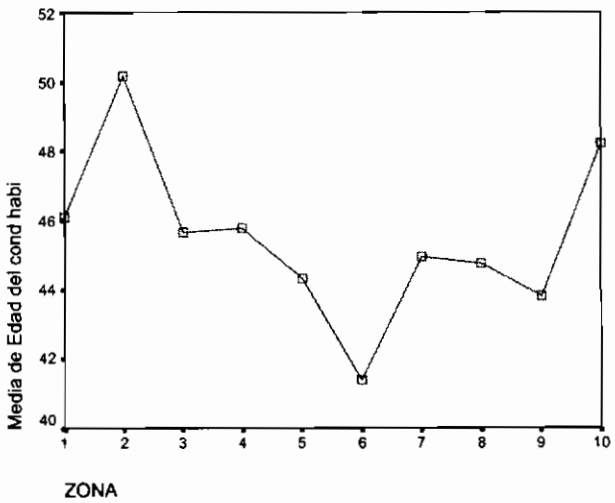
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	20	46.1089	9.8168	2.1951	41.5145	50.7033	31.47	68.38
2	25	50.1785	10.2805	2.0561	45.9349	54.4221	33.53	70.57
3	97	45.6908	10.4789	1.0640	43.5788	47.8028	22.27	72.55
4	47	45.8169	10.7518	1.5683	42.6601	48.9737	23.36	67.55
5	40	44.3216	11.0483	1.7469	40.7882	47.8551	28.10	76.97
6	69	41.4087	8.1039	.9756	39.4620	43.3555	27.99	61.27
7	109	44.9794	10.9761	1.0513	42.8955	47.0633	27.70	73.80
8	11	44.7841	8.6368	2.6041	38.9818	50.5863	31.78	61.72
9	14	43.8039	9.2318	2.4673	38.4736	49.1342	28.07	55.40
10	23	48.2106	14.7150	3.0683	41.8474	54.5738	25.10	73.19
Total	455	45.0760	10.6007	.4970	44.0994	46.0526	22.27	76.97

ANOVA

Edad del cond habi

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1936.026	9	215.114	1.950	.044
Intra-grupos	49081.838	445	110.296		
Total	51017.864	454			

Gráfico de las medias



Edad y tipo de vehículo

Descriptivos

Edad del cond habi

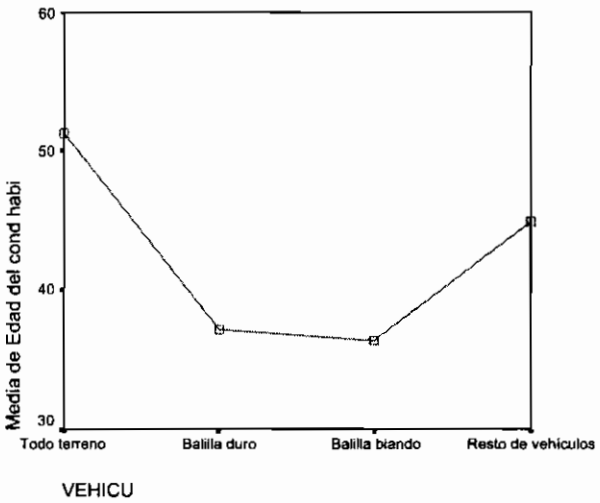
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Todo terreno	22	51.2682	9.7392	2.0764	46.9501	55.5863	36.59	68.59
Balilla duro	2	37.1521	5.2481	3.7110	-10.0001	84.3043	33.44	40.86
Balilla blando	6	36.3525	3.0558	1.2475	33.1456	39.5594	32.01	39.31
Resto de vehículos	425	44.9159	10.5860	.5135	43.9066	45.9252	22.27	76.97
Total	455	45.0760	10.6007	.4970	44.0994	46.0526	22.27	76.97

ANOVA

Edad del cond habi

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1436.631	3	478.877	4.356	.005
Intra-grupos	49581.233	451	109.936		
Total	51017.864	454			

Gráfico de las medias



Edad y forma de pago

Descriptivos

Edad del cond habi

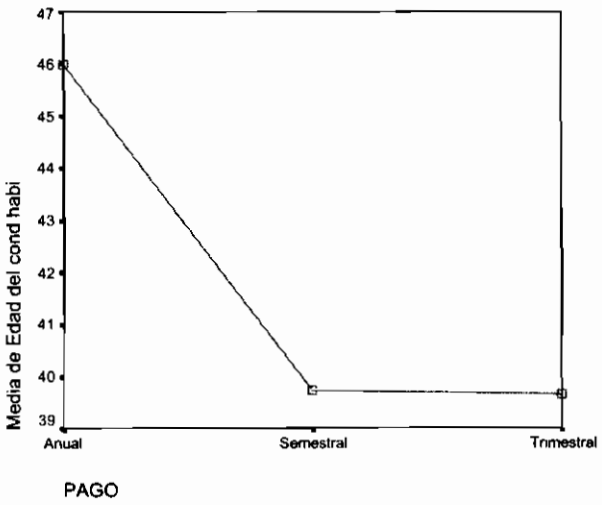
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	388	46.0042	10.5973	.5380	44.9464	47.0620	25.10	76.97
Semestral	34	39.7376	9.7928	1.6794	36.3207	43.1544	23.36	67.55
Trimestral	33	39.6629	8.1429	1.4175	36.7756	42.5503	22.27	53.12
Total	455	45.0760	10.6007	.4970	44.0994	46.0526	22.27	76.97

ANOVA

Edad del cond habi

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	2270.195	2	1135.098	10.525	.000
Intra-grupos	48747.668	452	107.849		
Total	51017.864	454			

Gráfico de las medias



Antigüedad carnet y sexo

Descriptivos

ANTICARN

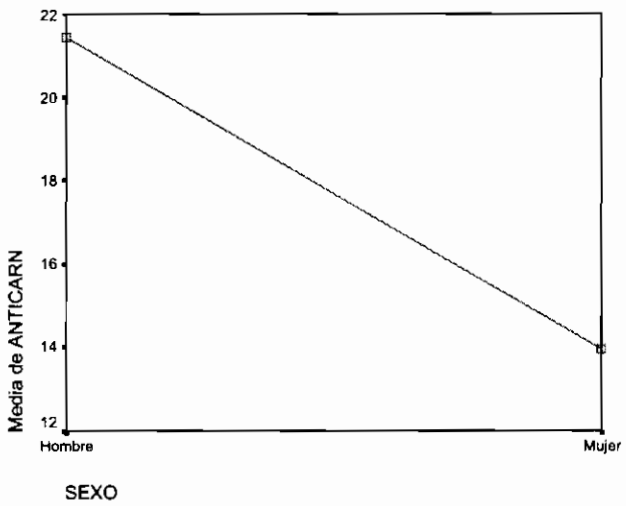
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Hombre	380	21.4282	8.0203	.4114	20.6192	22.2372	4.58	40.78
Mujer	75	13.9657	6.8988	.7966	12.3784	15.5529	2.75	35.92
Total	455	20.1981	8.3150	.3898	19.4321	20.9642	2.75	40.78

ANOVA

ANTICARN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	3488.260	1	3488.260	56.636	.000
Intra-grupos	27900.895	453	61.591		
Total	31389.154	454			

Gráfico de las medias



Antigüedad carnet y tipo de vehículo

Descriptivos

ANTICARN

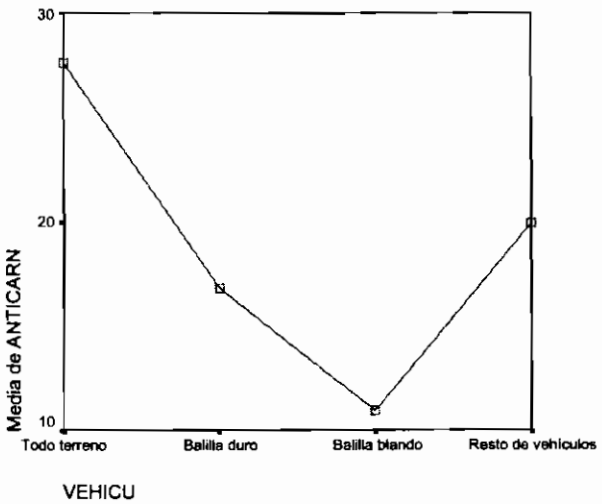
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Todo terreno	22	27.6121	6.4413	1.3733	24.7562	30.4680	15.99	37.86
Balilla duro	2	16.8452	7.3094	5.1685	-48.8267	82.5171	11.68	22.01
Balilla blando	6	10.9384	4.8864	1.9949	5.8104	16.0663	6.12	19.19
Resto de vehículos	425	19.9609	8.2094	.3982	19.1781	20.7436	2.75	40.78
Total	455	20.1981	8.3150	.3898	19.4321	20.9642	2.75	40.78

ANOVA

ANTICARN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1770.139	3	590.046	8.984	.000
Intra-grupos	29619.015	451	65.674		
Total	31389.154	454			

Gráfico de las medias



Antigüedad carnet y forma de pago

Descriptivos

ANTICARN

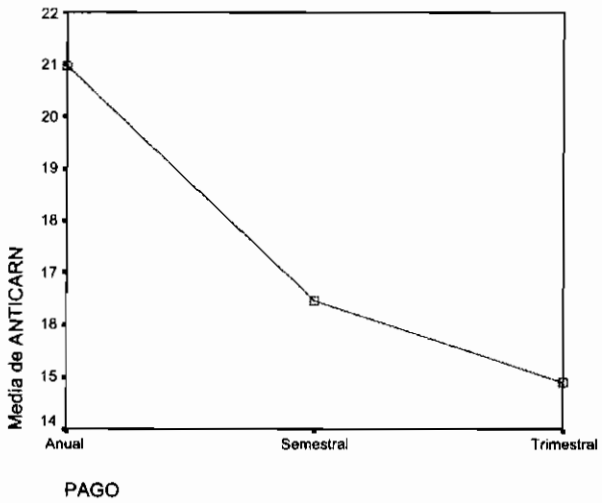
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	388	20.9745	8.1502	.4138	20.1610	21.7880	2.99	40.78
Semestral	34	16.4696	8.7648	1.5032	13.4114	19.5278	4.58	34.27
Trimestral	33	14.9117	6.8889	1.1992	12.4690	17.3545	2.75	27.50
Total	455	20.1981	8.3150	.3898	19.4321	20.9642	2.75	40.78

ANOVA

ANTICARN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1628.720	2	814.360	12.368	.000
Intra-grupos	29760.434	452	65.842		
Total	31389.154	454			

Gráfico de las medias



Malus y sexo

Descriptivos

MALUS

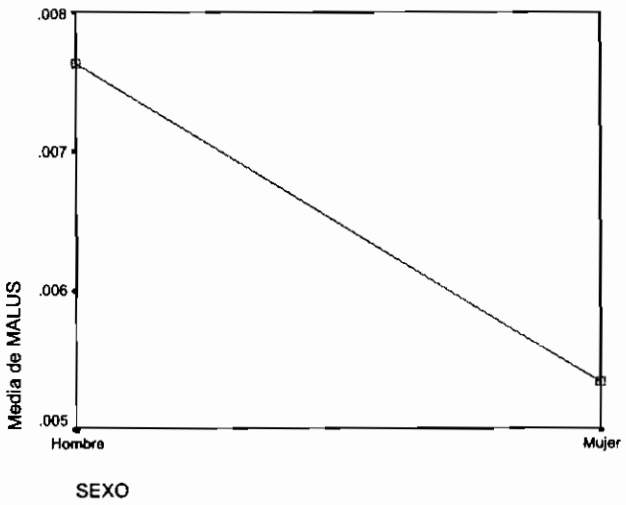
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Hombre	380	7.63E-03	5.4066E-02	2.774E-03	2.1781E-03	1.31E-02	.000	.700
Mujer	75	5.33E-03	3.6367E-02	4.199E-03	-3.03384E-03	1.37E-02	.000	.300
Total	455	7.25E-03	5.1542E-02	2.416E-03	2.5042E-03	1.20E-02	.000	.700

ANOVA

MALUS

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	3.308E-04	1	3.308E-04	.124	.725
Intra-grupos	1.206	453	2.662E-03		
Total	1.206	454			

Gráfico de las medias



Malus y zona

Descriptivos

MALUS

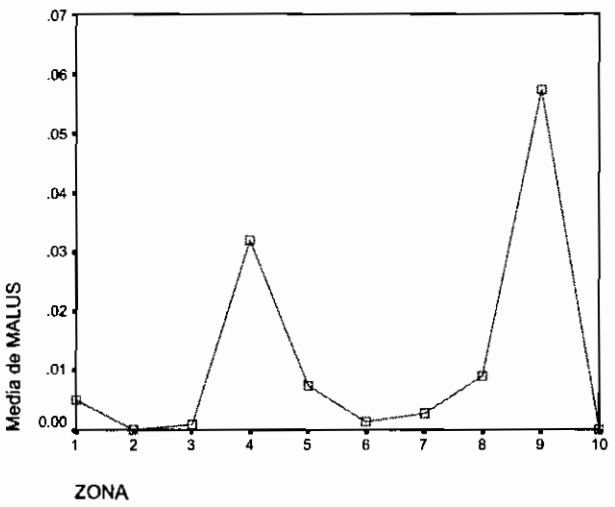
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	20	5.00E-03	2.2361E-02	5.000E-03	-5.46512E-03	1.55E-02	.000	.100
2	25	.00000	.00000	.00000	.00000	.00000	.000	.000
3	97	1.03E-03	1.0153E-02	1.031E-03	-1.01545E-03	3.08E-03	.000	.100
4	47	3.19E-02	.11629	1.696E-02	-2.23058E-03	6.61E-02	.000	.700
5	40	7.50E-03	4.7434E-02	7.500E-03	-7.67018E-03	2.27E-02	.000	.300
6	69	1.45E-03	1.2039E-02	1.449E-03	-1.44271E-03	4.34E-03	.000	.100
7	109	2.75E-03	2.8735E-02	2.752E-03	-2.70323E-03	8.21E-03	.000	.300
8	11	9.09E-03	3.0151E-02	9.091E-03	-1.11649E-02	2.93E-02	.000	.100
9	14	5.71E-02	.15046	4.021E-02	-2.97285E-02	.14401	.000	.500
10	23	.00000	.00000	.00000	.00000	.00000	.000	.000
Total	455	7.25E-03	5.1542E-02	2.416E-03	2.5042E-03	1.20E-02	.000	.700

ANOVA

MALUS

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	7.439E-02	9	8.265E-03	3.250	.001
Intra-grupos	1.132	445	2.543E-03		
Total	1.206	454			

Gráfico de las medias



Malus y uso del vehículo

Descriptivos

MALUS

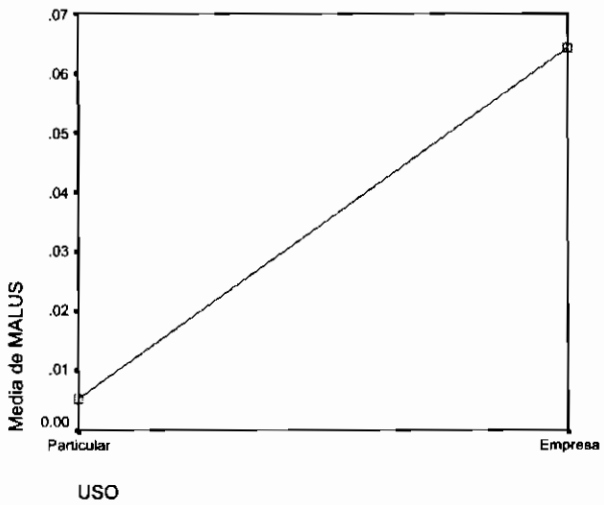
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Particular	441	5.44E-03	4.4388E-02	2.114E-03	1.2879E-03	9.60E-03	.000	.700
Empresa	14	6.43E-02	.14991	4.006E-02	-2.22688E-02	.15084	.000	.500
Total	455	7.25E-03	5.1542E-02	2.416E-03	2.5042E-03	1.20E-02	.000	.700

ANOVA

MALUS

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	4.698E-02	1	4.698E-02	18.363	.000
Intra-grupos	1.159	453	2.559E-03		
Total	1.206	454			

Gráfico de las medias



5.3. Aplicación 3. Datos de Baxter referentes al seguro del automóvil: cuantía de un siniestro para daños propios

Los datos relativos a esta aplicación son los descritos en el apartado 2.3.3 (tabla 2.2 del anexo 2.2), y hacen referencia al seguro del automóvil, concretamente a la cuantía por siniestro para la cobertura de daños propios. Estos datos suelen utilizarse para ilustrar el uso del MLG en el campo actuarial [Baxter, Coutts y Ross (1980); Hipp (2000); McCullagh y Nelder (1989)]. Con esta aplicación tan sólo pretendemos estudiar el comportamiento de los resultados del *bootstrap* del proceso de selección propuesto para la RBDP con datos agregados. El estudio podría completarse tratando el efecto de las interacciones de segundo y tercer orden.

5.3.1. Relaciones entre pares de variables

Disponemos de datos agregados con una respuesta continua, las medias de las cuantías de los 123 perfiles, y tres predictores cualitativos (grupo de vehículo, CG, antigüedad del vehículo, VA y antigüedad de la póliza, PA), para n = 8 902 siniestros.

Asociación de las cuantías con cada predictor

Calculamos (3.4),

- Y – PA: $\eta = 0.3564$
- Y – VA: $\eta = 0.5851$
- Y – CG: $\eta = 0.6441$

y observamos que el orden de asociación es CG, VA, PA, y que comparativamente PA es la menos asociada. En el anexo 5.8 encontramos las correspondientes anovas, de las que se deduce que las cuantías medias en las clases de los predictores son distintas. En los gráficos de medias se observa el sentido de las citadas diferencias. Respecto a los dos predictores cuantitativos discretizados, VA y PA, visualmente tenemos que a mayor antigüedad del vehículo (VA) y a mayor antigüedad de la póliza (PA), menores cuantías medias. Dada tal relación, podríamos decidir si utilizar o no tan sólo un coeficiente lineal por predictor en las regresiones.

Veamos ahora la relación de la respuesta con las variables de "interacción" entre predictores que construimos al igual que hicimos en la aplicación 1:

- Y – PA*VA: $\eta = 0.7081$
- Y – PA*CG: $\eta = 0.7771$
- Y – CG*VA: $\eta = 0.8140$
- Y – PA*VA*CG: $\eta = 1$

Observamos que la relación de la respuesta con la variable "interacción de tercer orden" es 1. Esto es lógico, ya que los datos de partida son datos agregados, y por tanto, la respuesta se corresponde con las medias de cada clase cruzada.

Observamos que la interacción de segundo orden de CG con VA es la más asociada con las cuantías, seguida de la interacción de CG con PA y por último la interacción de PA con VA, donde no interviene la variable CG.

Asociación predictor con predictor

Calculamos (3.6), (3.7), (3.8) y la correlación canónica r_1:

- PA – VA: $C = 0.0493,\ T = 0.0399,\ P = 0.0851,\ r_1 = 0.0731.$
- PA – CG: $C = 0.0826,\ T = 0.0668,\ P = 0.1416,\ r_1 = 0.1293.$
- CG – VA: $C = 0.1504,\ T = 0.1504,\ P = 0.2522,\ r_1 = 0.2440.$

Los predictores más asociados, según todas las mediadas calculadas, son el CG y el VA, aunque no en una medida excesiva. Este hecho nos va bien, pues como veremos, después de realizar los procesos de selección, las dos variables más significativas serán CG y VA. De este modo no correremos el riesgo de introducir información redundante en las regresiones.

5.3.2. Modelo lineal generalizado

Se han construido, en primer lugar, las variables binarias disjuntas sobre las que realizar el estudio, de

igual modo que procedimos en la aplicación 1: se han creado 7 para PA ,3 para VA, y 3 para CG.

Si se hubieran incluido en el estudio las interacciones deberíamos haber tenido en cuenta: 21 para PA*VA, 21 para PA*CG, 9 para CG*VA y 63 para PA*VA*CG. En total: 1 + 7 + 3 + 3 + 21 + 21 + 9 + 63 = 128 – 5 en relación a las celdas vacías = 123 variables binarias con los correspondientes parámetros. Tendríamos con todas ellas un modelo saturado, con el mismo número de parámetros que de observaciones.

Estos datos suelen utilizarse para ilustrar el uso del MLG tal y como ya hemos indicado. En Hipp (2000) pp. 1-31 y McCullagh y Nelder (1989) pp. 296-300 encontramos un estudio haciendo uso de la distribución Gamma, $\zeta = 2$, con el enlace inverso, $\lambda = -1$, que en el caso de la Gamma es el correspondiente enlace canónico. Se analiza la importancia de los factores y de todas sus interacciones, llegando a la conclusión de que el modelo correcto es el que tan sólo incluye los efectos principales. Encontramos en estos trabajos con detalle la tabla de desvianzas y los coeficientes resultantes del predictor lineal. Al tratarse del enlace inverso, $\eta_i = \mu_i^{-1}$, se interpreta que un coeficiente positivo (o grande) en el predictor lineal lleva a una reducción (grande) de la media de cuantía base.

Adicionalmente, McCullagh y Nelder (1989) utilizan estos datos para ilustrar en su libro el tratamiento de las familias paramétricas de distribuciones, (3.65), y enlaces, (3.26), caracterizadas respectivamente por:

$$V(\mu_i) = \mu_i^{\zeta}$$

y

$$\eta_i = g(\mu_i) = \begin{cases} \mu_i^{\lambda} & \text{para} \quad \lambda \neq 0 \\ \log(\mu_i) & \text{para} \quad \lambda = 0 \end{cases}.$$

Realizan tres análisis específicos:

a) En pp. 377, fijada la Gamma, $V(\mu_i) = \mu_i^2$, y los tres efectos principales, varían la λ implicada en enlace paramétrico, $\eta_i = \mu_i^{\lambda}$, con el fin de analizar los valores que proporcionan una menor desvianza. Reproducimos el gráfico con las desvianzas en función de λ:

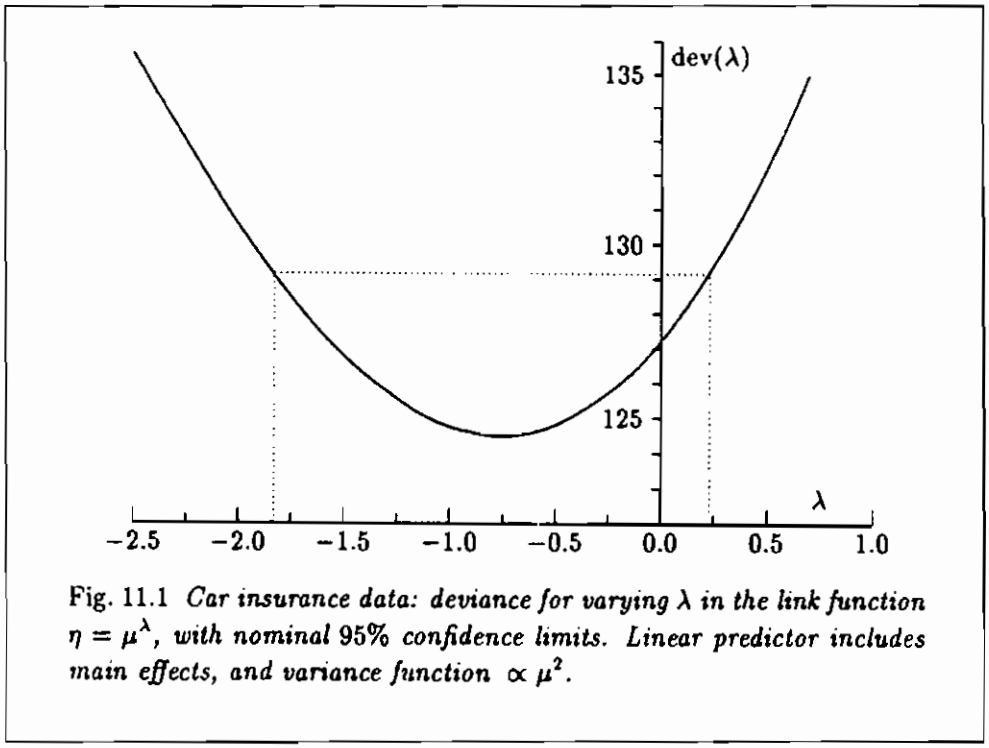


Fig. 11.1 *Car insurance data: deviance for varying λ in the link function* $\eta = \mu^{\lambda}$, *with nominal 95% confidence limits. Linear predictor includes main effects, and variance function* $\propto \mu^2$.

Figura 5.4. Reproducción de la figura 11.1 de McCullagh y Nelder (1989).

Obtienen que, para un nivel de confianza del 95%, λ puede ser variado entre –1.8 y 0.25 aproximadamente, observándose un mínimo hacia –0.8. Dentro de este rango encontramos el enlace canónico y el logarítmico. Nosotros calculamos la desvianza para algunas λ:

- $\lambda = 1$ (enlace identidad, aditivo): 139.900486
- $\lambda = \log$ (enlace logarítmico, multiplicativo): 127.059307
- $\lambda = -0.8$ (aproximadamente el mínimo): 123.843973
- $\lambda = -1$ (enlace inverso, canónico): 123.961177
- $\lambda = -1.5$ (otro valor dentro del rango): 125.687813

obteniendo unos resultados en concordancia con el gráfico.

b) En pp. 400, haciendo uso de la versión extendida de la cuasi-desvianza, fijado el enlace inverso, $\eta_i = \mu_i^{-1}$, y los tres efectos principales, buscan el parámetro ζ de la función de varianza $V(\mu_i) = \mu_i^{\zeta}$ que ofrezca un mínimo en la desvianza. Reproducimos el gráfico con las desvianzas en función de ζ:

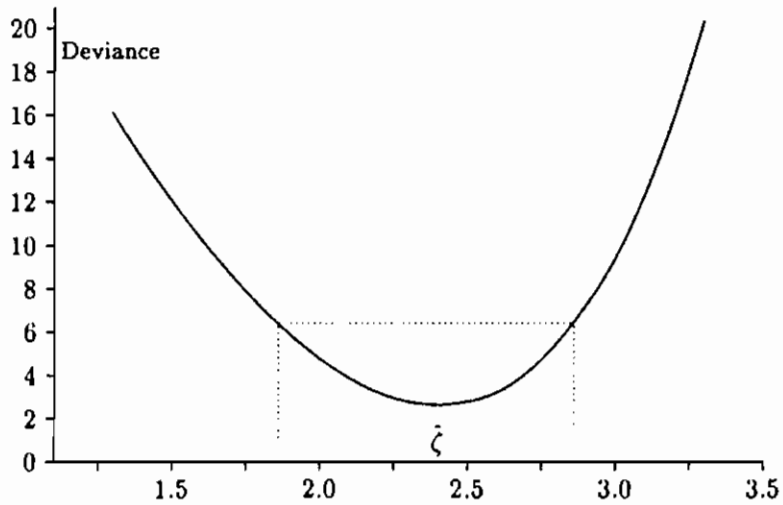


Fig. 12.6. *The profile extended quasi-likelihood curve plotted against ζ for the power-family of variance functions applied to the car-insurance data.*

Figura 5.5. Reproducción de la figura 12.6 de McCullagh y Nelder (1989).

Obtienen que, para un nivel de confianza del 95%, el parámetro puede variar entre 1.87 y 2.85, alcanzando un mínimo hacia el 2.4. Este rango incluye a la distribución Gamma, $\zeta = 2$, utilizada inicialmente.

c) En pp. 413-414, buscan conjuntamente la combinación (ζ, λ) que ofrezca un mínimo en la versión extendida de la cuasi-desvianza, el cual se alcanza para $\zeta = 2.4$ y $\lambda = 0.75$. Reproducimos el gráfico que ofrece diferentes contornos variando los niveles de confianza:

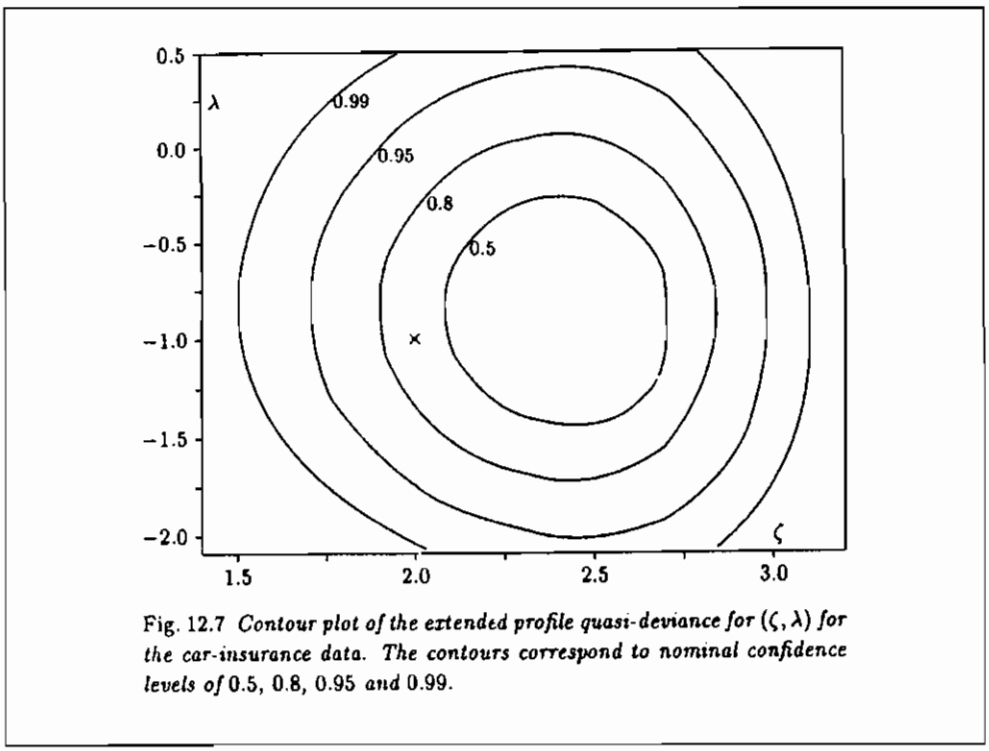


Fig. 12.7 *Contour plot of the extended profile quasi-deviance for* (ζ, λ) *for the car-insurance data. The contours correspond to nominal confidence levels of* 0.5, 0.8, 0.95 *and* 0.99.

Figura 5.6. Reproducción de la figura 12.7 de McCullagh y Nelder (1989).

Observamos que la combinación inicial, $(2,-1)$, cae dentro del contorno del 95%. También cae dentro de este nivel la combinación $(2,0)$ que se corresponde con el enlace logarítmico, aunque es una combinación menos buena.

Respecto a la selección de predictores, puesto que el detalle de los resultados de la Gamma con el enlace inverso lo encontramos en las referencias citadas, aquí hemos realizado los procesos para la distribución Gamma con los enlaces Logarítmico e Identidad, los cuales nos ofrecen una tarifa multiplicativa y aditiva respectivamente, y adicionalmente para el clásico Normal, $\zeta = 0$, Identidad, $\lambda = 1$. Éste último lo realizamos con el objetivo de utilizar sus resultados en el siguiente apartado para

la RBDP. En cualquier caso tan sólo incluimos los efectos principales.

Los procesos de selección se encuentran detallados en el anexo 5.9. En él encontramos los cálculos realizados utilizando tanto la distribución asintótica *ji*-cuadrado, (3.40), como la distribución asintótica *F*, (3.41). Puesto que en el resto de aplicaciones siempre hacemos uso de la *F*, en esta aplicación utilizamos ambas para corroborar que los resultados obtenidos respecto al conjunto de predictores seleccionados y respecto al orden de entrada y salida de los predictores en los procesos, dada una distribución del error y una función de enlace, son los mismos independientemente de la distribución asintótica utilizada.

Recogemos a continuación los resultados de los procesos para un nivel de significación $\alpha^* = 0.05$:

- Para la distribución Gamma, tanto con el enlace logarítmico como para el identidad:
  - F(1) = Antigüedad del vehículo,VA
  - F(2) = Grupo de vehículo, CG
  - F(3) = Antigüedad de la póliza, PA
- Para la distribución Normal con el enlace identidad:
  - F(1) = Grupo de vehículo, CG
  - F(2) = Antigüedad del vehículo,VA
  - F(3) = Antigüedad de la póliza, PA

Sin ser exhaustivos, validamos los modelos tratados en este apartado incluyendo sólo los efectos principales. Plasmamos en la siguiente tabla el *p*-valor resultante del estadístico (3.45) y el *pseudo* R^2 (3.46):

Modelo	**Poder predictivo**
$\zeta = 2$ y $\lambda = -1.5$	$R^2 = 0.7993$, *p*-valor = 4.9046E-32
$\zeta = 2$ y $\lambda = -1$	$R^2 = 0.8020$, *p*-valor = 2.3509E-32
$\zeta = 2$ y $\lambda = -0.8$	$R^2 = 0.8022$, ***p*-valor = 2.2356E-32**
$\zeta = 2$ y $\lambda = \log$	$R^2 = 0.7971$, *p*-valor = 8.7313E-32
$\zeta = 2$ y $\lambda = 1$	$R^2 = 0.7766$, *p*-valor = 1.4424E-29
$\zeta = 0$ y $\lambda = 1$	$R^2 = 0.7565$, *p*-valor = 1.3684E-27

Como era de esperar, el "mejor" modelo fijado $\zeta = 2$ es para $\lambda = -0.8$, y el "peor" para la identidad, $\lambda = 1$.

Para finalizar plasmamos los coeficientes estimados del predictor lineal para los modelos $(\zeta = 2, \lambda = 0)$ y $(\zeta = 2, \lambda = -0.8)$:

Términos	**Coeficientes para** $(2,0)$	**Coeficientes para** $(2,-0.8)$
Constante[27]	5.149358	0.018343
VA de 0-3	0.687540	–0.009021
VA de 4-7	0.599938	–0.008126
VA de 8-9	0.344454	–0.005171
CG para A	–0.401286	0.003532
CG para B	–0.400316	0.003607
CG para C	–0.244143	0.002028
PA de 17-20	0.223188	–0.002208
PA de 21-24	0.225143	–0.002023
PA de 25-29	0.153541	–0.001348
PA de 30-34	0.109888	–0.001082
PA de 35-39	–0.091584	0.001075
PA de 40-49	–0.016853	0.000152
PA de 50-59	0.002817	0.000010

Observamos la coherencia cuantitativa de los coeficientes: un coeficiente grande y positivo para el enlace logarítmico se corresponde con un coeficiente negativo y pequeño para el enlace –0.8, y uno negativo y pequeño para el logarítmico y con uno positivo y grande para el enlace –0.8, ya que la relación entre la estimación de las cuantías y el predictor lineal es $\log(\hat{\mu}_i) = \hat{\eta}_i$ y $(\hat{\mu}_i)^{-0.8} = \hat{\eta}_i$ en cada

[27] Para el término constante hemos utilizado a VA de 10 & over, a CG para D, y a PA de 60 & over.

caso, y por lo tanto $\hat{\mu}_i = \exp(\hat{\eta}_i)$ y $\hat{\mu}_i = \left(\frac{1}{\hat{\eta}_i}\right)^{1.25}$.

5.3.3. Regresión basada en distancias

Se ha utilizado como función de distancias entre individuos el coeficiente de coincidencias (4.7). Tal y como se detalla en el apartado 4.2.3 de casos particulares de la RBD (caso 3): si las variables explicativas son categóricas y el coeficiente de similaridad utilizado es el de coincidencias, la predicción obtenida mediante RBD y regresión múltiple clásica utilizando tantas variables binarias como clases, coinciden.

Al igual que hicimos en la aplicación 1, vamos a utilizar este resultado como pauta para corroborar los resultados obtenidos con el proceso de selección, en este caso ponderado. Recordemos que en el apartado anterior acabamos de realizar el estudio para el MLG con distribución del error Normal y enlace identidad, el cual coincide también con el modelo de regresión clásico.

Se han generado dos bloques de muestras, cada una con 500 muestras: $B = B1 + B2 = 500+500 = 1000$ simulaciones. En las siguientes tablas se detalla el resultado separado, *B1* y *B2*, y global, *B*, para observar la convergencia en la estimación del *p*-valor:

- ***B1 y B2:***

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
PA	0.040 y 0.042	0.050 y 0.052	**0.095 y 0.099**
VA	0.019 y 0.019	**0.028 y 0.028**	--------
CG	**0.003 y 0.007**	--------	--------
	F(1) = CG	F(2) = VA	F(3) = PA

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
PA	**0.095 y 0.099**	--------	--------
VA	0.028 y 0.030	**0.028 y 0.028**	--------
CG	0.013 y 0.017	0.018 y 0.020	**0.003 y 0.007**
	F[1] = PA	F[2] = VA	F[3] = CG

➢ ***B:***

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
PA	0.041	0.051	**0.097**
VA	0.019	**0.028**	--------
CG	**0.005**	--------	--------
	F(1) = CG	F(2) = VA	F(3) = PA

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
PA	**0.097**	--------	--------
VA	0.029	**0.028**	--------
CG	0.015	0.019	**0.005**
	F[1] = PA	F[2] = VA	F[3] = CG

Los resultados coinciden para *B1* y B2, al menos con dos decimales. Con *B2* se obtienen, en general, unos *p*-valores algo más elevados que para *B1*, pero vemos como los resultados son coherentes. Notamos que estamos simulando la distribución exacta del estadístico (4.57), y por lo tanto el resultado obtenido es tan sólo una estimación.

El conjunto resultante de la selección es el mismo que para la distribución Normal con el enlace identidad, como era de esperar:

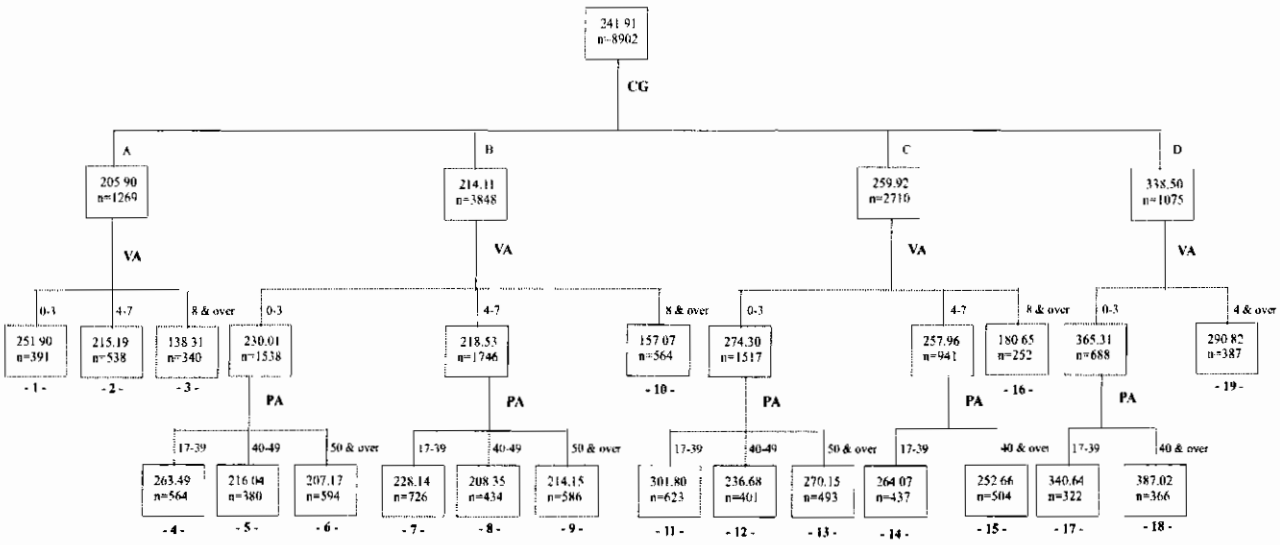
- F(1) = Grupo de vehículo, CG
- F(2) = Antigüedad del vehículo,VA
- F(3) = Antigüedad de la póliza, PA

5.3.4. Análisis de segmentación

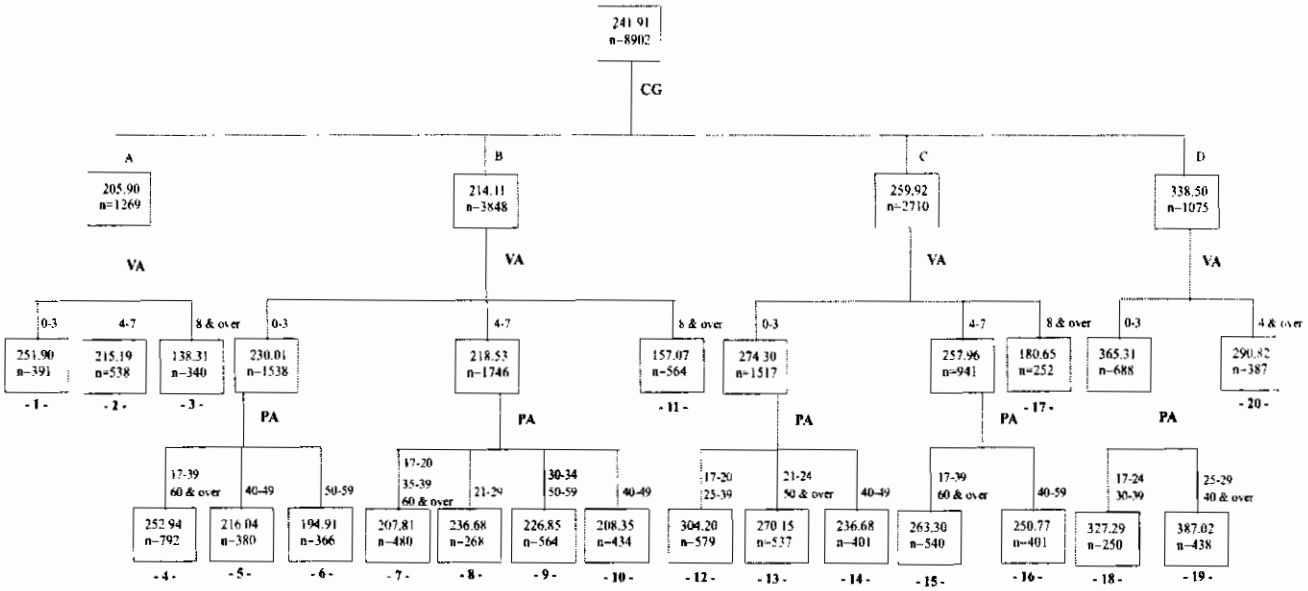
Hemos utilizado el algoritmo SPSS CHAID ordinal. Para ello hemos discretizado las cuantías en 31 clases haciendo uso de método de Ward, y hemos asignado como puntuaciones las medias correspondientes. Se ha fijado un nivel de significación tanto para la fase de agrupación como para la de selección del mejor predictor de 0.05. El tamaño mínimo para analizar un nodo de 500 y el mínimo para segmentar de 250. En el árbol 1 se ha definido al CG como libre y al VA y PA como monótonos. Y en el árbol 2 se han definido los tres predictores como libres.

A continuación plasmamos los árboles resultantes. En ellos se detallan las medias de las cuantías de la variable discretizada y el número de siniestros de cada nodo:

Árbol 1



Árbol 2



Observamos que la única diferencia es que el predictor PA, en el tercer nivel, ha agrupado sus clases en el árbol 2 de manera no monótona. Si observamos los gráficos de medias de las anovas del anexo 5.8, ya intuimos que la relación del PA con las cuantías medias, no es tan lineal como la del VA, al menos en algunas de sus clases.

ANEXO 5.8. Anovas de los datos de Baxter

Car Group (CG)

Descriptivos

mean claim amount

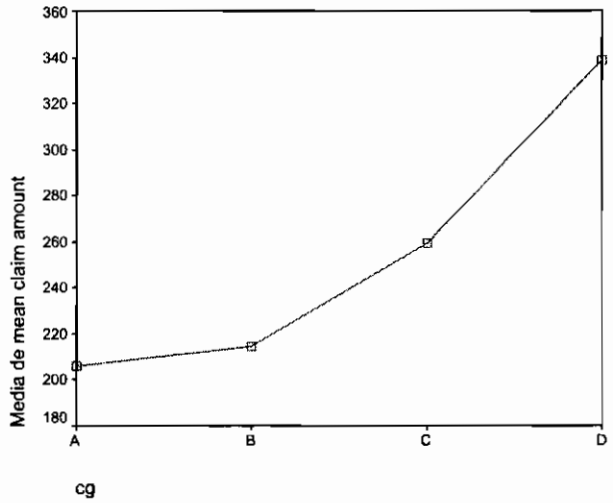
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
A	1269	206.00	56.178	1.577	202.91	209.10	98	302
B	3848	214.39	40.894	.659	213.10	215.68	11	420
C	2710	259.54	42.212	.811	257.95	261.13	104	343
D	1075	338.34	77.448	2.362	333.70	342.97	65	850
Total	8902	241.90	64.554	.684	240.56	243.25	11	850

ANOVA

mean claim amount

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	15388109	3	5129369.6	2102.846	.000
Intra-grupos	21704457	8898	2439.251		
Total	37092566	8901			

Gráfico de las medias



Vehicle Age (VA)

Descriptivos

mean claim amount

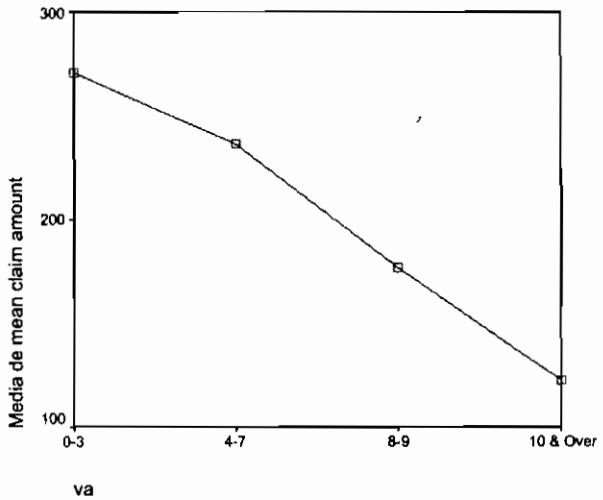
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
0-3	4134	270.96	64.602	1.005	268.99	272.93	189	763
4-7	3549	236.42	40.147	.674	235.10	237.75	129	850
8-9	822	176.95	31.978	1.115	174.76	179.14	116	346
10 & Over	397	122.79	38.547	1.935	118.99	126.60	11	636
Total	8902	241.90	64.554	.684	240.56	243.25	11	850

ANOVA

mean claim amount

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	12697396	3	4232465.2	1543.768	.000
Intra-grupos	24395170	8898	2741.646		
Total	37092566	8901			

Gráfico de las medias



Policy Holder Age (PA)

Descriptivos

mean claim amount

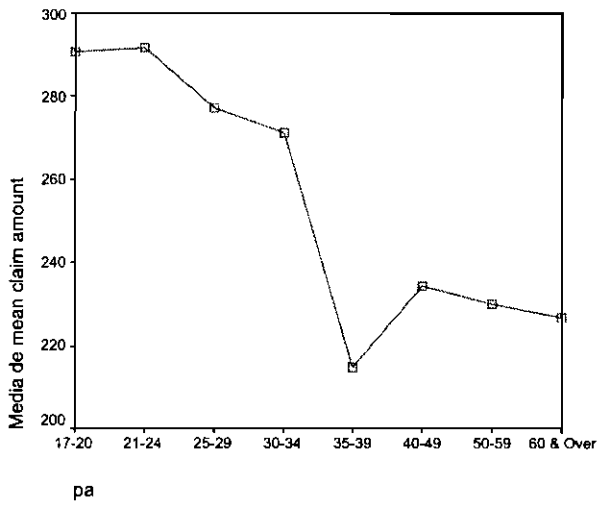
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
17-20	89	290.61	139.427	14.779	261.24	319.98	11	850
21-24	370	291.60	83.040	4.317	283.11	300.09	104	420
25-29	930	277.07	57.237	1.877	273.39	280.76	110	636
30-34	1101	271.12	71.542	2.156	266.89	275.35	65	400
35-39	1177	215.02	41.503	1.210	212.65	217.40	113	325
40-49	2238	234.45	59.758	1.263	231.98	236.93	98	387
50-59	1791	230.21	55.924	1.321	227.62	232.80	98	391
60 & Over	1206	226.70	56.441	1.625	223.51	229.89	101	385
Total	8902	241.90	64.554	.684	240.56	243.25	11	850

ANOVA

mean claim amount

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	4713019.9	7	673288.562	184.939	.000
Intra-grupos	32379546	8894	3640.606		
Total	37092566	8901			

Gráfico de las medias



ANEXO 5.9. Procesos de selección para los MLG

- Utilizando la distribución asintótica *F* de Fisher (3.41):

Gamma | Log:

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
PA	0.031282	0.000995	**1.0237E-15**
VA	**4.7132E-15**	--------	--------
CG	4.1343E-12	**4.2037E-15**	--------
	F(1) = VA	F(2) = CG	F(3) = PA

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
PA	**1.0263E-7**	--------	--------
VA	1.0574E-21	5.5815E-18	**4.7131E-15**
CG	7.9603E-19	**4.2037E-15**	--------
	F[1] = PA	F[2] = CG	F[3] = VA

Gamma | Identidad:

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
PA	0.031282	0.003859	**2.7203E-6**
VA	**4.7132E-15**	--------	--------
CG	4.1343E-12	**2.3318E-14**	--------
	F(1) = VA	F(2) = CG	F(3) = PA

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
PA	**2.7203E-6**	--------	--------
VA	1.9933E-19	3.1079E-17	**4.7132E-15**
CG	2.9209E-17	**2.3318E-14**	--------
	F[1] = PA	F[2] = CG	F[3] = VA

Normal | Identidad:

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
PA	0.025475	0.00058	**4.0057E-7**
VA	7.7339E-11	**1.9019E-12**	--------
CG	**8.0994E-14**	--------	--------
	F(1) = CG	F(2) = VA	F(3) = PA

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
PA	**4.0057E-7**	--------	--------
VA	2.0848E-15	**1.9019E-12**	--------
CG	1.2738E-18	2.3366E-15	**8.0994E-14**
	F[1] = PA	F[2] = VA	F[3] = CG

➢ Utilizando la distribución asintótica *ji*-cuadrado (3.40):

Gamma | Log:

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
PA	0.020684	0.001459	**1.1436E-9**
VA	**1.2260E-17**	--------	--------
CG	1.2892E-16	**6.4234E-19**	--------
	F(1) = VA	F(2) = CG	F(3) = PA

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
PA	**1.1436E-9**	--------	--------
VA	1.0031E-35	5.0547E-24	**1.2260E-17**
CG	7.0358E-29	**6.4234E-19**	--------
	F[1] = PA	F[2] = CG	F[3] = VA

Gamma | Identidad:

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
PA	0.020684	0.005382	**6.0249E-8**
VA	**1.2260E-17**	--------	--------
CG	1.2892E-16	**4.2611E-18**	--------
	F(1) = VA	F(2) = CG	F(3) = PA

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
PA	**6.0249E-8**	--------	--------
VA	1.6642E-31	3.6618E-23	**1.2260E-17**
CG	1.4778E-26	**4.2611E-18**	--------
	F[1] = PA	F[2] = CG	F[3] = VA

Normal | Identidad:

Proceso de introducción progresiva:

Variable	*p*-valor	*p*-valor | F(1)	*p*-valor | F(1)F(2)
PA	0.019159	0.000206	**1.0271E-8**
VA	2.2651E-13	**5.3092E-16**	--------
CG	**3.5447E-18**	--------	--------
	F(1) = CG	F(2) = VA	F(3) = PA

Proceso de eliminación progresiva:

Variable	*p*-valor	*p*-valor / F[1]	*p*-valor / F[1]F[2]
PA	**1.0271E-8**	--------	--------
VA	1.2998E-21	**5.3092E-16**	--------
CG	3.1539E-28	4.5997E-21	**3.5447E-18**
	F[1] = PA	F[2] = VA	F[3] = CG

5.4. Aplicación 4. Cartera C2 de responsabilidad civil de automóviles: número de siniestros para daños materiales

Utilizamos los datos descritos en el apartado 2.3.2 que hacen referencia a la cartera C2 de responsabilidad civil del automóvil. Analizamos el número de siniestros para daños materiales. Son datos desagregados, y los factores de riesgo están todos inicialmente discretizados.

Para escoger una modalidad realizamos la tabla de contingencia que cruza la modalidad con el número de siniestros de las pólizas que han estado vivas durante todo el período:

Recuento

		NÚMERO DE SINIESTROS						
		0	1	2	3	4	5	Total
Modalidad	Terceros	6299	441	50	10	2		6802
	Terceros + Rotura lunas	13771	1041	138	41	5	1	14997
	Terceros + Rotura de lunas + Incendios	3951	280	37	15			4283
	Todo riesgo	1757	169	31	3	1	1	1962
	Todo riesgo con franquicia	2834	242	32	12	2		3122
	Otra	351	30	2	2			385
Total		28963	2203	290	83	10	2	31551

Estudiamos la modalidad de terceros en la que tenemos 6 802 pólizas.

A modo informativo, la media y la varianza del número de siniestros de estas pólizas son:

$$E[N]=0.0851 \text{ y } Var(N)=0.105\text{, así, } V(N)=1.16\cdot E[N].$$

5.4.1. Agrupación de zonas

Realizamos una agrupación de zonas, ya que la agrupación inicial queda demasiado dispersa para estos datos. En la siguiente tabla detallamos las zonas iniciales (en total 31), el número de pólizas de cada zona, la media del número de siniestros en cada zona y la correspondiente desviación típica:

NUMSIN

	N	Media	Desviación típica
1	9	.00	.000
2	37	.14	.419
3	20	.05	.224
4	70	.09	.408
5	15	.07	.258
6	146	.07	.279
8	229	.06	.267
9	230	.14	.405
10	313	.07	.261
11	166	.05	.297
12	662	.10	.349
13	388	.09	.370
14	138	.08	.402
15	113	.05	.262
16	431	.06	.244
17	88	.05	.209
18	146	.11	.355
19	194	.07	.251
21	310	.10	.345
22	1793	.08	.304
23	1	.00	.
24	558	.07	.317
25	107	.10	.335
26	40	.15	.427
27	80	.09	.284
28	463	.13	.410
29	21	.10	.436
30	23	.04	.209
31	11	.00	.000
Total	6802	.09	.324

Agrupamos las zonas utilizando el algoritmo ordinal de SPSS. Definimos a la zona como predictor libre. Exigimos un mínimo de 450 pólizas por zona construida, y un nivel de significación para la fase de agrupación de categorías de 0.5 (no se ha restringido más, ya que al exigirle no encontraba diferencias en las clases). Con esta agrupación poco estricta permitiremos, en la aplicación de la segmentación conjunta del apartado 5.4.7, que sea posible distinguir dentro de las zonas mediante otros factores.

Las 5 zonas resultantes provienen de las siguientes agrupaciones de zonas iniciales:

- *Zona 1*: 1,3,8,11,15,16,17,23,30,31
- *Zona 2*: 2,9,26,28
- *Zona 3*: 4,13,14,22,27
- *Zona 4*: 5,6,10,19,24
- *Zona5*: 12,18,21,25,29.

Hemos pasado de una asociación de $\eta = 0.005$ con las 31 zonas iniciales y un p-valor en la anova de 0.140, a una asociación de $\eta = 0.071$ con las 5 zonas resultantes de la segmentación y un p-valor en la anova de 0 (anexo 5.10). De lo que deducimos que la agrupación realizada es más correcta para incluir en el estudio que la inicial.

5.4.2. Marcas de clase para los factores cuantitativos

Detallamos las marcas de clase que hemos utilizado para el cálculo de las relaciones entre pares de variables del siguiente apartado:

- Número de plazas: 2, 3, 4, 5, 6, 7, 8 y 9. Hemos utilizado directamente los valores discretos que toma la variable.
- Nivel de bonus-malus: -50, -40, -20, -10, 0, 10, 20, 30, 35, 40, 45 y 50. También hemos utilizado los valores discretos que toma la variable.
- Antigüedad de la póliza. Ésta está tomada de año en año excepto para el último intervalo, al cual le asignamos el valor 8.5:

[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	**[7,...)**
0.5	1.5	2.5	3.5	4.5	5.5	6.5	**8.5**

- Edades. Éstas están tomadas aproximadamente en intervalos de 5 años, por lo que no podremos realizar intervalos de menor amplitud en el estudio. Cuando analizamos la edad del primer conductor observamos que no hay ningún asegurado en el intervalo [18,20], sin embargo, en la edad del segundo conductor, en la cual tenemos 6 091 valores *missing*, encontramos que en el intervalo de [18,20] aparecen 46 pólizas. Tal y como justificaremos en el apartado 5.4.4 es conveniente construir la variable "edad de máximo riesgo", definida como el mínimo de ambas edades en el caso de existencia de segundo conductor. Por lo que, en lo que sigue, trabajaremos con la edad del primer conductor y con la edad de máximo riesgo. Respecto a las marcas de clase utilizamos los puntos medios, excepto para el intervalo superior al cual le asignamos el valor 81:

[18,20]	[21,25]	[26,30]	[31,35]	[36,40]	[41,45]	[46,50]	[51,55]	[56,60]	[61,65]	[66,70]	[71,75]	[76,80]	**[81,99]**
19	23	28	33	38	43	48	53	58	63	68	73	78	**81**

- Antigüedad del carnet. Ésta, al igual que la antigüedad de la póliza, está tomada de año en año excepto para el último intervalo, al cual asignamos el valor 15. Observamos que en el primer intervalo [0,1) no hay pólizas, puesto que tampoco tenemos asegurados de menos de 21 años.

[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)	**[10,...)**
0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	**15**

- Antigüedad del vehículo. Ésta, al igual que las otras dos antigüedades, está tomada de año en año excepto para el último intervalo, al cual le asignamos el valor 14:

[0,1)	[1,2)	[2,3)	[3,4)	[4,5)	[5,6)	[6,7)	[7,8)	[8,9)	[9,10)	[10,11)	**[11,...)**
0.5	1.5	2.5	3.5	4.5	5.5	6.5	7.5	8.5	9.5	10.5	**14**

- Potencia. Ésta está tomada en intervalos de diferentes amplitudes, seguramente porque no es necesario afinar mucho más. Al primer intervalo le asignamos el valor 25, y al último el valor 220:

[...,28]	[29,33]	[34,42]	[43,53]	[54,75]	[76,94]	[95,118]	[119,215]	**[216,...]**
25	31	38	48	64.5	85	106.5	167	**220**

5.4.3. Relaciones entre pares de variables

En este apartado, previo a la agregación de los datos que se realizará en el apartado 5.4.4, vamos a estudiar las relaciones entre el número de siniestros y los siguientes factores:

- Factores cuantitativos (utilizando los valores del apartado anterior): número de plazas, nivel de bonus-malus, antigüedad de la póliza, edad del primer conductor, edad de máximo riesgo (calculada como el mínimo de la edad del primer y segundo conductor), antigüedad del carnet del primer conductor, antigüedad del vehículo y potencia del vehículo.
- Factores cualitatitvos: Forma de pago, sexo y zona (utilizando la agrupación en 5 zonas realizada en el apartado 5.4.1)

Adicionalmente estudiamos las relaciones entre tales factores distinguiendo su naturaleza, cuantitativa o cualitativa.

5.4.3.1. Número de siniestros con factores

Medidas de asociación del número de siniestros (discreto cuantitativo) con cada factor potencial uno a uno:

Asociación del número de siniestros con cada factor cuantitativo

Calculamos (3.2),

Variables	ρ
N – Plazas	**0.032**
N – Bonus-malus	**–0.205**
N – Antipol	–0.016
N – Edad1	0.002
N – Edadr	–0.015
N – Anticarn	–0.027
N – Antivehi	–0.029
N – Potencia	**0.035**

Cabe destarcar:

- La más asociada con el número de siniestros es el nivel de bonus-malus, con una correlación negativa, como es de esperar, pues a mejor escala menor número de siniestros. Esto implica que el sistema bonus-malus aplicado está funcionando bien. Hay que tener en cuenta que el período de observación de un año de esta cartera es anterior al fichero histórico de siniestralidad de conductores. Posiblemente con datos actuales la correlación sería mucho mayor.
- La siguiente es la potencia, a mayor potencia mayor número de siniestros.
- La sigue el número de plazas con una correlación positiva, a mayor número de plazas mayor número de siniestros. Y la correlación sería mayor si no fuera por los de 2 y 3 plazas, que son un caso especial, y que luego se contabilizaran a parte.

En un segundo orden tenemos las siguientes asociaciones:

- La antigüedad del vehículo con una correlación negativa: a mayor antigüedad del vehículo menor número de siniestros.

- La antigüedad del carnet con una correlación negativa: a mayor antigüedad del carnet menor número de siniestros.

Y para terminar:

- La antigüedad de la póliza con correlación negativa: a mayor antigüedad de la póliza menor número de siniestros.
- La edad del primer conductor con una correlación positiva y pequeña, 0.002, que pasa a una correlación algo mayor y negativa, –0.015, con la edad de máximo riesgo, en la que a menor edad mayor número de siniestros.

Asociación del número de siniestros con cada factor cualitativo

Estudiamos la asociación del número con cada predictor cualitativo haciendo uso de (3.4), adicionalmente detallamos el *p*-valor del anova correspondiente (anexo 5.10), y obtenemos:

Variables	η	*p*-valor
N – Pago	**0.053**	**0**
N – Sexo	0	0.984
N – Zona	**0.071**	**0**

De las tres variables cualitativas, las dos más asociadas con el número de siniestros son la zona seguida de la forma de pago. El sexo parece no tener relación.

En el gráfico de medias de la forma de pago, que encontramos en el anexo 5.10, observemos un número medio de siniestros mayor a mayor fraccionamiento en el pago.

5.4.3.2. Factor con factor

Cualitativo con cualitativo

Calculamos sólo el coeficiente de contingencia de Pearson (3.8),

P =	Pago	Sexo	Zona
Pago	1	0.030	**0.120**
Sexo	------	1	0.047
Zona	------	------	1

y obtenemos que las más asociadas son la zona con la forma de pago.

Cuantitativo con cuantitativo

Calculamos (3.2),

$\rho =$	Plazas	Bonus-malus	Antipol	Edad1	Edadr	Anticarn	Antivehi	Potencia
Plazas	1	–0.006	–0.012	–0.026	–0.012	0.013	–0.006	0.040
Bonus-malus	------	1	**0.123**	**0.082**	**0.068**	**0.123**	**0.118**	**–0.065**
Antipol	------	------	1	**0.216**	**0.177**	**0.116**	**0.043**	0
Edad1	------	------	------	1	**0.837**	**0.255**	**0.047**	–0.016
Edadr	------	------	------	------	1	**0.211**	0.036	0.004
Anticarn	------	------	------	------	------	1	0.006	0.033
Antivehi	------	------	------	------	------	------	1	**–0.167**
Potencia	------	------	------	------	------	------	------	1

Destacamos las siguientes correlaciones:

- A mayor nivel de bonus-malus mayor antigüedad de la póliza, mayor edad del conductor, mayor antigüedad del carnet del primer conductor, mayor antigüedad del vehículo y menor potencia.
- A mayor antigüedad de la póliza mayores edades, mayor antigüedad del carnet del primer conductor y mayor antigüedad del vehículo.
- A mayor edad del primer conductor mayor antigüedad del carnet del primer conductor y mayor antigüedad del vehículo
- La correlación de 0.837 entre la edad del primer conductor y la de máximo riesgo es lógica, pues de los 6 802 valores, 6 091 son iguales. También observamos que la correlación entre la edad de máximo riesgo y la de la antigüedad de carnet del primer conductor es algo menor que la correlación entre la edad del primer conductor y su antigüedad del carnet.
- A mayor potencia peor nivel de bonus-malus y menor antigüedad del vehículo.

Cualitativo con cualitativo

En la siguiente tabla detallamos la medida (3.4) y el *p*-valor del anova correspondiente (anexo 5.11), y obtenemos los siguientes resultados:

$\eta =$ $p-\text{valor} =$	Pago	Sexo	Zona
Plazas	0.022 0.181	0.023 0.062	0.034 0.105
Bonus-malus	0.191 0	0.006 0.625	0.171 0
Antipol	0.186 0	0.027 0.025	0.109 0
Edad1	0.149 0	0.142 0	0.070 0
Edadx	0.152 0	0.119 0	0.062 0
Anticarn	0.136 0	0.151 0	0.057 0
Antivehi	0.041 0.003	0.002 0	0.056 0
Potencia	0.033 0.023	0.081 0	0.081 0

Obtenemos que la variable cuantitativa más asociada con la zona es el nivel de bonus-malus. Esto es así, ya que ambas variables han sido construidas a partir de la siniestralidad histórica de los conductores. El bonus-malus a nivel individual y la zona a nivel global. Nosotros adicionalmente hemos reagrupado las zonas iniciales en función del número de siniestros de esta cartera.

En el anexo 5.11 encontramos las anovas asociadas a la forma de pago y al sexo con las variables cuantitativas consideradas. Podemos realizar los siguientes comentarios, conjuntamente con la información desprendida a partir de los gráficos de medias:

- Forma de pago con número de plazas: las formas de pago fraccionadas tienen un menor número de plazas, aunque el gráfico no es concluyente, ya que el *p*-valor del anova indica que no se aprecian diferencias significativas.
- Sexo del primer conductor con número de plazas: la media del número de plazas de las

mujeres es menor que la de los hombres.

- Nivel de bonus-malus con forma de pago: la media del nivel de bonus-malus es peor contra más fraccionado sea el pago.
- Nivel de bonus-malus con sexo del primer conductor: la media del nivel de bonus-malus es peor en los hombres que en las mujeres, aunque las diferencias no resultan significativas.
- Antigüedad de la póliza con forma de pago: Las antigüedades medias son menores para los pagos fraccionados que para el anual.
- Antigüedad de la póliza con sexo del primer conductor: la antigüedad media de las pólizas de los hombres es menor que la de las mujeres.
- Edad del primer conductor con forma de pago: los pagos fraccionados están asociados a edades menores.
- Edad con sexo del primer conductor: la media de edad de las mujeres es menor que la de los hombres.
- Edad de máximo riesgo con forma de pago: Éste es un gráfico similar al de la edad del primer conductor con la forma de pago. Observamos que la medida de asociación calculada en la tabla anterior indica que la edad de máximo riesgo está algo más asociada con la forma de pago que la edad del primer conductor.
- Edad de máximo riesgo con sexo del primer conductor: Éste es un gráfico similar al de la edad del primer conductor con la forma de pago.
- Antigüedad del carnet del primer conductor con forma de pago: Los pagos fraccionados están asociados a antigüedades medias menores.
- Antigüedad del carnet y sexo del primer conductor: la media de antigüedad del carnet es menor en las mujeres, del mismo modo que tenían menores edades.
- Antigüedad del vehículo con forma de pago: La forma de pago semestral tiene asociada una media de antigüedad del vehículo mayor que la anual, y ésta última mayor que la trimestral.
- Antigüedad del vehículo con sexo del primer conductor: no existen diferencias en las medias de antigüedad del vehículo según el sexo.
- Potencia con forma de pago: la mayor potencia media es la asociada a los pagos trimestrales.
- Potencia con sexo del primer conductor: la potencia media asociada a las mujeres es menor que la asociada a los hombres.

5.4.4. Agregación de los datos

Para poder realizar el estudio de la frecuencia de siniestralidad procedemos, de manera detallada, al proceso de agregación de los datos.

En primer lugar nos hacemos una idea de la información desprendida por los factores disponibles, los cuales son:

- *Factores relativos al vehículo*: tipo de combustible, número de plazas, antigüedad del vehículo y potencia.
- *Factores relativos al conductor*: forma de pago, sexo, nivel de bonus-malus, antigüedad de la póliza, edad del primer conductor, edad del segundo conductor, antigüedad del carnet del primer conductor.
- *Factores relativos a la circulación*: zona de circulación

Empezamos por los factores que descartamos directamente del análisis por falta de datos:

- el tipo de combustible lo descartamos, ya que en la porción de datos con que operamos tan sólo tenemos de diesel (5 133 vehículos) y de *missing* (1 669 vehículos), por lo que dentro de los que no conocemos el tipo de combustible también podrían haber de diesel,
- el valor del vehículo, ya que de las 6 802 pólizas, 6 398 son *missing*.

Puesto que todos ellos han sido anotados de manera discreta en el fichero cedido de SPSS, nos es fácil realizar un estudio previo de anova incluyendo los gráficos de medias, del número de siniestros con cada factor individual. Éste estudio se realiza en el anexo 5.10, y se comenta a continuación:

- Respecto a *los factores relativos al vehículo*:

  - Las medias del número de siniestros según el número de plazas del vehículo son significativamente diferentes, p-valor = 0. Si observamos el gráfico de medias observamos una tendencia creciente del número de plazas con respecto del número de siniestros, excepto para los de 2 y 3 plazas y para los de 9 plazas. La menor media es para los de 4 plazas y la mayor para los de 2, 7 y 8 plazas. Los de 3, 5, 6 y 9 plazas tienen una media similar. Por ello el número de plazas lo tendremos en cuenta para la agregación haciéndolo pasar previamente por

una agrupación de sus clases tratándolo como predictor libre. Tal agrupación la detallamos en el siguiente apartado 5.4.4.1.

En la siguiente tabla de contingencia que cruza la potencia con el número de plazas, observamos como en general a mayor número de plazas mayor potencia, excepto para los de 2 y 3 plazas, y los de 9 plazas. Esta tabla nos ayuda a entender la tendencia en las medias del número de siniestros respecto del número de plazas.

Tabla de contingencia POTENCIA * PLAZA

Recuento

		PLAZA								
		2	3	4	5	6	7	8	9	Total
POTENCIA	[...,28]			36	5					41
	[29,33]			5	43					48
	[34,42]			5	413					418
	[43,53]			22	863		1			886
	[54,75]	2	4	25	2829	8	5	16	38	2927
	[76,94]	4	2	4	1242	1	41		9	1303
	[95,118]			2	618		31	2	5	658
	[119,215]	3		3	445		19	2		472
	[216,...]	5		3	41					49
Total		14	6	105	6499	9	97	20	52	6802

- La correlación de la antigüedad del vehículo con el número de siniestros sale negativa, $\rho = -0.029$, lo que nos indica que a mayor antigüedad del vehículo menor número de siniestros. Pero sería de esperar que los vehículos más antiguos ocasionaran mayor número de siniestros porque estuvieran en peor estado. Sin embargo esto no ocurre así, ya que la antigüedad está negativamente correlacionada con la potencia del vehículo y positivamente características del conductor en las que se refleja la experiencia (el nivel de bonus-malus, con la antigüedad de la póliza y con la edad del primer conductor). Por ello no tendremos en cuenta a la antigüedad para la agregación, ya que el efecto de ésta es debido a otros factores, los cuales sí tendremos en cuenta.
- La potencia tiene una relación creciente y bastante lineal con respecto el número de siniestros. A mayor potencia mayor número de siniestros. El p-valor del anova es de 0.025 y nos indica que las diferencias en las medias del número de siniestros son significativas globalmente con respecto a la potencia, pero observamos que algunas son similares, especialmente en las últimas clases. Creemos conveniente hacer pasar a la potencia por una agrupación previa de

clases, siendo tratada como predictor monótono. Tal agrupación la detallamos en el siguiente apartado 5.4.4.1.

➢ Respecto a *los factores relativos al conductor*:

- La forma de pago y el sexo los incorporamos tal cual para la agregación.
- El nivel de bonus-malus, lo haremos pasar por un proceso de agrupación de clases en el siguiente apartado 5.4.4.1, ya que: en el gráfico de medias se observa una tendencia lineal decreciente con respecto al número de siniestros, como es de esperar, excepto para la clase de escala malus –20 donde las pólizas no han sufrido siniestro. Adicionalmente las tres últimas clases, que son –50, –40 y –20 tienen tan sólo dos o tres pólizas cada una. También observamos una siniestralidad similar en las clases contiguas de 0 y 10, y en las de 20 y 30. Por ello procede una agrupación previa de clases para conseguir una mayor linealidad en el decrecimiento y una mejor agregación de datos.

  Aprovechamos para notar que si nos fijamos en los valores máximos del número de siniestros en los descriptivos del anexo 5.10, resulta que el máximo para las pólizas de escala malus ha sido de 0, de 1 y como mucho de 2. Esto nos lleva a pensar que en esta cartera una vez se está en la escala de malus, el asegurado es prudente. En el contexto del MLG de estructura de error Poisson para la frecuencia de siniestralidad, modelo (3.54), este hecho equivale a relajar la hipótesis de Poisson. El efecto que tiene es una infradispersión en el modelo, que llevado al caso extremo se corresponde con el caso Binomial. Resaltamos esto, ya que una vez realizado el proceso de selección en el apartado 5.4.5 con el modelo de estructura de error Poisson (3.54), y contrastamos que es disperso, realizamos una estimación del parámetro de dispersión común a las celdas utilizando la media de desvianzas (3.36), y obtenemos $\hat{\phi} < 1$. Por lo tanto aquí tenemos una posible justificación de este resultado.

- No tiene sentido poner la antigüedad de la póliza. Si nos fijamos, la mayoría de pólizas se concentran en antigüedades de uno a tres años. Esto es así, ya que la antigüedad de la póliza no se refiere a los años de antigüedad del asegurado en la compañía, sino a la antigüedad del vehículo asegurado en la compañía.

El historial de siniestralidad del asegurado se ve reflejado en el nivel bonus-malus, el cual ya incluimos para realizar la agregación, y no de iniciar un nuevo contrato. En la siguiente tabla observamos como, por ejemplo, los de antigüedad de menos de 1 año están en diferentes niveles de bonus, ya que se les ha mantenido el historial de siniestralidad. La antigüedad de la póliza no tiene que ver ni con las características del conductor ni con las del vehículo.

Tabla de contingencia ANTIPOL * BONUS

Recuento

		BONUS												
		-50	-40	-20	-10	0	10	20	30	35	40	45	50	Total
ANTIPO	[0,1)					4	8	11	26	28	20	40	135	272
	[1,2)			1	9	32	82	81	153	192	163	286	938	1937
	[2,3)			1	2	21	17	78	82	116	174	96	802	1389
	[3,4)		1		3	13	21	37	90	54	96	135	504	954
	[4,5)				1	8	16	18	7	75	44	77	354	600
	[5,6)	1			2	9	12	39	24	22	80	71	308	568
	[6,7)	1	1	1	5	4	8	9	29	41	21	102	358	580
	[7,...)					1	5	7	3	32	18	26	410	502
Total		2	2	3	22	92	169	280	414	560	616	833	3809	6802

Adicionalmente el *p*-valor del anova es de 0.254.

- Respecto a la edad del primer conductor observamos lo siguiente: lo primero que nos llama la atención es que en las primeras edades, tan sólo tenemos a segundos conductores. Este hecho nos lleva a observar la siniestralidad de tales pólizas. Realizamos el cruce de la edad del primer conductor y de la del segundo para las pólizas que han declarado un segundo conductor:

Tabla de contingencia EDAD1 * EDAD2

Recuento

		EDAD2													
		[18,20]	[21,25]	[26,30]	[31,35]	[36,40]	[41,45]	[46,50]	[51,55]	[56,60]	[61,65]	[66,70]	[71,75]	[76,80]	Total
EDAD1	[21,25]		1												1
	[26,30]			8	2	1			1		1				13
	[31,35]		6	12	27	4	1	2				2			54
	[36,40]	5	1	6	16	19	6	3				1			57
	[41,45]	8	18	7	2	15	17	4	3			1			75
	[46,50]	15	40	10		5	5	17	1		1			1	95
	[51,55]	8	55	39	4	6	5	6	8	6	1				138
	[56,60]	5	29	34	14	3	5	5	9	4	1				109
	[61,65]	4	15	23	18	9	1	1	1	4	5	2	1		84
	[66,70]		9	19	11	7	1			2	2	1			52
	[71,75]		3	4	5	2	2	2	1	1	2	1	1		24
	[76,80]	1		2	1			1	1	1			1		8
	[81,99]							1							1
Total		46	177	164	100	71	43	42	25	18	13	8	3	1	711

y observamos que fuera de la diagonal, la tabla de contingencia está llena especialmente en el triángulo inferior, es decir, la mayoría de los que han declarado segundo conductor, éste es de menor edad. Los del triángulo superior son de aproximadamente edades parecidas a las del primer conductor, excepto para algún caso. Por ello creamos la edad de máximo riesgo, calculada como el mínimo de ambas edades en el caso de que se haya declarado segundo conductor; sino la edad coincide con la del primer conductor. Si observamos los gráficos de medias del número medio de siniestros en la edad del primer conductor y en la edad de máximo riesgo, observamos claramente como las pólizas que han declarado segundo conductor tienen un mayor número de siniestros, en especial para las primeras edades. Por lo tanto trabajaremos con la edad de máximo riesgo y no con la del primer conductor.

- La antigüedad del carnet del primer conductor no es representativa de la realidad en estos datos, debería ser "la antigüedad de máximo riesgo", por lo que no la tendremos en cuenta por sí sola. Sin embargo vamos a crear en el próximo apartado una nueva variable a la que denominaremos Edad|Anticarn, la cual nos definirá diferentes situaciones de riesgo.

  Esta nueva variable, Edad|Anticarn será creada combinando la edad de máximo riesgo con la antigüedad del carnet del primer conductor. Vamos a justificar porqué lo realizamos así. La antigüedad del carnet refleja la experiencia del conductor, se espera que a mayor antigüedad en el carnet menor número de siniestros. En las siguientes tablas de contingencia hemos cruzado la antigüedad del carnet del primer conductor con la edad del primer conductor y con la edad de máximo riesgo, y observamos que los que tienen antigüedad del carnet inferior a 10 años y que son mayores de 30 años, o bien no han declarado segundo conductor o bien el segundo conductor es una edad similar. Sin embargo, los que han declarado segundo conductor han sido los de antigüedad mayor a 10 años. Para verlo tan sólo nos hemos de fijar en los de 18 a 25 años (edad de máximo riesgo) de la segunda tabla, en que todos (46 + 172 pólizas) tienen antigüedad mayor a 10 años. Nos referimos al siguiente apartado en el que construimos la nueva variable para observar la idoneidad de ésta para describir las diferentes situaciones de riesgo.

Tabla de contingencia ANTICARN * EDAD1

Recuento

	EDAD1													Total
	[21,25]	[26,30]	[31,35]	[36,40]	[41,45]	[46,50]	[51,55]	[56,60]	[61,65]	[66,70]	[71,75]	[76,80]	[81,99]	
ANTICAR [1,2)		2		1										3
[2,3)		2	2	2					1					7
[3,4)		2	3	6			1							12
[4,5)	1	7	1	1		2	1		1					14
[5,6)		11	5	1	1	3	2	1		1				25
[6,7)		14	15	6	1	1	1	1						39
[7,8)	2	20	15	10	11	4	3	1	2	1				69
[8,9)		20	22	11	4	4	4	4	3					72
[9,10)		36	43	23	4	7	4	3	3		1			124
[10,...		109	602	846	820	826	815	791	662	529	293	120	24	6437
Total	3	223	708	907	841	847	831	801	672	531	294	120	24	6802

Tabla de contingencia ANTICARN * EDADX

Recuento

	EDADX														Total
	[18,20]	[21,25]	[26,30]	[31,35]	[36,40]	[41,45]	[46,50]	[51,55]	[56,60]	[61,65]	[66,70]	[71,75]	[76,80]	[81,99]	
ANTICAR [1,2)			2		1										3
[2,3)			2	2	2					1					7
[3,4)			2	3	6			1							12
[4,5)		2	7	1	1		2			1					14
[5,6)			12	5	1	1	3	2			1				25
[6,7)			14	15	6	1	1	1	1						39
[7,8)		5	20	17	8	10	4	1	1	2	1				69
[8,9)			20	23	10	4	4	4	4	3					72
[9,10)			39	42	23	5	4	4	3	3		1			124
[10,...	46	172	261	653	868	789	770	707	696	590	479	271	112	23	6437
Total	46	179	379	761	926	810	788	720	705	600	481	272	112	23	6802

Cabe notar que el cruce de la edad con el sexo en esta cartera no tiene sentido. Deberíamos tener en cuenta "el sexo de máximo riesgo", el cual no ha sido anotado.

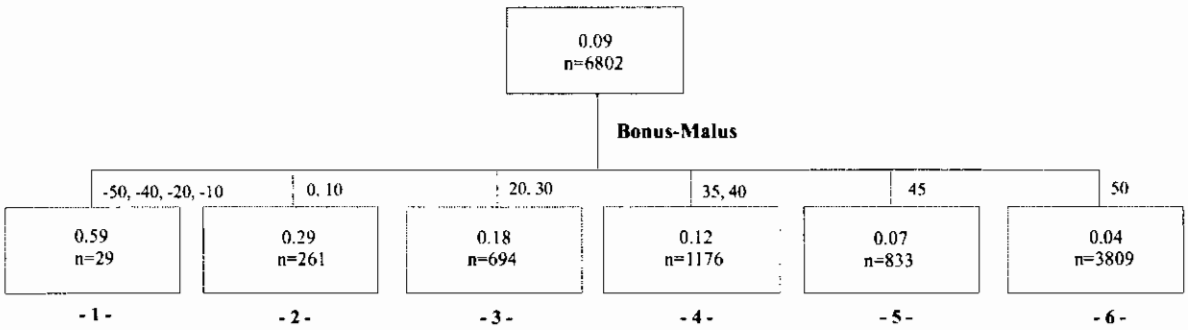
- Respecto a *los factores relativos a la circulación*, no hay problema, tan sólo tenemos la zona, la cual ya ha sido tratada en el primer apartado de este capítulo. Así para la agregación de datos la tendremos en cuenta con las cinco categorías en que ha sido agrupada.

5.4.4.1. Discretización de los factores

Hemos agrupado los siguientes predictores haciendo uso del algoritmo ordinal de SPSS:

Bonus-Malus

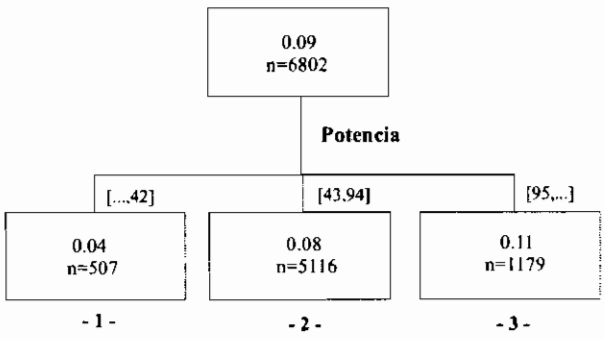
Lo hemos definido como predictor monótono, ya que pretendemos darle tratamiento cuantitativo en las regresiones. Hemos fijado un nivel de significación para la fase de agrupación de categorías del 5%, y no hemos exigido un número mínimo de pólizas para formar un grupo, obteniendo el siguiente resultado:



En lo que sigue denominaremos Bonus|Malus a esta nueva variable reagrupada en 6 clases.

Potencia

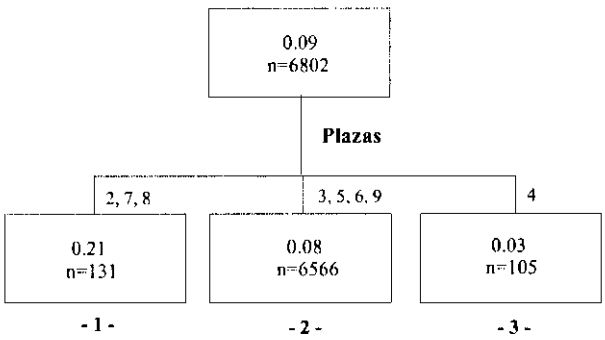
La hemos definido como predictor monótono (si lo definimos como libre obtenemos el mismo resultado), ya que pretendemos darle tratamiento cuantitativo en las regresiones. Hemos fijado un nivel de significación para la fase de agrupación de categorías del 5%, y no hemos exigido un número mínimo de pólizas para formar un grupo, obteniendo el siguiente resultado:



En lo que sigue denominaremos Potencia a esta nueva variable reagrupada en 3 clases.

Plazas

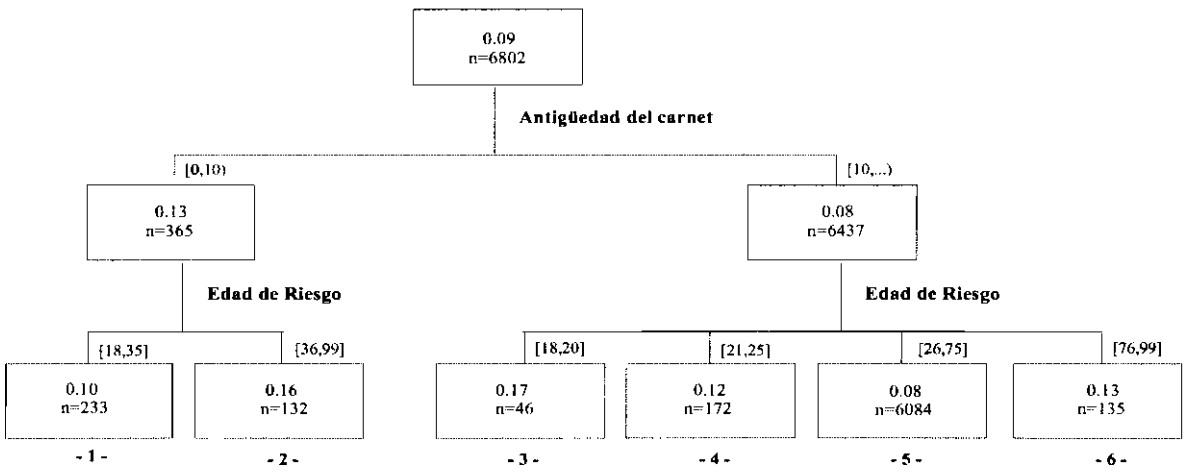
La hemos definido como predictor libre. Hemos fijado un nivel de significación para la fase de agrupación de categorías del 5%, y no hemos exigido un número mínimo de pólizas para formar un grupo, obteniendo el siguiente resultado:



En lo que sigue denominaremos Plazas a esta nueva variable reagrupada en 3 clases.

Edad de máximo riesgo y antigüedad del carnet del primer conductor

Hemos definido ambos predictores como monótonos. Hemos fijado un nivel de significación para la fase de agrupación de categorías del 25% con el objetivo de que no junte en exceso las clases de las variables y, al igual que en los otros casos no hemos exigido un número mínimo de pólizas para formar un grupo, obteniendo el siguiente resultado:



Observamos la idoneidad de esta nueva variable para determinar las situaciones de riesgo:

⇒ Si la antigüedad del carnet del primer conductor es mayor a 10 años, distinguiremos

- Segmento 3: si la edad de máximo riesgo es de 18 a 20 años, situación en la que tenemos un segundo conductor muy joven, tenemos el máximo número medio de siniestros, 0.17.
- Segmento 4: si la edad es de 21 a 25, el número disminuye pero sigue siendo elevado, 0.12, ya que seguimos teniendo un segundo conductor joven.
- Segmento 5: si la edad es de 26 a 75 años el número medio es el menor, de 0.08.
- Segmento 6: si pasa de 76 años la siniestralidad vuelve a subir más o menos como para los de 21 a 25 años, es de 0.13

⇒ Si la antigüedad del carnet del primer conductor es menor a 10 años, distinguiremos:

- Segmento 1: si la edad de máximo riesgo es de 18 a 35 años, la media será de 0.10, la cual es menor que en el caso de declarar un segundo conductor entre 18 y 25 años. Cuando la antigüedad de estas pólizas sea superior a 10 años, pasará a una media menor de 0.08 de manera razonable, pues su experiencia será mayor.
- Segmento 2: si la edad es más de 36 años, la media será elevada, de 0.16.

Respecto a las primeras edades:

Consideramos razonable asignar, según esta nueva variable, a un asegurado de antigüedad del carnet menor a 10 años y de edad de 18 a 35 años (segmento 1) una siniestralidad media menor que a un asegurado de antigüedad mayor a 10 años con un segundo conductor entre 18 y 25 años (segmentos 3 y 4). Recordemos que en esta cartera tan sólo tenemos a 3 asegurados con edad del primer conductor entre 18 y 25 años, más concretamente no tenemos ninguno de edad de 18 a 21 años y tenemos 3 entre 22 y 25 años. Esta cartera pretende estar depurada de tales conductores jóvenes. Sin embargo, estos se han introducido como segundos conductores ocasionando la mayor siniestralidad de la cartera y pagando primas bajas, ya que, si observamos la siguiente tabla de contingencia que cruza la edad de máximo riesgo y el nivel de bonus,

Tabla de contingencia EDADX * BONUS

Recuento

		BONUS												
		-50	-40	-20	-10	0	10	20	30	35	40	45	50	Total
EDADX	[18,20]					1		1	4	3	5	5	27	46
	[21,25]					1	4	4	11	18	27	28	86	179
	[26,30]			1	3	6	8	29	26	37	37	49	183	379
	[31,35]	1		1	3	6	32	31	48	67	79	106	387	761
	[36,40]				3	11	30	30	60	86	72	117	517	926
	[41,45]			1	2	10	24	34	47	52	67	103	470	810
	[46,50]				1	36	23	34	74	60	66	89	405	788
	[51,55]		1		1	5	14	30	30	65	77	88	409	720
	[56,60]	1	1		6	6	13	43	40	52	60	83	400	705
	[61,65]				1	6	9	25	33	46	59	75	346	600
	[66,70]				2	2	4	8	25	41	39	60	300	481
	[71,75]					1	5	7	10	19	25	19	186	272
	[76,80]					1	3	3	4	13	2	10	76	112
	[81,99]							1	2	1	1	1	17	23
Total		2	2	3	22	92	169	280	414	560	616	833	3809	6802

las pólizas con los segundos conductores de edades entre 18 y 25 años, tienen asignados niveles elevados de bonificaciones debidas al primer conductor que lo abala, y en ningún caso están en escala de malus. Por lo tanto, al declarar el segundo conductor se aplica en la prima un descuento debido a la variable bonus, y por lo tanto la siniestralidad máxima de 0.17 y 0.12 de los segmentos 3 y 4 queda reducida, así que se iguala a la del segmento 1.

En lo que sigue denominaremos Edad|Anticarn a esta nueva variable reagrupada en 6 clases.

5.4.4.2. Resultado

Teniendo en cuenta los análisis anteriores, de toda la información original, nos quedamos finalmente como factores potenciales de riesgo para la agregación de los datos con:

- Pago, con 3 clases
- Sexo, con 2 clases
- Zona, con 5 clases
- Plazas, con 3 clases
- Edad|Anticarn, con 6 clases
- Bonus|Malus, con 6 clases
- Potencia, con 3 clases

Hemos procedido a la agregación de las 6 802 pólizas a partir de los factores anteriores. Para ello hemos implementados el programa cummodel.m que encontramos en el anexo 4.2, el cual nos agrega los datos creando la nueva matriz de datos, en la que como respuesta se tiene la frecuencia de siniestralidad para las celdas no vacías junto con el correspondiente peso. La matriz de predictores puede hacer referencia a la necesaria para operar con la RBD, o a la necesaria para operar con el MLG, la cual incluye todas la variables binarias necesarias. Hemos obtenido un total de 712 celdas no vacías, por lo que hemos reducido considerablemente el volumen de los datos.

En el anexo 5.12 encontramos las anovas entre la frecuencia de siniestralidad generada con la agregación y los factores discretos que han sido utilizados para la misma. Las medias se corresponden con las medias de la frecuencia de siniestralidad calculada y no con la media real del número de siniestros, ya que hemos pasado a trabajar con datos agregados y por ello hemos perdido información.

En la siguiente tabla encontramos la medida de asociación (3.4) calculada entre tal frecuencia de siniestralidad y cada factor empleado, junto con el correspondiente p-valor de las anovas que encontramos en el anexo 5.12:

Variable	η	p-valor
Y_u – Pago	0.112	0
Y_u – Sexo	0.001	0.966
Y_u – Zona	0.150	0
Y_u – Plazas	0.125	0
Y_u – Edad|Anticarn	0.109	0
Y_u – Bonus|Malus	0.445	0
Y_u – Potencia	0.1	0

Observamos que el sexo es la variable que diferencia menos las medias de la frecuencia de siniestralidad. Según los p-valores el resto de factores proporcionan medias diferentes. Según (3.4) el orden de asociación individual es el siguiente: Bonus|Malus, Zona, Plazas, Pago, Edad|Anticarn, Potencia y finalmente Sexo.

Para el estudio de la frecuencia de siniestralidad tanto con el MLG como con la RBD el Bonus|Malus y la Potencia serán tratados como predictores cuantitativos dada su relación lineal con el número de

siniestros. En los gráficos de medias de estas nuevas variables (anexo 5.12) observamos una mayor linealidad que en los de las variables discretizadas iniciales (anexo 5.10).

Las puntuaciones cuantitativas para sus clases que utilizaremos en los modelos de regresión serán las medias de las variables cuantitativas creadas en el apartado 5.4.2 (recordemos que inicialmente todas las variables habían sido anotadas como discretizaciones):

BONUSR	Media
1	-15,8621
2	6,4751
3	25,9654
4	37,6190
5	45,0000
6	50,0000
Total	42,8440

POTR	Media
1	36,2860
2	66,8637
3	135,4377
Total	76,4705

5.4.5. Modelo lineal generalizado sobre la frecuencia de siniestralidad

Tenemos en cuenta como factores cuantitativos el Bonus|Malus y la Potencia, y como cualitativos el Pago, el Sexo, la Zona y la Edad|Anticarn. Manejamos en total 17 parámetros (1 + 1 + 1 + 2 + 1 + 4 + 5). Utilizamos el modelo con estructura de error Poisson ponderado (3.54) junto con el enlace logarítmico para la realización del proceso de selección de predictores.

Proceso de selección paso a paso

Hacemos uso de la distribución asintótica F de Fisher (3.41) en la contrastación de (3.38). A continuación detallamos los resultados del proceso de selección paso a paso:

Fases de introducción:

Variable	*p*-valor	*p*-valor|F(1)	*p*-valor |F(1)F(2)	*p*-valor|F(1)...F(3)
Pago	0.000802	0.517265	0.521682	0.591576
Sexo	0.984094	0.836044	0.869323	0.937068
Zona	9.3065E-7	0.001155	**0.000847**	--------
Plazas	7.4355E-5	**1.0764E-5**	--------	--------
Edad|Anticarn	0.009767	0.004325	0.001990	**0.002061**
Bonus|Malus	**2.1788E-53**	--------	--------	--------
Potencia	0.000941	0.002527	0.014096	0.035619
	F(1) = **Bonus|Malus**	F(2) = **Plazas**	F(3) = **Zona**	F(4) = **Edad|Anticarn**

Variable	*p*-valor|F(1)...F(4)	*p*-valor|F(1)...F(5)	*p*-valor|F(1)...F(6)
Pago	0.637266	**0.66714**	--------
Sexo	0.853554	0.937431	**0.928369**
Zona	--------	--------	--------
Plazas	--------	--------	--------
Edad|Anticarn	--------	--------	--------
Bonus|Malus	--------	--------	--------
Potencia	**0.014506**	--------	--------
	F(5) = **Potencia**	F(6) = **Pago**	F(7) = **Sexo**

Fases de eliminación (expresadas como un proceso de eliminación progresiva):

Variable	*p*-valor	*p*-valor/F[1]	*p*-valor/F[1]F[2]	*p*-valor/F[1]...F[3]
Pago	0.667231	**0.667143**	--------	--------
Sexo	**0.928369**	--------	--------	--------
Zona	0.002151	0.002136	0.001918	0.000889
Plazas	1.8457E-5	1.7104E-5	1.6563E-5	3.2766E-6
Edad|Anticarn	0.001167	0.001143	0.001048	**0.002061**
Bonus|Malus	5.0519E-48	4.1345E-48	4.1303E-50	1.8998E-50
Potencia	0.015754	0.015428	**0.014506**	--------
	F[1] = **Sexo**	F[2] = **Pago**	F[3] = **Potencia**	F[4] = **Edad|Anticarn**

Variable	*p*-valor/F[1]...F[4]	*p*-valor/F[1]...F[5]	*p*-valor/F[1]...F[6]
Pago	--------	--------	--------
Sexo	--------	--------	--------
Zona	**0.000847**	--------	--------
Plazas	7.9375E-6	**1.0764E-5**	--------
Edad|Anticarn	--------	--------	--------
Bonus|Malus	2.4185E-50	4.3204E-54	**2.1788E-53**
Potencia	--------	--------	--------
	F[5] = **Zona**	F[6] = **Plazas**	F[7] = **Bonus|Malus**

Si fijamos por ejemplo un nivel de significación del 5%, obtenemos como resultado en la selección las variables:

Bonus|Malus, Plazas, Zona, Edad|Anticarn y Potencia

Nota respecto al enlace identidad: Hemos intentado realizar el proceso para el enlace identidad en lugar del logarítmico, y nos ha sido imposible. Hemos obtenido estimaciones de 0 para la frecuencia de siniestralidad. Si nos fijamos en la expresión de la desvianza para la distribución de Poisson de la

tabla 3.3 del anexo 3.2, observamos que cuando se obtienen estimaciones de 0 nos encontramos con ese 0 en un denominador. Tal y como indican Brockman y Wright (1992), el enlace logarítmico es apropiado para obtener, al menos demostrado empíricamente, siempre estimaciones positivas sea cual sea la distribución del error empleada. Así, también es adecuado utilizar el enlace logarítmico en el caso de la cuantía por siniestro: en el caso de la Gamma si obtenemos alguna estimación negativa, en el cálculo de las desvianzas necesitaríamos calcular logaritmos de valores negativos, y en el caso de la Gaussiana Inversa, al igual que para el Poisson, con estimaciones de 0 tendríamos una indeterminación. Esto puede ser una luz de alarma para darnos cuenta de que el enlace no es el adecuado, o quizá no lo sea la distribución.

Validación del modelo resultante

Validación general

Respecto a la validación del modelo, vemos primeramente la significación conjunta de los coeficientes finalmente incluidos, calculando el cociente de verosimilitudes (3.46) y el *p*-valor del estadístico (3.45) asociado al contraste (3.44):

$$R^2 = \frac{(884.3231 - 576.4465)}{884.3231} = 0.3481$$

$$F_{13,698} = \frac{(884.3231 - 576.4465)/13}{576.4465/698} = 28.6767 \quad p\text{-valor} = 1.5091 \times 10^{-56}$$

El *p*-valor obtenido nos indica que los predictores incluidos tienen conjuntamente poder predictivo en la regresión.

Dispersión

La media y la varianza de la frecuencia de siniestralidad son: $E[Y_u] = 0.0851$ y $Var(Y_u) = 0.024$.

Hemos aplicado **el contraste de dispersión basado en un modelo de regresión** detallado en el apartado 3.5.3.1.1. Obtenemos los siguientes *p*-valores en la contrastación de los coeficientes individuales en cada regresión (haciendo uso de la distribución *F*):

$\Rightarrow$ al regresar sobre la predicción: p-valor = 2.2224×10^{-29}

$\Rightarrow$ al regresar sobre la constante: p-valor = 8.4304×10^{-71}

De lo que concluimos que en ambos casos se rechaza la hipótesis nula de igualdad entre media y varianza, por lo que el modelo es (infra o sobre) disperso.

Siguiendo a Albrecht (1983a)

Siguiendo los pasos que propone Albretch, partimos de los datos agregados, con $n=6802$ pólizas, $m=712$ celdas no vacías, y $r=16$ predictores incluidos en la regresión. Realizamos el ajuste del modelo (3.54) combinado con el enlace logarítmico. Y calculamos (3.62) haciendo uso de la información desagregada, (3.61) y (3.63) haciendo uso de la agregada, y finalmente (3.64). Por lo que primero contrastamos la validez de las hipótesis del modelo Poisson estimado con todos los predictores potenciales que nos han servido para agregar:

$$Q=7653.7074\sim\chi^2_{6785}\quad p\text{-}valor=3.8448\times10^{-13}$$

$$Q_w=6802.0452\sim\chi^2_{6090}\quad p\text{-}valor=2.4504\times10^{-10}$$

$$Q_d=851.6621\sim\chi^2_{695}\quad p\text{-}valor=4.0775\times10^{-5}$$

$$F=\frac{851.6621/695}{6802.0452/6090}=\frac{1.2254}{1.1169}=1.0971\sim F_{(695,6090)}\qquad p\text{-}valor=0.04724$$

Según Albrecht con ninguno de los estadísticos obtenemos un buen resultado, excepto con la F, que dependiendo del nivel de significación que establezcamos podemos aceptar o rechazar la validez de la regresión.

Posteriormente, suponiendo que aceptáramos el modelo, Albrecht (1983a) propone la selección de variables. Realizamos el proceso, tal y como se detalla en el apartado 5.4.5, y nos quedamos con $r=13$ predictores. Llamamos $\hat{\lambda}_{(r=16,m=712)}$ a la predicción con la que hemos validado, y $\hat{\lambda}_{(r=13,m=712)}$ a la predicción que obtenemos con los $r=13$ predictores a partir de las 712 observaciones ponderadas iniciales. Ahora consideramos que deberíamos volver a validar con los 13 predictores. Pero tan sólo tiene sentido calcular:

$$Q = 7530.7588 \sim \chi^2_{6788} \quad p\text{-}valor = 3.5912 \times 10^{-10}$$

el cual observamos que ha mejorado algo, ya que los predictores utilizados, aunque han sido menos en número, han resultado más adecuados para la estimación.

Para poder recalcular Q_w y Q_d necesitamos previamente calcular las medias de siniestralidad de las nuevas $m = 354$ celdas obtenidas a partir de los $r = 13$ predictores, y lo lógico sería volver a calcular $\hat{\lambda}_{(r=13,m=354)}$, es decir, deberíamos agregar los datos con los $r = 13$ predictores para obtener las $m = 354$ nuevas observaciones ponderadas a partir de las cuales realizar una nueva estimación. Sino los grados de libertad no cuadran. Es decir, si nos quedamos con $\hat{\lambda}_{(r=13,m=712)}$, aunque partimos de $m = 712$ celdas, tan sólo obtendremos $m = 354$ estimaciones diferentes, adicionalmente si utilizamos $\bar{N}_u$ obtenida a partir de $m = 712$, resultará que $Q \neq Q_w + Q_d$.

Consideramos no lógico validar el modelo antes de realizar la selección de predictores tal y como se propone en Albrecht (1983a), puesto que si el modelo no se ajuste puede ser debido tanto a la distribución del error empleada, a la función de enlace y/o al conjunto de predictores seleccionados. Y en nuestro caso el objetivo es la selección de las variables de tarifa.

Estimación del parámetro de dispersión

El test de dispersión nos ha indicado que el modelo es disperso. Ahora suponemos un parámetro constante y por lo tanto igual para cada celda, y calculamos su estimación mediante (3.36), que realiza una media con las desvianzas del modelo, y resulta $\hat{\phi} = 0.8259$.

El parámetro es menor que 1, lo que en conjunto implica infradispersión. Como ya se ha comentado, esta estimación es común a todas las celdas, lo que no implica que para cada perfil sea el mismo sino de manera conjunta. Para hacernos una idea de que cada celda aporta una desvianza diferente realizamos el gráfico de los residuos individuales de desvianza del modelo:

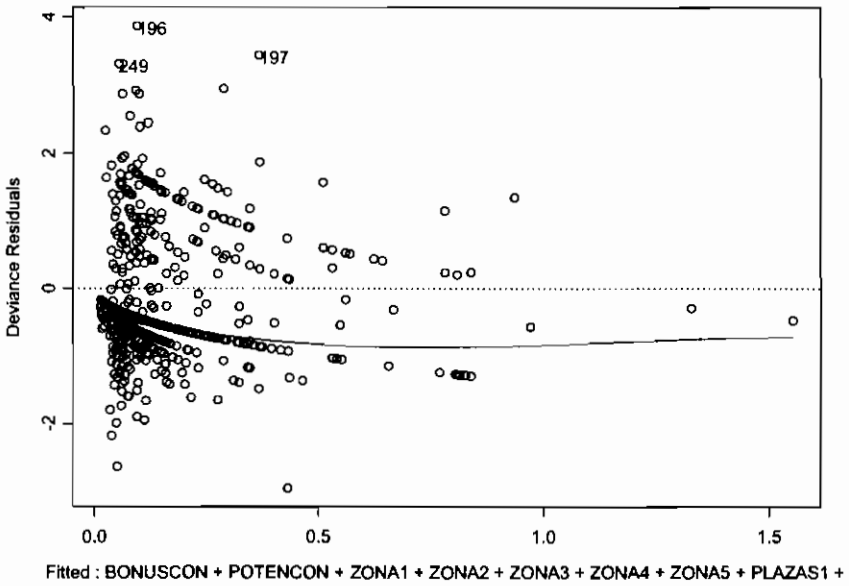


Figura 5.7. Gráfico de los residuos de desvianza.

Si seguimos a Brockman y Wright (1992), cuando la hipótesis que no se cumple es la de Poisson, es decir, que las intensidades no son uniformes para todo el período de observación, este hecho se recoge en una infradispersión del modelo, y puede ser recogida razonablemente de manera común a las celdas.

Tal y como hemos contado en el apartado 5.4.4, si nos fijamos en los valores máximos del número de siniestros en los descriptivos del bonus-malus inicial (anexo 5.10), observamos que el máximo para las pólizas de escala malus ha sido de 0, de 1 y como mucho de 2 siniestros. Esto nos lleva a pensar que en esta cartera una vez se está en la escala de malus, el asegurado es más prudente, por lo que la intensidad no es constante a lo largo del período para todas las pólizas. Ésta puede ser una de las razones de la infradispersión. Pero como hemos detallado en el capítulo 3, pueden no cumplirse también el resto de hipótesis, o no ser adecuada la función de enlace empleada. Por lo que la conclusión más prudente es que los efectos de todas estas causas nos llevan a aplicar el modelo de Poisson infradisperso (3.59).

Los coeficientes resultantes del predictor lineal, y los errores estándar de éstos son respectivamente:

Término	**Coeficiente**	**Error Estándar para $\hat{\phi} = 0.8259$**	**Error Estándar para $\phi = 1$**
Constante[28]	–1.3647	0.5817	0.6401
Bonus	–0.0395	0.0021	0.0024
Plazas1	1.7829	0.5545	0.6102
Plazas2	0.9367	0.5266	0.5795
Zona 1	–0.4042	0.1448	0.1594
Zona 2	0.0712	0.1215	0.1338
Zona 3	–0.1598	0.1039	0.1144
Zona 4	–0.3329	0.1268	0.1396
Edad 1	–0.6710	0.2901	0.3193
Edad 2	–0.1979	0.2956	0.3253
Edad 3	0.1847	0.3902	0.4294
Edad 4	–0.1591	0.2968	0.3267
Edad 5	–0.6559	0.2254	0.2480
Potencia	0.0031	0.0013	0.0014

5.4.6. Regresión basada en distancias sobre la frecuencia de siniestralidad

Hacemos uso del índice de similaridad de Gower, (4.8). Damos tratamiento cuantitativo a la potencia y al Bonus|Malus. Damos tratamiento cualitativo al sexo, a la zona de circulación, a la forma de pago, a la Edad|Anticarn y al número de plazas. Utilizamos B = 712 muestras para realizar los cálculos relacionados con la metodología *bootstrap* en la estimación de los valores de probabilidad del proceso de selección.

Proceso de selección paso a paso

Fases de introducción:

[28] Plazas 3, Zona 5 y Edad 6.

Variable	*p*-valor	*p*-valor|F(1)	*p*-valor |F(1)F(2)	*p*-valor|F(1)...F(3)	*p*-valor|F(1)...F(4)
Pago	0.5133	**0.5133**	--------	--------	--------
Sexo	0.9773	0.9743	0.9040	0.9226	0.9266
Zona	0.5560	0.5600	0.5706	**0.5720**	--------
Plazas	**0.4853**	--------	--------	--------	--------
Edad|Anticarn	0.6586	0.6586	0.6693	0.6626	0.6640
Bonus|Malus	0.5693	0.5680	0.5707	0.5760	**0.5613**
Potencia	0.5226	0.5453	**0.5426**	--------	--------
	F(1) = **Plazas**	F(2) = **Pago**	F(3) = **Potencia**	F(4) = **Zona**	F(5) = **Bonus|Malus**

Fases de eliminación:

Variable	*p*-valores
F(1)	**0.4853**
F(2) | F(1)	**0.5133**
F(1) | F(2)	0.4840
F(3) | F(1)F(2)	**0.5426**
F(1) | F(2)F(3)	0.4853
F(2) | F(1)F(3)	0.5106
F(4) | F(1)F(2)F(3)	**0.5720**
F(1) | F(2)F(3)F(4)	0.4920
F(2) | F(1)F(3)F(4)	0.5066
F(3) | F(1)F(2)F(4)	0.5680
F(5) | F(1)F(2)F(3)F(4)	0.5613
F(1) | F(2)F(3)F(4)F(5)	0.4946
F(2) | F(1)F(3)F(4)F(5)	0.5746
F(3) | F(1)F(2)F(4)F(5)	**0.5920**
F(4) | F(1)F(2)F(3)F(5)	0.5866

Por lo que F[1] = F(3) = Potencia. Si realizamos la fase de introducción siguiente obtenemos que la siguiente variable a entrar es también la Potencia:

Variable	*p*-valor|F(1)F(2)F(4)F(5)
Pago	--------
Sexo	0.7933
Zona	--------
Plazas	--------
Edad|Anticarn	0.6813
Bonus|Malus	--------
Potencia	**0.5920**
	F(6) = **Potencia**

Por lo que finalizamos el proceso con:

Plazas, Pago, Bonus|Malus y Zona.

Realizamos el proceso de eliminación progresiva completo, obteniendo que las cuatro últimas variables a ser eliminadas coinciden con la selección final anterior:

Variable	*p*-valor	*p*-valor/F[1]	*p*-valor/F[1]F[2]	*p*-valor/F[1]...F[3]
Pago	0.6133	0.6133	0.5746	0.5600
Sexo	**0.8533**	--------	--------	--------
Zona	0.5800	0.5746	0.5866	**0.5826**
Plazas	0.4960	0.4933	0.4946	0.4906
Edad|Anticarn	0.6813	**0.6826**	--------	--------
Bonus|Malus	0.5760	0.5773	0.5613	0.5626
Potencia	0.5773	0.5786	**0.5920**	--------
	F[1] = **Sexo**	F[2] = **Edad|Anticarn**	F[3] = **Potencia**	F[4] = **Zona**

Variable	*p*-valor/F[1]...F[4]	*p*-valor/F[1]...F[5]	*p*-valor/F[1]...F[6]
Pago	0.5480	**0.5133**	--------
Sexo	--------	--------	--------
Zona	--------	--------	--------
Plazas	0.4866	0.4840	**0.4853**
Edad|Anticarn	--------	--------	--------
Bonus|Malus	**0.5706**	--------	--------
Potencia	--------	--------	--------
	F[5] = **Bonus|Malus**	F[6] = **Pago**	F[7] = **Plazas**

Observando el proceso de eliminación decidimos que las variables seleccionadas incluirán finalmente a la potencia:

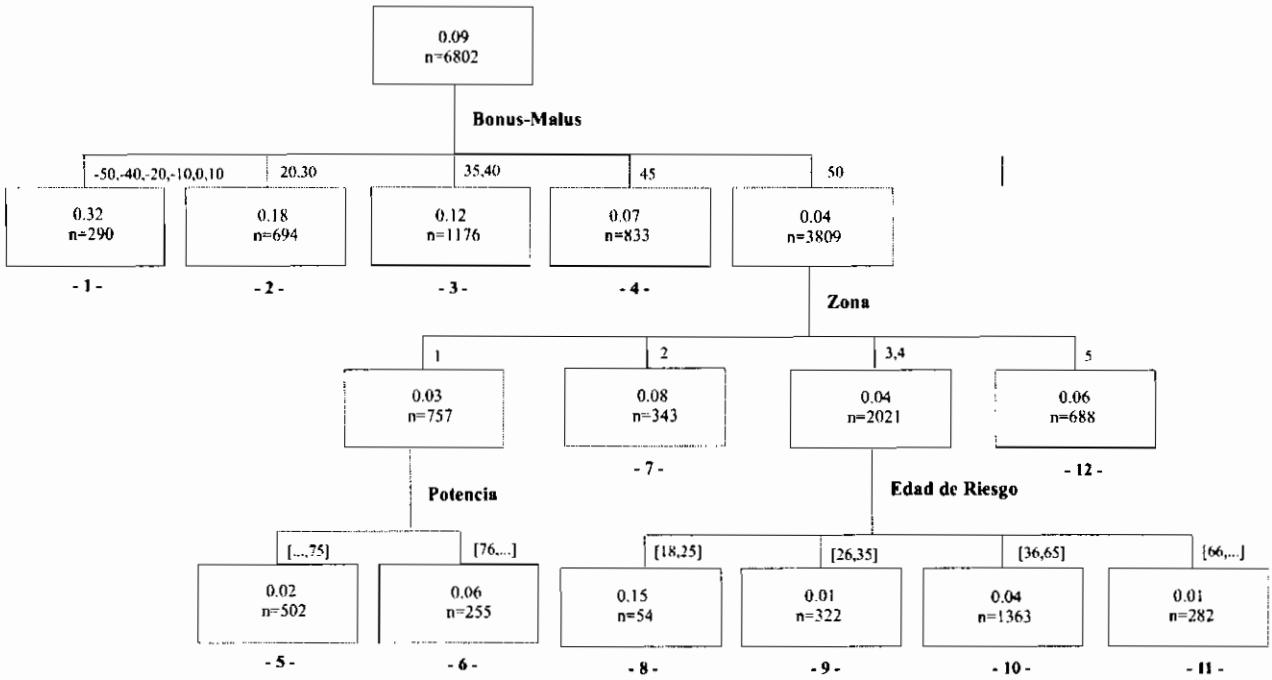
Plazas, Pago, Bonus|Malus, Zona y Potencia

El *p*-valor del contraste global, (4.58), es de 0.5893, un valor cuantitativamente elevado, pero en línea con lo esperado según las distribuciones de los *p*-valores simuladas. El coeficiente de determinación, (4.54), es de $R^2 = 0.2261$, de una magnitud parecida al resultante del MLG, $R^2 = 0.3481$.

Hemos calculado adicionalmente la medida de robustez B^2, obteniendo $B^2 = 0.9845$, cercana a 1, lo que indica que es un modelo robusto respecto a la predicción.

5.4.7. Análisis de segmentación sobre el número de siniestros

Hemos utilizado el SPSS CHAID ordinal, utilizando como respuesta el número de siniestros individual original. Los predictores utilizados han sido también los iniciales con sus clases discretas originales. Hemos definido como predictores monótonos a la potencia, el nivel de bonus-malus, la antigüedad del carnet y la edad de máximo riesgo. Como predictores libres a la zona, la forma de pago, el sexo y el número de plazas. Hemos determinado los niveles de significación tanto para la fase de agrupación de categorías como para la de selección del mejor predictor ambos del 5%. Fijamos un tamaño mínimo para analizar de 100 y uno para crear grupos de 50. Dejamos que opere hasta 12 niveles, con lo que no ponemos restricción. Obteniendo el siguiente árbol de segmentación:



Anexo 5.10. Anovas del número de siniestros con los factores discretos iniciales

Forma de pago

Descriptivos

NUMSIN

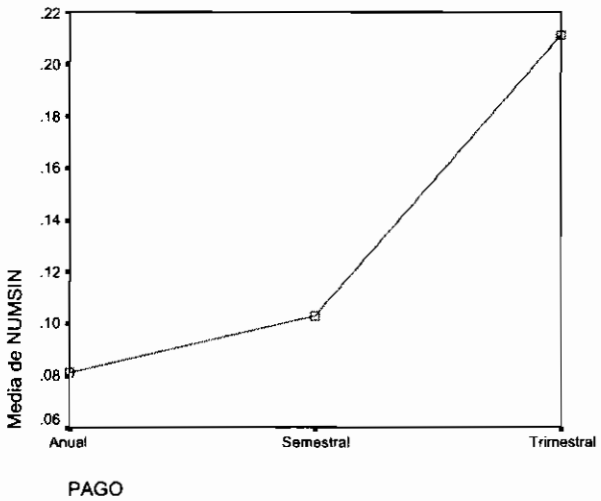
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	6080	8.11E-02	.32	4.08E-03	7.31E-02	8.91E-02	0	4
Semestral	613	.10	.33	1.33E-02	7.66E-02	.13	0	2
Trimestral	109	.21	.53	5.06E-02	.11	.31	0	3
Total	6802	8.51E-02	.32	3.93E-03	7.74E-02	9.28E-02	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	2.017	2	1.009	9.637	.000
Intra-grupos	711.697	6799	.105		
Total	713.714	6801			

Gráfico de las medias



Tipo de combustible

Descriptivos

NUMSIN

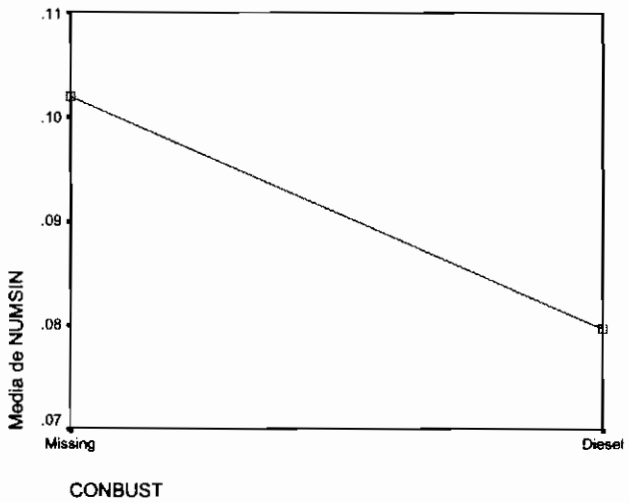
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Missing	1669	.10	.34	8.32E-03	8.55E-02	.12	0	3
Diesel	5133	7.97E-02	.32	4.44E-03	7.10E-02	8.84E-02	0	4
Total	6802	8.51E-02	.32	3.93E-03	7.74E-02	9.28E-02	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	.619	1	.619	5.907	.015
Intra-grupos	713.095	6800	.105		
Total	713.714	6801			

Gráfico de las medias



Sexo del primer conductor

Descriptivos

NUMSIN

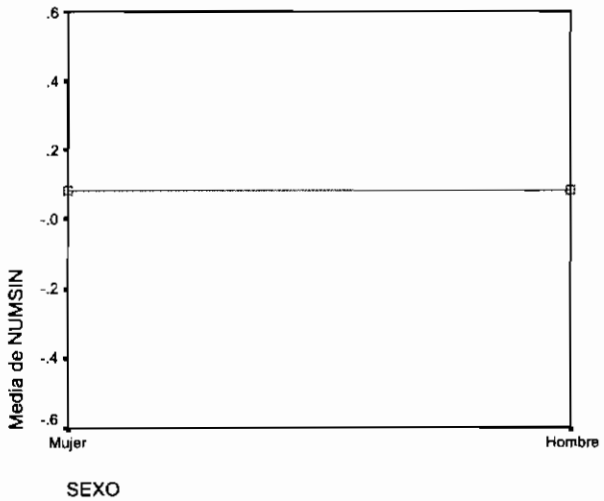
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Mujer	942	8.49E-02	.31	1.03E-02	6.48E-02	.11	0	3
Hombre	5860	8.52E-02	.33	4.25E-03	7.68E-02	9.35E-02	0	4
Total	6802	8.51E-02	.32	3.93E-03	7.74E-02	9.28E-02	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	4.215E-05	1	4.215E-05	.000	.984
Intra-grupos	713.714	6800	.105		
Total	713.714	6801			

Gráfico de las medias



Zona de circulación

Descriptivos

NUMSIN

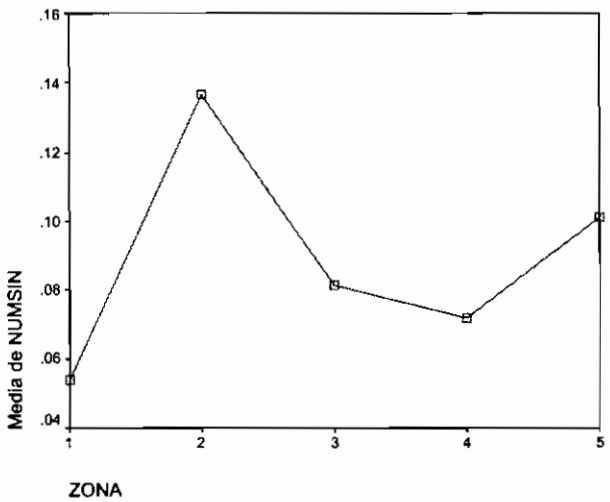
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	1091	.05	.253	.008	.04	.07	0	3
2	770	.14	.409	.015	.11	.17	0	3
3	2469	.08	.324	.007	.07	.09	0	4
4	1226	.07	.288	.008	.06	.09	0	4
5	1246	.10	.349	.010	.08	.12	0	3
Total	6802	.09	.324	.004	.08	.09	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	3.645	4	.911	8.722	.000
Intra-grupos	710.070	6797	.104		
Total	713.714	6801			

Gráfico de las medias



Número de plazas

Descriptivos

NUMSIN

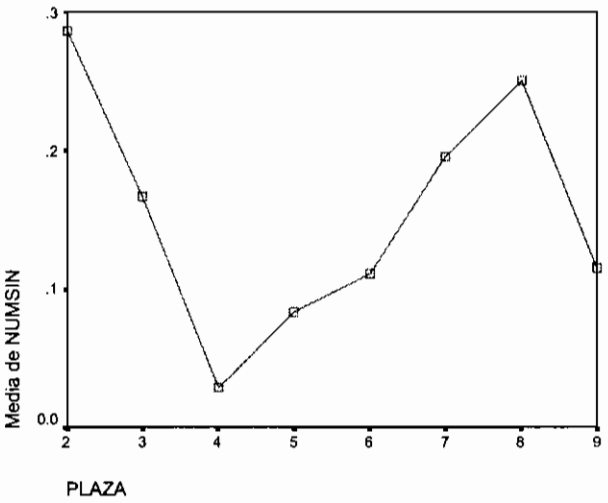
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
2	14	.29	.61	.16	-6.72E-02	.64	0	2
3	6	.17	.41	.17	-.26	.60	0	1
4	105	2.86E-02	.17	1.63E-02	-3.82E-03	6.10E-02	0	1
5	6499	8.31E-02	.32	3.94E-03	7.54E-02	9.08E-02	0	4
6	9	.11	.33	.11	-.15	.37	0	1
7	97	.20	.59	5.98E-02	7.72E-02	.31	0	4
8	20	.25	.55	.12	-7.46E-03	.51	0	2
9	52	.12	.38	5.25E-02	1.00E-02	.22	0	2
Total	6802	8.51E-02	.32	3.93E-03	7.74E-02	9.28E-02	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	2.753	7	.393	3.758	.000
Intra-grupos	710.961	6794	.105		
Total	713.714	6801			

Gráfico de las medias



Nivel de bonus-malus

Descriptivos

NUMSIN

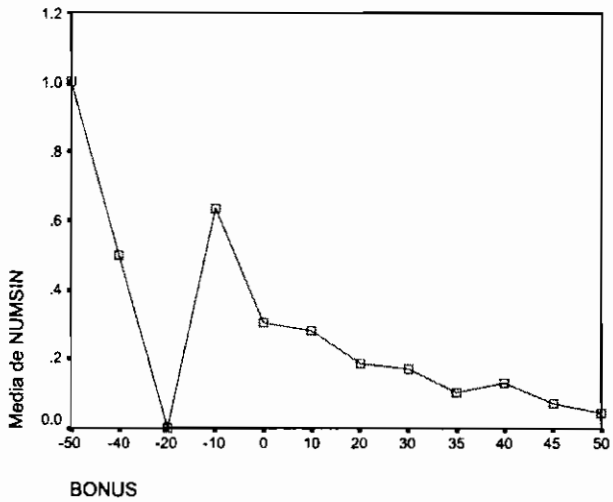
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
-50	2	1.00	.00	.00	1.00	1.00	1	1
-40	2	.50	.71	.50	-5.85	6.85	0	1
-20	3	.00	.00	.00	.00	.00	0	0
-10	22	.64	.66	.14	.34	.93	0	2
0	92	.30	.59	6.13E-02	.18	.43	0	3
10	169	.28	.54	4.13E-02	.20	.37	0	2
20	280	.19	.49	2.96E-02	.13	.24	0	3
30	414	.17	.48	2.35E-02	.12	.22	0	4
35	560	.10	.36	1.53E-02	7.17E-02	.13	0	4
40	616	.13	.39	1.58E-02	.10	.16	0	3
45	833	7.08E-02	.30	1.04E-02	5.04E-02	9.12E-02	0	3
50	3809	4.38E-02	.22	3.58E-03	3.68E-02	5.09E-02	0	2
Total	6802	8.51E-02	.32	3.93E-03	7.74E-02	9.28E-02	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	33.725	11	3.066	30.614	.000
Intra-grupos	679.990	6790	.100		
Total	713.714	6801			

Gráfico de las medias



Antigüedad de la póliza

Descriptivos

NUMSIN

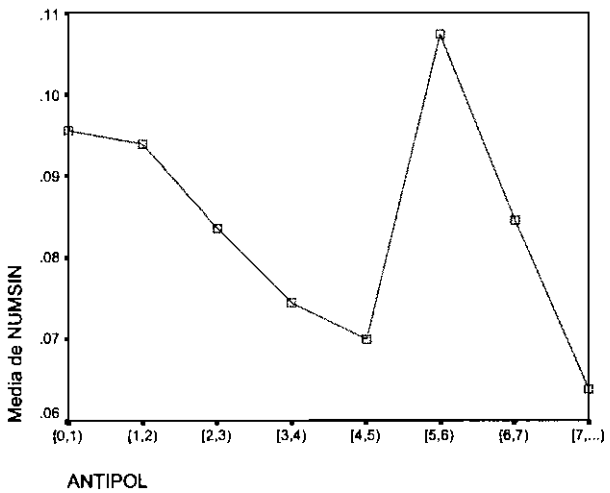
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[0,1)	272	9.56E-02	.33	2.00E-02	5.62E-02	.13	0	3
[1,2)	1937	9.40E-02	.34	7.71E-03	7.88E-02	.11	0	4
[2,3)	1389	8.35E-02	.33	8.89E-03	6.61E-02	.10	0	4
[3,4)	954	7.44E-02	.30	9.60E-03	5.56E-02	9.33E-02	0	3
[4,5)	600	7.00E-02	.29	1.17E-02	4.71E-02	9.29E-02	0	2
[5,6)	568	.11	.37	1.56E-02	7.67E-02	.14	0	2
[6,7)	580	8.45E-02	.31	1.30E-02	5.89E-02	.11	0	3
[7,...)	502	6.37E-02	.28	1.26E-02	3.90E-02	8.85E-02	0	3
Total	6802	8.51E-02	.32	3.93E-03	7.74E-02	9.28E-02	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	.942	7	.135	1.283	.254
Intra-grupos	712.772	6794	.105		
Total	713.714	6801			

Gráfico de las medias



Edad del primer conductor

Descriptivos

NUMSIN

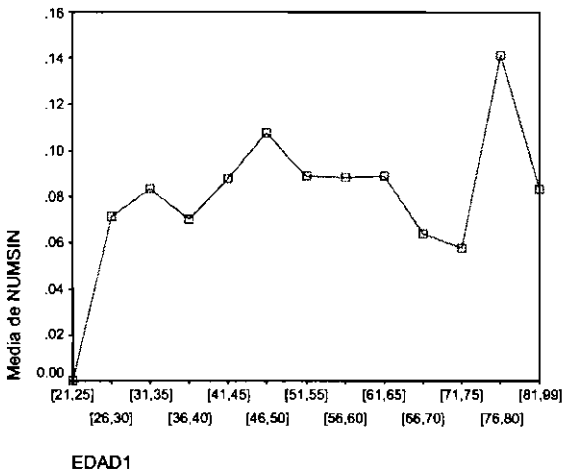
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[21,25]	3	.00	.000	.000	.00	.00	0	0
[26,30]	223	.07	.259	.017	.04	.11	0	1
[31,35]	708	.08	.328	.012	.06	.11	0	3
[36,40]	907	.07	.318	.011	.05	.09	0	4
[41,45]	841	.09	.348	.012	.06	.11	0	3
[46,50]	847	.11	.366	.013	.08	.13	0	4
[51,55]	831	.09	.313	.011	.07	.11	0	3
[56,60]	801	.09	.314	.011	.07	.11	0	3
[61,65]	672	.09	.324	.013	.06	.11	0	3
[66,70]	531	.06	.281	.012	.04	.09	0	2
[71,75]	294	.06	.248	.014	.03	.09	0	2
[76,80]	120	.14	.436	.040	.06	.22	0	2
[81,99]	24	.08	.282	.058	-.04	.20	0	1
Total	6802	.09	.324	.004	.08	.09	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1.558	12	.130	1.238	.250
Intra-grupos	712.156	6789	.105		
Total	713.714	6801			

Gráfico de las medias



Edad del segundo conductor

Descriptivos

NUMSIN

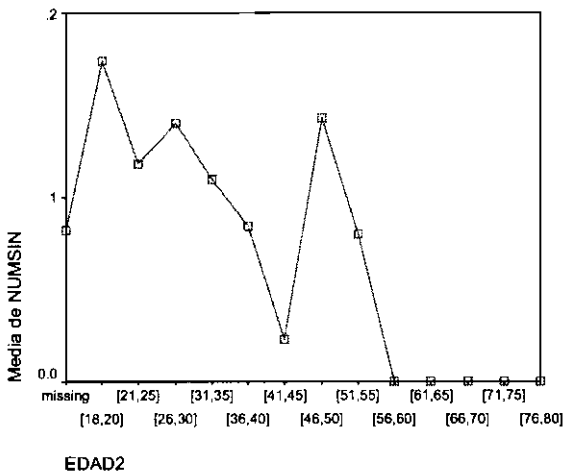
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
missing	6091	.08	.319	.004	.07	.09	0	4
[18,20]	46	.17	.383	.057	.06	.29	0	1
[21,25]	177	.12	.341	.026	.07	.17	0	2
[26,30]	164	.14	.427	.033	.07	.21	0	3
[31,35]	100	.11	.399	.040	.03	.19	0	3
[36,40]	71	.08	.327	.039	.01	.16	0	2
[41,45]	43	.02	.152	.023	-.02	.07	0	1
[46,50]	42	.14	.472	.073	.00	.29	0	2
[51,55]	25	.08	.277	.055	-.03	.19	0	1
[56,60]	18	.00	.000	.000	.00	.00	0	0
[61,65]	13	.00	.000	.000	.00	.00	0	0
[66,70]	8	.00	.000	.000	.00	.00	0	0
[71,75]	3	.00	.000	.000	.00	.00	0	0
[76,80]	1	.00	.	.	.	.	0	0
Total	6802	.09	.324	.004	.08	.09	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1.789	13	.138	1.312	.197
Intra-grupos	711.926	6788	.105		
Total	713.714	6801			

Gráfico de las medias



Edad de máximo riesgo

Descriptivos

NUMSIN

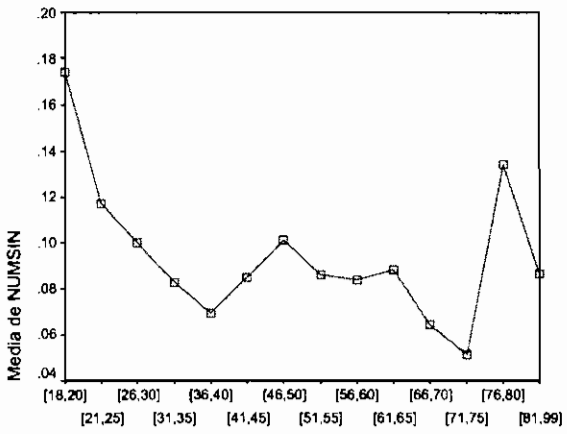
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[18,20]	46	.17	.383	.057	.06	.29	0	1
[21,25]	179	.12	.340	.025	.07	.17	0	2
[26,30]	379	.10	.342	.018	.07	.13	0	3
[31,35]	761	.08	.324	.012	.06	.11	0	3
[36,40]	926	.07	.318	.010	.05	.09	0	4
[41,45]	810	.09	.346	.012	.06	.11	0	3
[46,50]	788	.10	.363	.013	.08	.13	0	4
[51,55]	720	.09	.313	.012	.06	.11	0	3
[56,60]	705	.08	.306	.012	.06	.11	0	3
[61,65]	600	.09	.312	.013	.06	.11	0	2
[66,70]	481	.06	.278	.013	.04	.09	0	2
[71,75]	272	.05	.221	.013	.03	.08	0	1
[76,80]	112	.13	.414	.039	.06	.21	0	2
[81,99]	23	.09	.288	.060	-.04	.21	0	1
Total	6802	.09	.324	.004	.08	.09	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1.877	13	.144	1.377	.162
Intra-grupos	711.837	6788	.105		
Total	713.714	6801			

Gráfico de las medias



Antigüedad del carnet del primer conductor

Descriptivos

NUMSIN

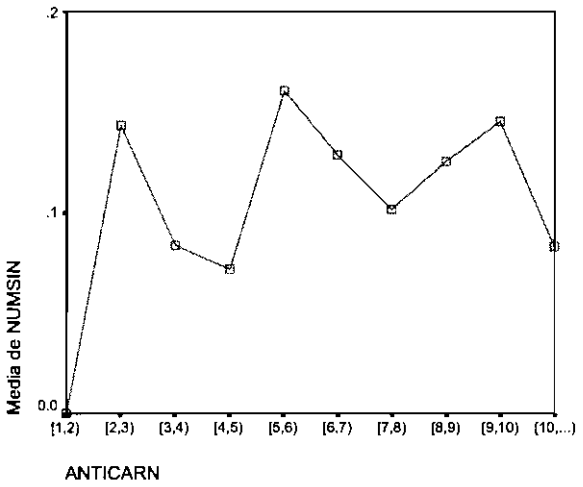
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[1,2)	3	.00	.00	.00	.00	.00	0	0
[2,3)	7	.14	.38	.14	-.21	.49	0	1
[3,4)	12	8.33E-02	.29	8.33E-02	-.10	.27	0	1
[4,5)	14	7.14E-02	.27	7.14E-02	-8.29E-02	.23	0	1
[5,6)	25	.16	.37	7.48E-02	5.55E-03	.31	0	1
[6,7)	39	.13	.34	5.42E-02	1.84E-02	.24	0	1
[7,8)	69	.10	.30	3.66E-02	2.84E-02	.17	0	1
[8,9)	72	.13	.56	6.54E-02	-5.43E-03	.26	0	4
[9,10)	124	.15	.44	3.92E-02	6.76E-02	.22	0	3
[10,...)	6437	8.28E-02	.32	3.96E-03	7.50E-02	9.06E-02	0	4
Total	6802	8.51E-02	.32	3.93E-03	7.74E-02	9.28E-02	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	.875	9	9.720E-02	.926	.501
Intra-grupos	712.840	6792	.105		
Total	713.714	6801			

Gráfico de las medias



Antigüedad del vehículo

Descriptivos

NUMSIN

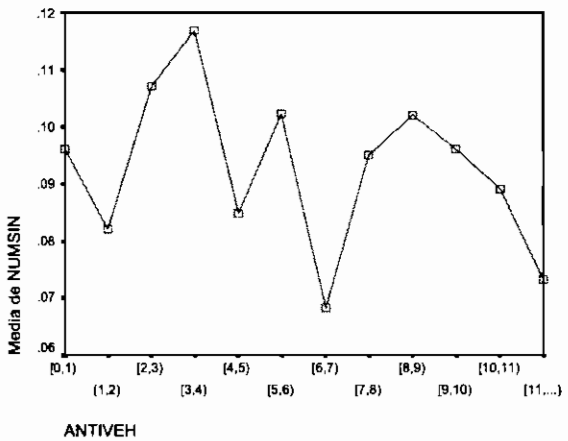
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[0,1)	52	9.62E-02	.36	4.96E-02	-3.38E-03	.20	0	2
[1,2)	365	8.22E-02	.38	1.97E-02	4.35E-02	.12	0	4
[2,3)	336	.11	.35	1.93E-02	6.91E-02	.15	0	3
[3,4)	291	.12	.42	2.48E-02	6.80E-02	.17	0	4
[4,5)	259	8.49E-02	.34	2.12E-02	4.31E-02	.13	0	2
[5,6)	303	.10	.33	1.92E-02	6.45E-02	.14	0	2
[6,7)	322	6.83E-02	.26	1.48E-02	3.93E-02	9.73E-02	0	2
[7,8)	484	9.50E-02	.35	1.60E-02	6.37E-02	.13	0	3
[8,9)	392	.10	.40	2.01E-02	6.25E-02	.14	0	3
[9,10)	479	9.60E-02	.33	1.53E-02	6.60E-02	.13	0	3
[10,11)	572	8.92E-02	.34	1.40E-02	6.16E-02	.12	0	3
[11,...)	2947	7.33E-02	.28	5.19E-03	6.31E-02	8.35E-02	0	3
Total	6802	8.51E-02	.32	3.93E-03	7.74E-02	9.28E-02	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1.284	11	.117	1.112	.346
Intra-grupos	712.430	6790	.105		
Total	713.714	6801			

Gráfico de las medias



Potencia

Descriptivos

NUMSIN

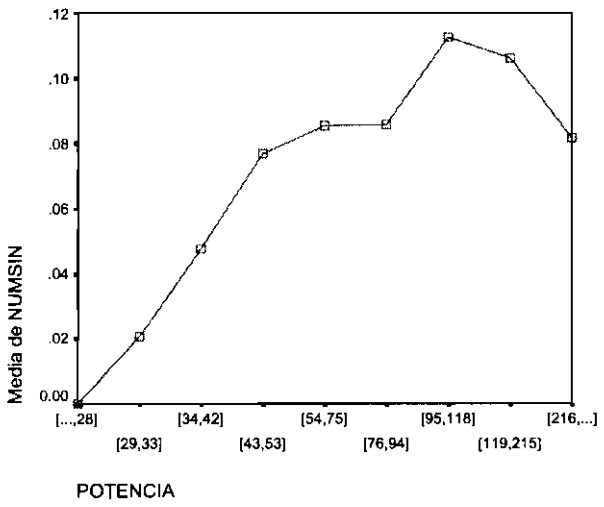
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[...,28]	41	.00	.00	.00	.00	.00	0	0
[29,33]	48	2.08E-02	.14	2.08E-02	-2.11E-02	6.27E-02	0	1
[34,42]	418	4.78E-02	.25	1.20E-02	2.43E-02	7.14E-02	0	3
[43,53]	886	7.67E-02	.31	1.04E-02	5.63E-02	9.72E-02	0	3
[54,75]	2927	8.54E-02	.32	6.00E-03	7.36E-02	9.72E-02	0	4
[76,94]	1303	8.60E-02	.31	8.70E-03	6.89E-02	.10	0	3
[95,118]	658	.11	.38	1.47E-02	8.36E-02	.14	0	3
[119,215]	472	.11	.37	1.71E-02	7.24E-02	.14	0	4
[216,...]	49	8.16E-02	.34	4.91E-02	-1.71E-02	.18	0	2
Total	6802	8.51E-02	.32	3.93E-03	7.74E-02	9.28E-02	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1.836	8	.230	2.190	.025
Intra-grupos	711.878	6793	.105		
Total	713.714	6801			

Gráfico de las medias



Valor del vehículo

Descriptivos

NUMSIN

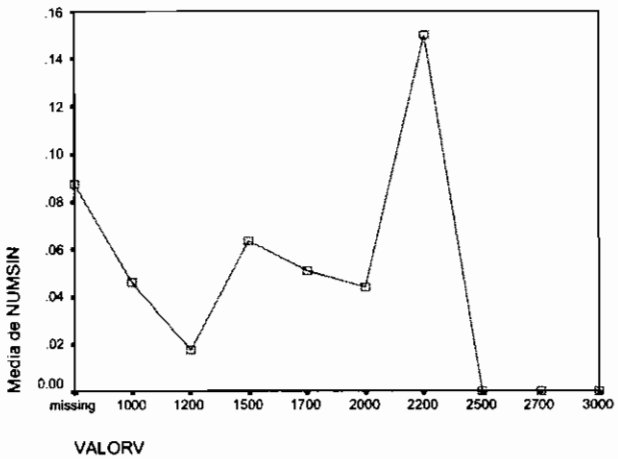
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
missing	6398	8.74E-02	.33	4.09E-03	7.93E-02	9.54E-02	0	4
1000	65	4.62E-02	.21	2.62E-02	-6.24E-03	9.85E-02	0	1
1200	58	1.72E-02	.13	1.72E-02	-1.73E-02	5.18E-02	0	1
1500	110	6.36E-02	.31	2.97E-02	4.83E-03	.12	0	2
1700	59	5.08E-02	.22	2.88E-02	-6.89E-03	.11	0	1
2000	68	4.41E-02	.27	3.27E-02	-2.11E-02	.11	0	2
2200	20	.15	.49	.11	-7.90E-02	.38	0	2
2500	13	.00	.00	.00	.00	.00	0	0
2700	7	.00	.00	.00	.00	.00	0	0
3000	4	.00	.00	.00	.00	.00	0	0
Total	6802	8.51E-02	.32	3.93E-03	7.74E-02	9.28E-02	0	4

ANOVA

NUMSIN

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	.891	9	9.898E-02	.943	.486
Intra-grupos	712.824	6792	.105		
Total	713.714	6801			

Gráfico de las medias



Anexo 5.11. Anovas de los factores cuantitativos con los factores cualitativos

Forma de pago con número de plazas

Descriptivos

PLAZAC

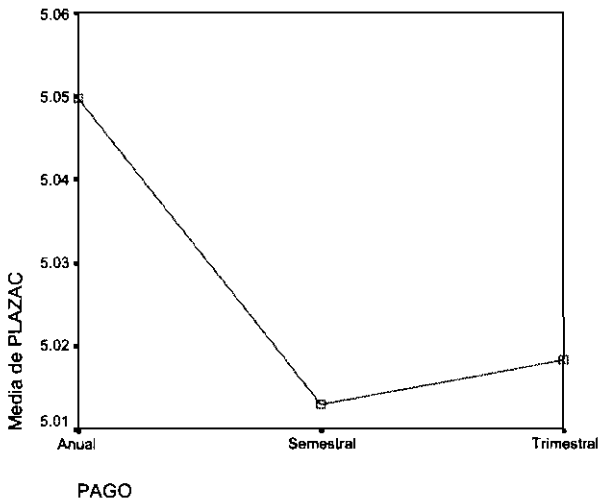
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	6080	5.0497	.5028	6.449E-03	5.0370	5.0623	2.00	9.00
Semestral	613	5.0131	.3917	1.582E-02	4.9820	5.0441	2.00	9.00
Trimestral	109	5.0183	.4078	3.906E-02	4.9409	5.0958	3.00	8.00
Total	6802	5.0459	.4925	5.972E-03	5.0342	5.0576	2.00	9.00

ANOVA

PLAZAC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	.831	2	.415	1.713	.181
Intra-grupos	1648.858	6799	.243		
Total	1649.689	6801			

Gráfico de las medias



Sexo del primer conductor con número de plazas

Descriptivos

PLAZAC

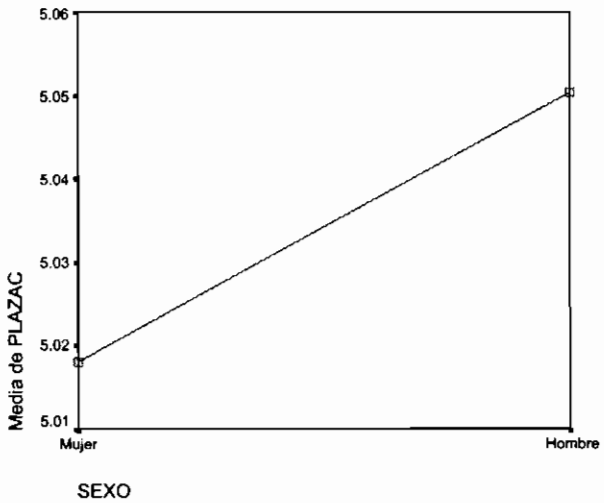
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Mujer	942	5.0180	.3271	1.066E-02	4.9971	5.0390	2.00	9.00
Hombre	5860	5.0503	.5140	6.715E-03	5.0372	5.0635	2.00	9.00
Total	6802	5.0459	.4925	5.972E-03	5.0342	5.0576	2.00	9.00

ANOVA

PLAZAC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	.846	1	.846	3.491	.062
Intra-grupos	1648.843	6800	.242		
Total	1649.689	6801			

Gráfico de las medias



Nivel de bonus-malus con forma de pago

Descriptivos

BONUSC

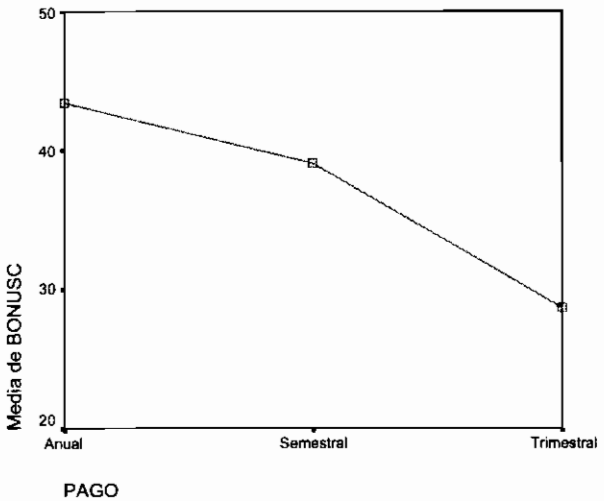
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Limite inferior	Límite superior		
Anual	6080	43.4770	11.0596	.1418	43.1989	43.7550	-50.00	50.00
Semestral	613	39.0865	13.0540	.5272	38.0510	40.1219	-10.00	50.00
Trimestral	109	28.6697	15.9229	1.5251	25.6466	31.6928	-10.00	50.00
Total	6802	42.8440	11.5560	.1401	42.5693	43.1187	-50.00	50.00

ANOVA

BONUSC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	32990.197	2	16495.099	128.140	.000
Intra-grupos	875217.304	6799	128.727		
Total	908207.501	6801			

Gráfico de las medias



Nivel de bonus-malus con sexo del primer conductor

Descriptivos

BONUSC

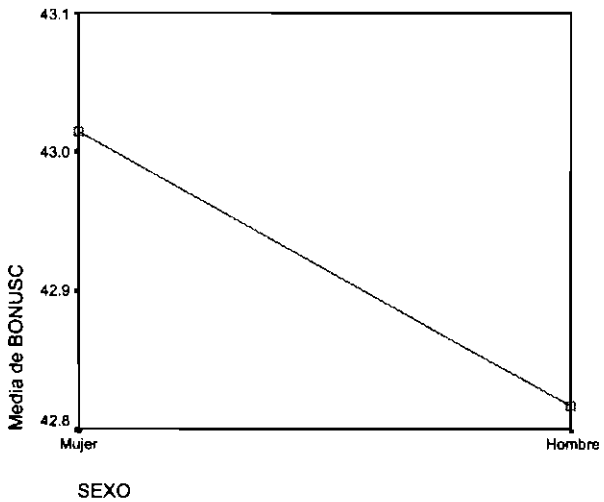
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Mujer	942	43.0149	11.2732	.3673	42.2940	43.7357	-50.00	50.00
Hombre	5860	42.8166	11.6015	.1516	42.5195	43.1137	-50.00	50.00
Total	6802	42.8440	11.5560	.1401	42.5693	43.1187	-50.00	50.00

ANOVA

BONUSC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	31.915	1	31.915	.239	.625
Intra-grupos	908175.586	6800	133.555		
Total	908207.501	6801			

Gráfico de las medias



Antigüedad de la póliza con forma de pago

Descriptivos

ANTIPOLC

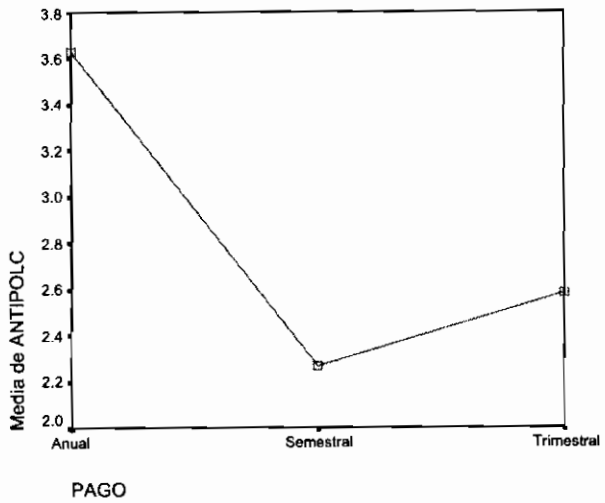
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	6080	3.6252	2.2027	2.825E-02	3.5698	3.6805	.50	8.50
Semestral	613	2.2700	1.4567	5.883E-02	2.1544	2.3855	.50	8.50
Trimestral	109	2.5826	1.7220	.1649	2.2556	2.9095	.50	8.50
Total	6802	3.4863	2.1768	2.639E-02	3.4346	3.5381	.50	8.50

ANOVA

ANTIPOLC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1113.154	2	556.577	121.628	.000
Intra-grupos	31112.574	6799	4.576		
Total	32225.728	6801			

Gráfico de las medias



Antigüedad de la póliza con sexo del primer conductor

Descriptivos

ANTIPOLC

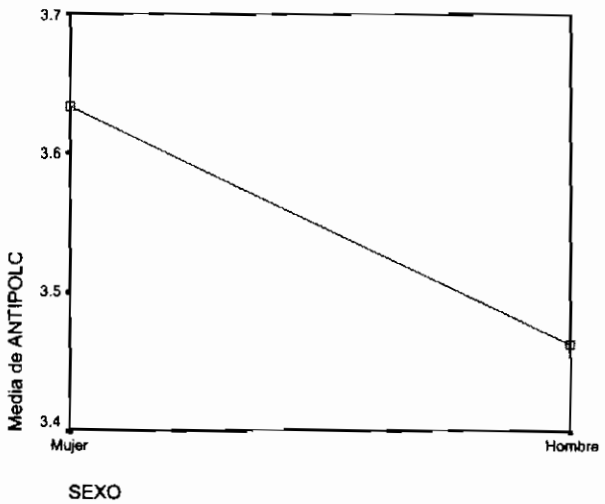
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Mujer	942	3.6338	2.1622	7.045E-02	3.4955	3.7720	.50	8.50
Hombre	5860	3.4626	2.1784	2.846E-02	3.4068	3.5184	.50	8.50
Total	6802	3.4863	2.1768	2.639E-02	3.4346	3.5381	.50	8.50

ANOVA

ANTIPOLC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	23.766	1	23.766	5.019	.025
Intra-grupos	32201.962	6800	4.736		
Total	32225.728	6801			

Gráfico de las medias



Edad del primer conductor con forma de pago

Descriptivos

EDAD1C

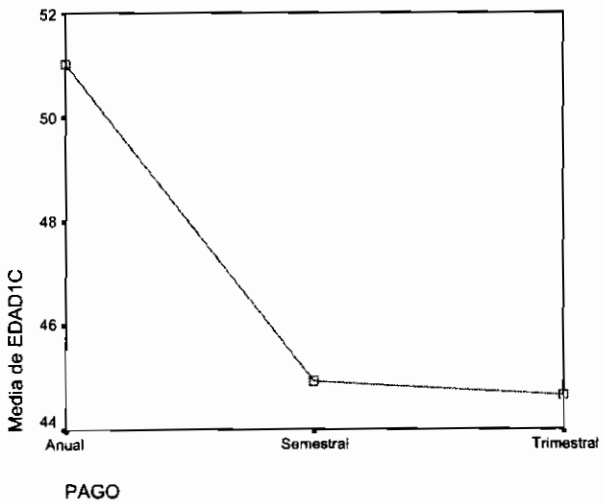
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	6080	51.0316	12.7089	.1630	50.7121	51.3511	23.00	81.00
Semestral	613	44.9168	11.8799	.4798	43.9745	45.8591	28.00	78.00
Trimestral	109	44.6514	11.0984	1.0630	42.5443	46.7585	23.00	73.00
Total	6802	50.3783	12.7523	.1546	50.0752	50.6814	23.00	81.00

ANOVA

EDAD1C

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	24454.262	2	12227.131	76.865	.000
Intra-grupos	1081533.4	6799	159.072		
Total	1105987.7	6801			

Gráfico de las medias



Edad con sexo del primer conductor

Descriptivos

EDAD1C

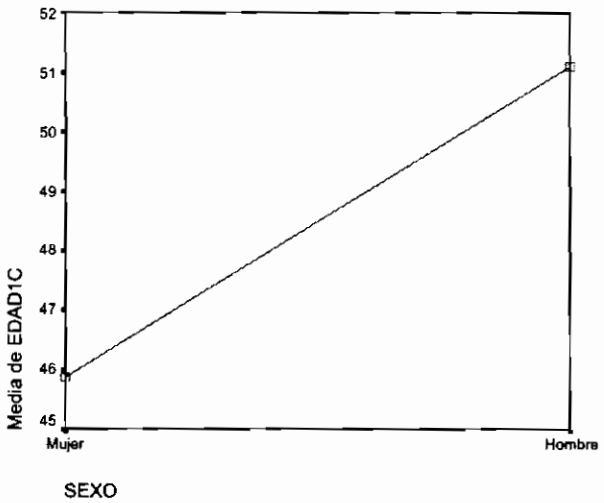
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Mujer	942	45.8715	11.7029	.3813	45.1233	46.6198	28.00	81.00
Hombre	5860	51.1027	12.7664	.1668	50.7758	51.4297	23.00	81.00
Total	6802	50.3783	12.7523	.1546	50.0752	50.6814	23.00	81.00

ANOVA

EDAD1C

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	22208.095	1	22208.095	139.341	.000
Intra-grupos	1083779.6	6800	159.379		
Total	1105987.7	6801			

Gráfico de las medias



Edad de máximo riesgo con forma de pago

Descriptivos

EDADXC

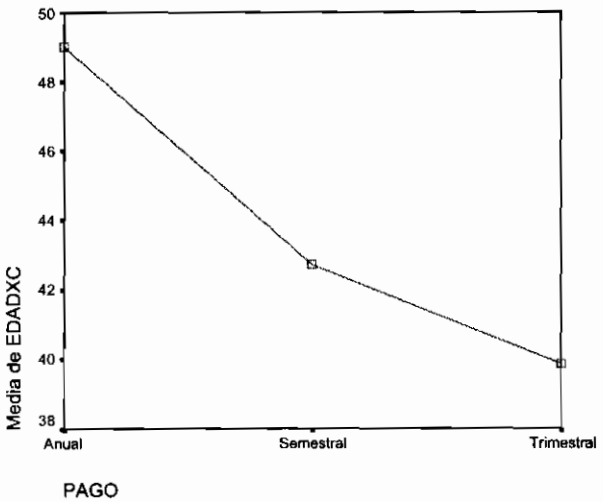
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	6080	49.0204	13.7898	.1769	48.6737	49.3671	19.00	81.00
Semestral	613	42.7129	12.4560	.5031	41.7249	43.7009	19.00	78.00
Trimestral	109	39.8532	12.3031	1.1784	37.5174	42.1891	19.00	73.00
Total	6802	48.3051	13.8115	.1675	47.9768	48.6333	19.00	81.00

ANOVA

EDADXC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	30067.415	2	15033.708	80.656	.000
Intra-grupos	1267282.6	6799	186.392		
Total	1297350.0	6801			

Gráfico de las medias



Edad de máximo riesgo con sexo del primer conductor

Descriptivos

EDADXC

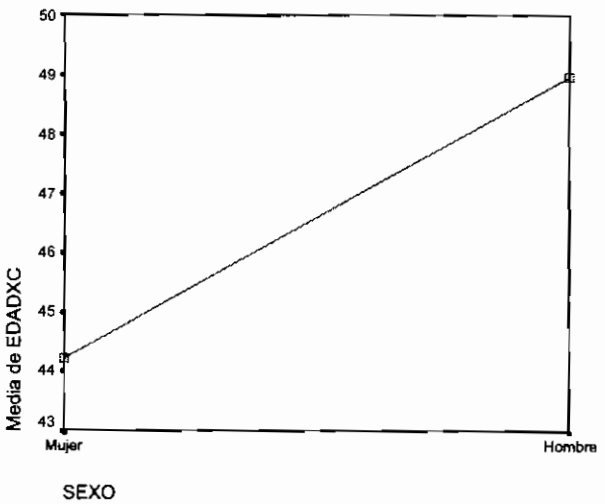
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Mujer	942	44.2197	12.2984	.4007	43.4334	45.0061	19.00	81.00
Hombre	5860	48.9618	13.9292	.1820	48.6051	49.3185	19.00	81.00
Total	6802	48.3051	13.8115	.1675	47.9768	48.6333	19.00	81.00

ANOVA

EDADXC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	18249.056	1	18249.056	97.016	.000
Intra-grupos	1279101.0	6800	188.103		
Total	1297350.0	6801			

Gráfico de las medias



Antigüedad del carnet del primer conductor con forma de pago

Descriptivos

ANTICARC

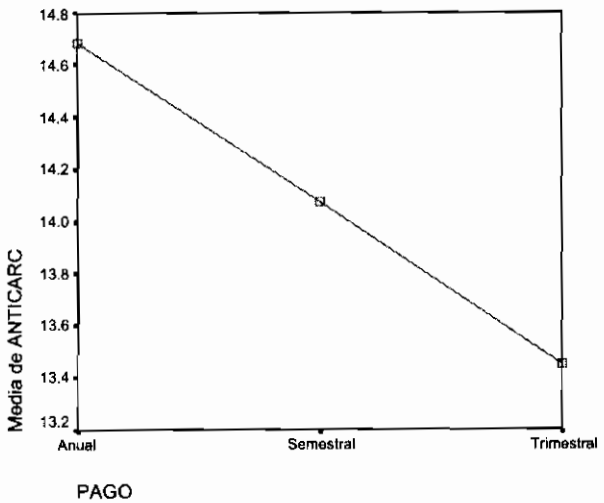
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	6080	14.6855	1.5017	1.926E-02	14.6478	14.7233	1.50	15.00
Semestral	613	14.0726	2.5932	.1047	13.8669	14.2783	2.50	15.00
Trimestral	109	13.4495	3.4761	.3329	12.7896	14.1095	1.50	15.00
Total	6802	14.6105	1.6927	2.052E-02	14.5702	14.6507	1.50	15.00

ANOVA

ANTICARC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	358.504	2	179.252	63.712	.000
Intra-grupos	19128.718	6799	2.813		
Total	19487.223	6801			

Gráfico de las medias



Antigüedad del carnet con sexo del primer conductor

Descriptivos

ANTICARC

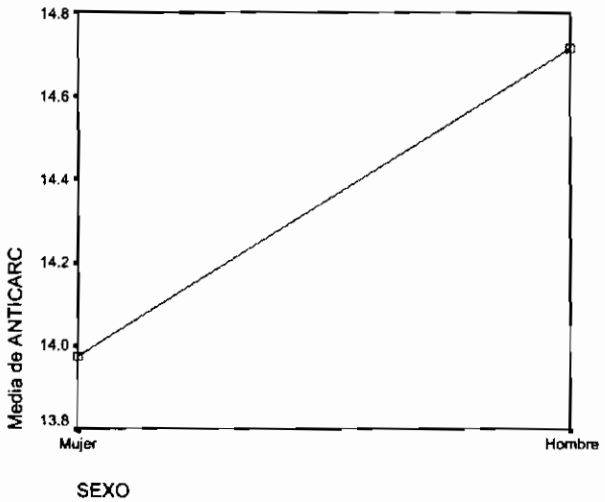
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Mujer	942	13.9740	2.6434	8.613E-02	13.8050	14.1430	1.50	15.00
Hombre	5860	14.7128	1.4588	1.906E-02	14.6754	14.7502	1.50	15.00
Total	6802	14.6105	1.6927	2.052E-02	14.5702	14.6507	1.50	15.00

ANOVA

ANTICARC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	442.970	1	442.970	158.168	.000
Intra-grupos	19044.253	6800	2.801		
Total	19487.223	6801			

Gráfico de las medias



Antigüedad del vehículo con forma de pago

Descriptivos

ANTIVEHC

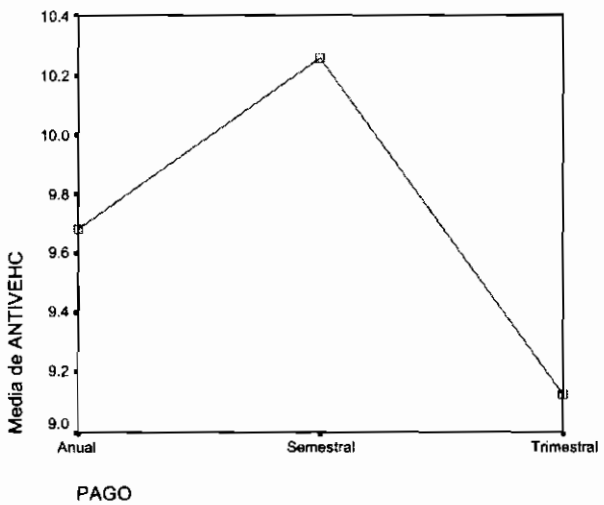
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	6080	9.6797	4.4005	5.643E-02	9.5691	9.7903	.50	14.00
Semestral	613	10.2553	4.1803	.1688	9.9237	10.5869	1.50	14.00
Trimestral	109	9.1239	4.6538	.4458	8.2403	10.0074	.50	14.00
Total	6802	9.7227	4.3884	5.321E-02	9.6183	9.8270	.50	14.00

ANOVA

ANTIVEHC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	224.224	2	112.112	5.830	.003
Intra-grupos	130748.315	6799	19.231		
Total	130972.539	6801			

Gráfico de las medias



Antigüedad del vehículo con sexo del primer conductor

Descriptivos

ANTIVEHC

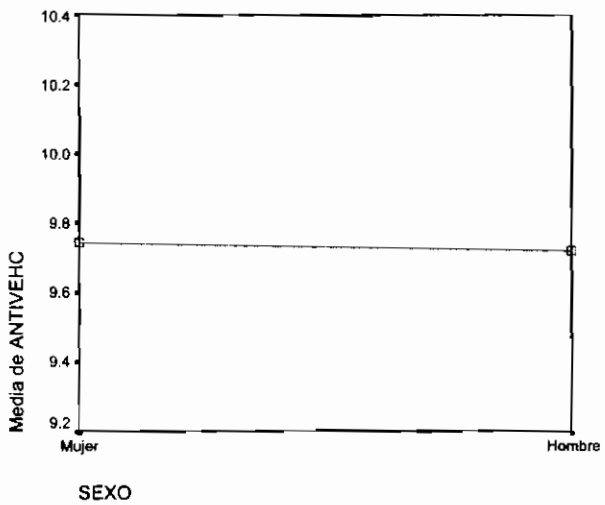
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Mujer	942	9.7426	4.3368	.1413	9.4653	10.0199	.50	14.00
Hombre	5860	9.7195	4.3970	5.744E-02	9.6069	9.8321	.50	14.00
Total	6802	9.7227	4.3884	5.321E-02	9.6183	9.8270	.50	14.00

ANOVA

ANTIVEHC

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	.434	1	.434	.023	.881
Intra-grupos	130972.105	6800	19.261		
Total	130972.539	6801			

Gráfico de las medias



Potencia con forma de pago

Descriptivos

POTENCON

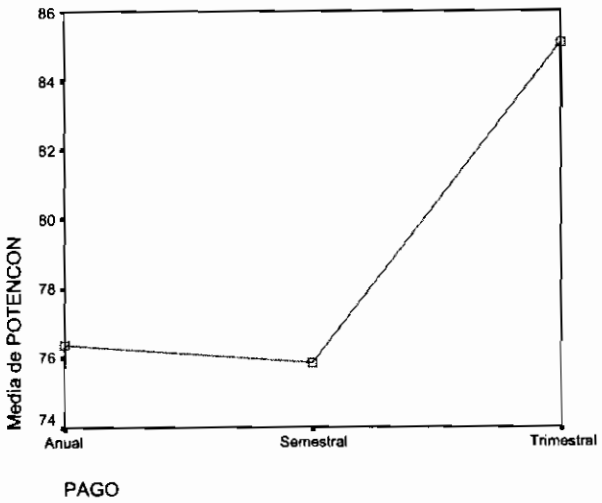
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	6080	76.3750	33.5350	.4301	75.5319	77.2181	25.00	220.00
Semestral	613	75.8842	30.9727	1.2510	73.4275	78.3409	25.00	220.00
Trimestral	109	85.0963	30.1354	2.8864	79.3749	90.8178	38.00	167.00
Total	6802	76.4705	33.2744	.4035	75.6796	77.2614	25.00	220.00

ANOVA

POTENCON

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	8376.325	2	4188.162	3.786	.023
Intra-grupos	7521614.0	6799	1106.282		
Total	7529990.3	6801			

Gráfico de las medias



Potencia con sexo del primer conductor

Descriptivos

POTENCON

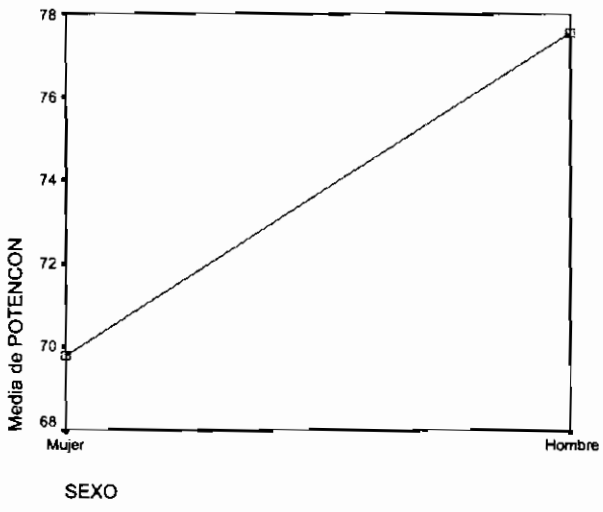
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Mujer	942	69.7818	26.9440	.8779	68.0590	71.5047	25.00	220.00
Hombre	5860	77.5457	34.0625	.4450	76.6734	78.4180	25.00	220.00
Total	6802	76.4705	33.2744	.4035	75.6796	77.2614	25.00	220.00

ANOVA

POTENCON

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	48918.177	1	48918.177	44.465	.000
Intra-grupos	7481072.2	6800	1100.158		
Total	7529990.3	6801			

Gráfico de las medias



Anexo 5.12. Anovas de la frecuencia de siniestralidad con los factores discretizados

Forma de pago

Descriptivos

YMEDIA

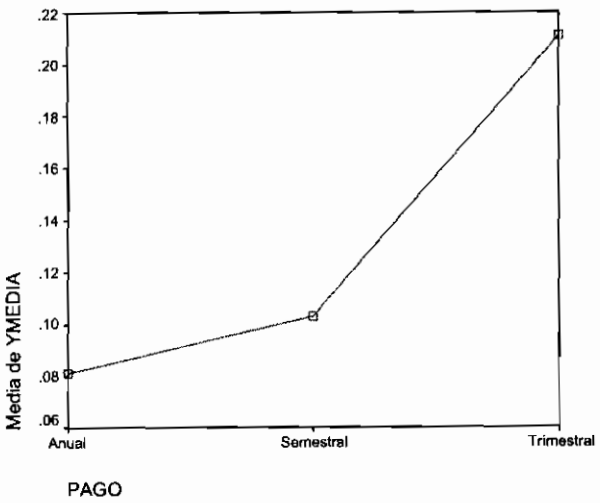
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Anual	6080	.0811	.13660	.00175	.0777	.0845	.00	4.00
Semestral	613	.1028	.19208	.00776	.0875	.1180	.00	2.00
Trimestral	109	.2110	.46544	.04458	.1226	.2994	.00	3.00
Total	6802	.0851	.15406	.00187	.0815	.0888	.00	4.00

ANOVA

YMEDIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	2.017	2	1.009	43.025	.000
Intra-grupos	159.401	6799	.023		
Total	161.419	6801			

Gráfico de las medias



Sexo del primer conductor

Descriptivos

YMEDIA

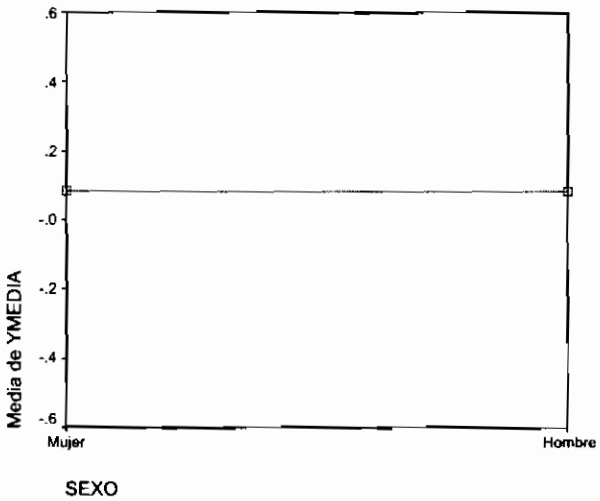
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
Mujer	942	.0849	.19639	.00640	.0724	.0975	.00	2.00
Hombre	5860	.0852	.14614	.00191	.0814	.0889	.00	4.00
Total	6802	.0851	.15406	.00187	.0815	.0888	.00	4.00

ANOVA

YMEDIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	.000	1	.000	.002	.966
Intra-grupos	161.419	6800	.024		
Total	161.419	6801			

Gráfico de las medias



Zona

Descriptivos

YMEDIA

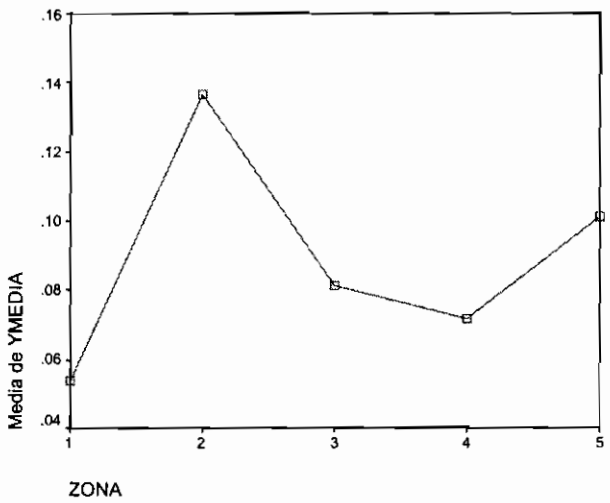
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	1091	.0541	.11815	.00358	.0471	.0611	.00	2.00
2	770	.1364	.17857	.00644	.1237	.1490	.00	1.00
3	2469	.0814	.15087	.00304	.0755	.0874	.00	4.00
4	1226	.0718	.13330	.00381	.0643	.0792	.00	1.00
5	1246	.1011	.17946	.00508	.0911	.1111	.00	2.00
Total	6802	.0851	.15406	.00187	.0815	.0888	.00	4.00

ANOVA

YMEDIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	3.645	4	.911	39.252	.000
Intra-grupos	157.774	6797	.023		
Total	161.419	6801			

Gráfico de las medias



Plazas

Descriptivos

YMEDIA

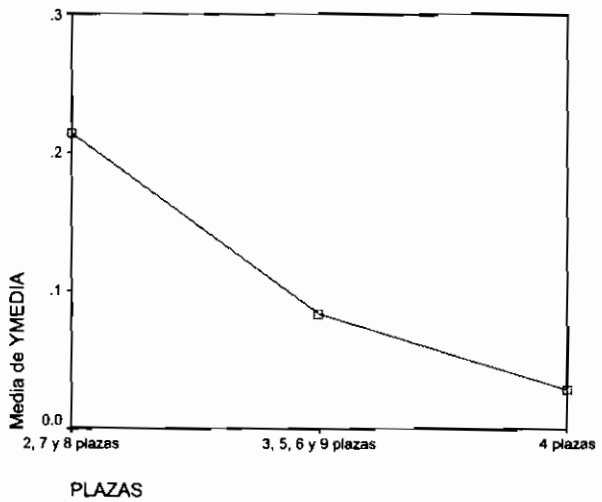
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
2, 7 y 8 plazas	131	.2137	.51004	.04456	.1256	.3019	.00	4.00
3, 5, 6 y 9 plazas	6566	.0835	.13683	.00169	.0801	.0868	.00	3.00
4 plazas	105	.0286	.14426	.01408	.0007	.0565	.00	1.00
Total	6802	.0851	.15406	.00187	.0815	.0888	.00	4.00

ANOVA

YMEDIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	2.521	2	1.261	53.935	.000
Intra-grupos	158.898	6799	.023		
Total	161.419	6801			

Gráfico de las medias



Bonus|Malus

Descriptivos

YMEDIA

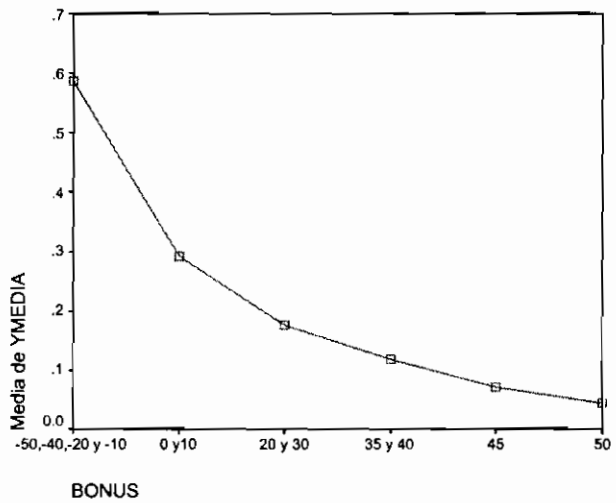
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
-50,-40,-20 y -10	29	.5862	.44935	.08344	.4153	.7571	.00	2.00
0 y10	261	.2912	.31908	.01975	.2523	.3301	.00	2.00
20 y 30	694	.1758	.20784	.00789	.1603	.1913	.00	2.00
35 y 40	1176	.1173	.16123	.00470	.1081	.1266	.00	4.00
45	833	.0708	.14654	.00508	.0609	.0808	.00	3.00
50	3809	.0438	.07063	.00114	.0416	.0461	.00	2.00
Total	6802	.0851	.15406	.00187	.0815	.0888	.00	4.00

ANOVA

YMEDIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	31.951	5	6.390	335.439	.000
Intra-grupos	129.467	6796	.019		
Total	161.419	6801			

Gráfico de las medias



Potencia

Descriptivos

YMEDIA

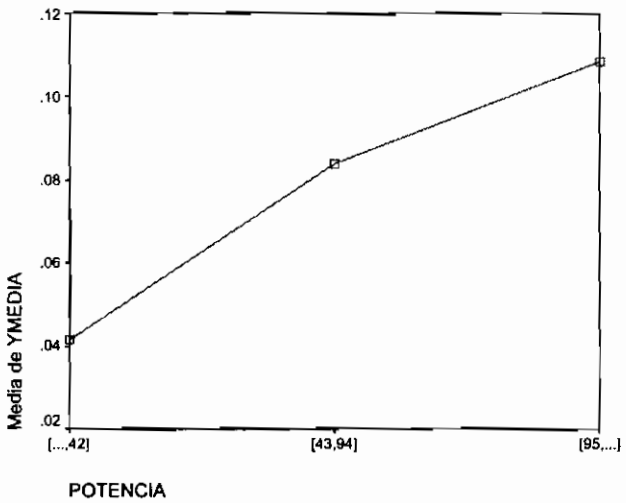
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
[...,42]	507	.0414	.09616	.00427	.0330	.0498	.00	1.00
[43,94]	5116	.0841	.14139	.00198	.0802	.0879	.00	3.00
[95,...]	1179	.1086	.21183	.00617	.0965	.1207	.00	4.00
Total	6802	.0851	.15406	.00187	.0815	.0888	.00	4.00

ANOVA

YMEDIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1.622	2	.811	34.511	.000
Intra-grupos	159.797	6799	.024		
Total	161.419	6801			

Gráfico de las medias



Edad|Anticarn

Descriptivos

YMEDIA

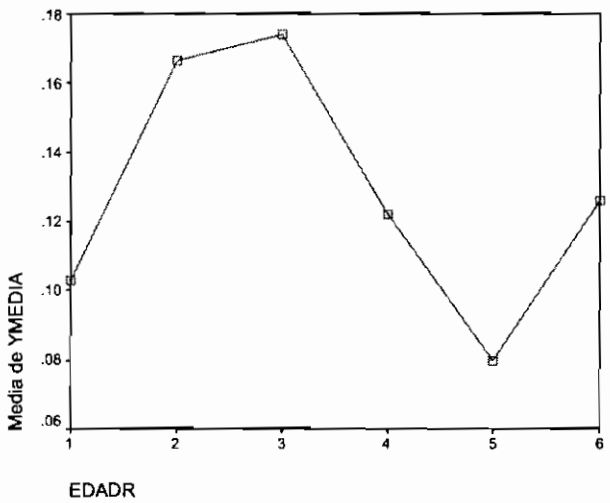
	N	Media	Desviación típica	Error típico	Intervalo de confianza para la media al 95%		Mínimo	Máximo
					Límite inferior	Límite superior		
1	233	.1030	.22558	.01478	.0739	.1321	.00	1.00
2	132	.1667	.50518	.04397	.0797	.2537	.00	4.00
3	46	.1739	.33140	.04886	.0755	.2723	.00	1.00
4	172	.1221	.24010	.01831	.0860	.1582	.00	1.00
5	6084	.0800	.11528	.00148	.0771	.0829	.00	2.00
6	135	.1259	.37266	.03207	.0625	.1894	.00	2.00
Total	6802	.0851	.15406	.00187	.0815	.0888	.00	4.00

ANOVA

YMEDIA

	Suma de cuadrados	gl	Media cuadrática	F	Sig.
Inter-grupos	1.932	5	.386	16.461	.000
Intra-grupos	159.487	6796	.023		
Total	161.419	6801			

Gráfico de las medias



Capítulo 6

Conclusiones: Sinopsis y Aportaciones

Destacamos, a continuación, las principales ideas de cada capítulo:

Capítulo 2

En este capítulo centramos el escenario actuarial de los seguros no vida remarcando, en especial, las peculiaridades del seguro del automóvil, y describimos el proceso de tarificación *a priori*, detallando sus fases y justificando la importancia de una correcta selección de los factores de riesgo como base del proceso.

Describimos los principios del cálculo de primas en lo que a la parte técnica actuarial se refiere: principios de equidad, solidaridad y suficiencia. El principio de equidad en el cálculo de la tarifa necesita de los factores de riesgo como variables exógenas del modelo. A través de ellos hemos de ser capaces de explicar la mayor parte de aleatoriedad del coste total esperado. Estos factores pueden ser seleccionados por separado respecto del número de siniestros o respecto de la cuantía de un siniestro.

Exponemos con detalle el estado actual del seguro del automóvil en España en lo que se refiere al efecto y al tratamiento de los datos en los estudios de tarificación, resaltando:

a) El gran avance que supone en la tarificación la constitución del fichero histórico de SINiestralidad de Conductores (SINCO). El fichero proporciona información objetiva sobre la siniestralidad del tomador del seguro referente a los últimos 5 años, por lo que permite a la entidad ajustar la prima que realmente corresponde. Empezó a funcionar en noviembre del 2000 y a medida que se van adhiriendo nuevas entidades y, las compañías van renovando las pólizas, pueden ser asignadas primas más equitativas a los riesgos realmente asegurados.

b) La problemática asociada al Convenio entre Entidades aseguradoras de automóviles para la Indemnización Directa de daños materiales a vehículos (CIDE), respecto a la selección de los factores de riesgo influyentes en la cuantía de un siniestro para daños materiales en el seguro de responsabilidad civil obligatoria.

Este convenio se implantó en España en enero de 1988, y tiene como objetivo acelerar la liquidación y pago entre compañías de los daños causados exclusivamente a los vehículos de aquellos accidentes de circulación que se producen por colisión directa entre dos de ellos. A grandes rasgos, la liquidación de siniestros vía CIDE es la siguiente: cuando una entidad es acreedora, anticipa la reparación en la cuantía peritada y recibe como base del recobro el coste medio establecido por el convenio para ese período en lugar del importe real; y cuando es deudora paga el coste medio independientemente de la cuantía real del siniestro, la cual no llega a conocer.

Por lo que respecto a la tarificación *a priori*, la aplicación de este convenio tiene mucha importancia: en responsabilidad civil de daños materiales el coste del siniestro es el pago total último que ha de realizar la entidad y un porcentaje elevado de los costes es de cuantía la que se establece en el coste medio sectorial del convenio (que encontramos detallado en la tabla 2.1 del anexo 2.1), que al fin y al cabo es el coste real para la entidad. Por ejemplo, en los datos de las carteras que hemos analizado, la mayoría de siniestros son de 75 000 pesetas para la cartera C1 y de 90 000 pesetas para la cartera C2. Debido a que no sabremos el coste real de los siniestros en los que la entidad es deudora, la selección de variables de tarifa respecto al coste de un siniestro no tiene sentido. Adicionalmente, a la entidad no le es de ningún interés saber cuales son las variables que explican el riesgo, ya que para ella, independientemente de las características de la póliza, el coste siempre será el que establezca el convenio.

Así, se debe suponer una cartera infinita en la que los costes por siniestro tiendan a la media sectorial y, entonces, la estabilidad económica de una entidad dependerá de que la media de los siniestros en los que es acreedora vía CIDE, también se corresponda con la media sectorial. Pero este hecho es no predecible ya que dependerá de las características de los causantes de los siniestros y no de sus propias pólizas.

De todo ello concluimos que, en el caso de daños materiales, respecto de una compañía, sólo será de interés el estudio de los factores de riesgo que influyen en el número de siniestros por póliza y no en la cuantía por siniestro.

Para poder llevar a cabo una correcta tarificación destacamos la necesidad de una buena base de datos como paso previo a la explotación estadística. Recomendamos partir de información desagregada para poder decidir su tratamiento dependiendo del objetivo del análisis y de la metodología empleada.

Capítulo 3

Dedicamos inicialmente atención al estudio de la asociación entre la variable univariante cuantitativa experiencia de siniestralidad (número de siniestros por póliza, o cuantía de un siniestro, o bien cuantía total de los siniestros) y cada factor de riesgo univariante, cuantitativo o cualitativo de forma individualizada. Para ello describimos medidas de asociación entre pares de variables, que también nos permiten estudiar las relaciones entre factores. Este estudio nos sirve para tener una primera aproximación de los factores que, uno a uno, están más asociados con la siniestralidad objeto de estudio.

Puesto que el objetivo es la selección del conjunto de variables de tarifa que mejor explique la estructura de riesgo, introducimos las técnicas de análisis estadístico multivariante, las cuales nos permiten organizar procesos de selección teniendo en cuenta simultáneamente el conjunto de factores. Algunas de las técnicas nos servirán para realizar una tarificación completa, y otras para cubrir sólo algunas de las fases de la obtención de la estructura de tarifa.

Citamos las diferentes metodologías de selección de variables de tarifa que se encuentran en la bibliografía actuarial, y dedicamos especial atención al funcionamiento técnico de dos de ellas por ser las más utilizadas: el análisis de segmentación y el modelo lineal generalizado. De ambas efectuamos un resumen de los fundamentos teóricos, prestando una especial atención a su aplicabilidad a la selección de variables de tarifa.

Análisis de segmentación:

El análisis de segmentación es una técnica estadística de cluster jerárquico divisivo que trabaja sobre datos tipo regresión. Las variables independientes deben ser categóricas, de tipo nominal u ordinal, y la variable dependiente puede ser cuantitativa o categórica. Se utiliza con fines exploratorios y descriptivos, con el objetivo básico de encontrar una clasificación de la población en grupos capaces de describir la variable dependiente de la mejor manera posible. El análisis de segmentación permite estudiar el efecto de las interacciones entre factores, aunque de forma jerárquica.

Puede ser utilizado para la predicción directamente a partir del árbol resultante simplemente con la media aritmética de la variable respuesta estudiada (número de siniestros o cuantía por siniestro). Y, adicionalmente, es útil como paso previo en la aplicación de otras técnicas especializadas para datos cualitativos como el análisis de correspondencias. En la aplicación 1 del capítulo 5 del trabajo lo

utilizamos para configurar los grupos de partida del modelo de credibilidad de Bülhmann-Straub en la estimación de primas. Del mismo modo lo podríamos utilizar para formar los grupos de tarifa que entrarían a modo de predictores en un modelo de regresión, sin tener que realizar entonces un proceso de selección de variables.

Modelo lineal generalizado:

El modelo lineal generalizado representa una extensión del modelo de regresión lineal clásico: por un lado la distribución del error no es necesariamente la Normal y, por otro, la función que relaciona el predictor lineal con la respuesta no es necesariamente la identidad. En la actualidad supone la metodología más propuesta en la bibliografía actuarial no sólo para la selección de variables de tarifa sino también para seguir cubriendo todo el proceso hasta la estimación de las primas puras.

Con este modelo de regresión es usual trabajar con datos agregados, por lo que el número de siniestros se estudia a través de la frecuencia de siniestralidad o número medio de siniestros por póliza en cada celda de la clasificación cruzada. Para el tratamiento de la frecuencia de siniestralidad, ponderamos los datos en el modelo con el número de pólizas con las que se han realizado las medias.

La distribución del error recomendada depende de la variable de siniestralidad objeto de estudio:

- Para la frecuencia de siniestralidad: se recomienda la distribución del error de Poisson ponderada que se basa en los supuestos de que el número individual de siniestros por póliza sigue una distribución de Poisson, de que las pólizas son independientes entre sí, y de que la agregación en las celdas ha supuesto una homogeneidad perfecta. Si alguna de las tres hipótesis anteriores no se cumple, este hecho se refleja en una infra o sobre dispersión en el modelo. Si el modelo es disperso no podemos averiguar exactamente a cuál de las causas es debido.

- Para la cuantía por siniestro: se recomiendan distribuciones que ofrezcan un rango positivo y con asimetría positiva, como son la Gamma y la Gaussiana Inversa.

La función de enlace, puede ser cualquier función monótona y diferenciable. Si utilizamos la identidad obtendremos una tarifa aditiva y si utilizamos la logarítmica obtendremos una multiplicativa.

Usualmente se indica que el modelo más apropiado para unos datos determinados es aquél que nos ofrece una menor desvianza. Como se intuye, tenemos diferentes maneras de reducirla:

- si variamos la función de enlace,
- si variamos la distribución del error,
- y/o si variamos los factores de riesgo incluidos en el predictor lineal.

Puesto que nosotros tenemos como objetivo la selección de variables de tarifa, fijaremos un enlace y un error y a partir de aquí realizaremos el proceso de selección: seleccionaremos los factores de riesgo para un modelo dado.

Para finalizar notamos que cada metodología incorpora unas hipótesis y, con ellas ventajas e inconvenientes. Generalmente se recomienda la utilización de varios métodos para decidir finalmente un "buen" subconjunto de variables tarificadoras. Los métodos utilizados deberían coincidir aproximadamente en los resultados obtenidos, y es importante extraer de cada uno, no sólo el resultado final, sino la información que se desprende durante los procesos respecto a relaciones entre factores seleccionados y no seleccionados.

Capítulo 4

Con el objetivo de proponer una herramienta estadística alternativa al resto de metodologías de selección de variables de tarifa, en este capítulo presentamos una metodología de selección de predictores en el modelo de regresión basado en distancias que permite trabajar directamente sobre factores potenciales de riesgo de tipo mixto.

Esta regresión fue inicialmente planteada por Cuadras (1989b) y Cuadras y Arenas (1990), y posteriormente desarrollada en Cuadras, Arenas y Fortiana (1996) y Cuadras y Fortiana (1998).

Brevemente, consiste en proyectar la respuesta continua en el espacio euclídeo obtenido mediante escalado multidimensional métrico a partir del conjunto de predictores. La información aportada por los factores de riesgo queda reflejada en una matriz de distancias sobre la que se opera. La regresión basada en distancias es una extensión del modelo clásico de regresión: si la distancia empleada es ℓ^2 y los predictores son cuantitativos se obtiene como caso particular el modelo de regresión lineal por mínimos cuadrados ordinarios.

En la tarificación, y en especial cuando se trabaja con modelos de regresión, es usual trabajar con datos agregados y por lo tanto ponderados, especialmente para el tratamiento del número de siniestros

mediante la frecuencia de siniestralidad. Por ello, para completar la metodología de selección a la tarificación, proponemos la versión ponderada de la regresión basada en distancias. En este capítulo la construimos mostrando su consistencia con la regresión basada en distancias usual. El modelo ponderado permite tratar también el caso heteroscedástico.

Planteamos un método de selección de predictores para la regresión basada en distancias. Para ello definimos las medidas y los tests estadísticos necesarios para la realización del proceso. Posteriormente construimos el proceso de selección paso a paso, análogamente que para el modelo lineal generalizado.

Puesto que no conocemos las distribuciones de los estadísticos de test para muestras finitas y, ciertamente, sería complicado obtenerlas, incluso aproximadamente, realizamos la estimación de los *p*-valores haciendo uso de la metodología *bootstrap*, a partir de estimaciones por simulación de las distribuciones de probabilidad de los estadísticos de los tests. Los modelos basados en distancias son especialmente adecuados para el empleo de *bootstrap*, pues el hecho que todas las interdistancias entre individuos de un remuestreo aparezcan ya en la matriz de distancias inicial nos permite vectorizar los remuestreos mediante matrices de multiplicidades, lo que es de gran economía computacional. Para la validación del modelo resultante empleamos diferentes criterios, incluidos los métodos de validación cruzada.

La metodología de selección que proponemos, además de cubrir la fase de selección de variables de tarifa, puede servir, si se desea, para completar la tarificación hasta la estimación de primas, al igual que ocurre con el modelo lineal generalizado.

Puesto que la regresión basada en distancias no está implementada en los programas usuales, hemos realizado la programación de prototipos con el lenguaje `octave`. El programa `octave` no tiene, en principio, restricción explícita en el número de pólizas ni en el de factores de riesgo. Sin embargo, es impracticable para una cartera real con un número de pólizas del orden de, por ejemplo, 10^6. En tal caso sería necesario:

⇒ Reemplazar los prototipos realizados en lenguaje interpretado (`octave`) por un programa ad-hoc en algún lenguaje compilado (fortran/C++) y

⇒ Realizar un muestreo, posiblemente estratificado, del conjunto de individuos a fin de evaluar el modelo y decidir los predictores significativos con un gasto computacional practicable.

Capítulo 5

En este capítulo hemos ilustrado el uso las tres metodologías estadísticas presentadas con detalle en el trabajo: el análisis de segmentación, el modelo lineal generalizado y la regresión basada en distancias.

Dado que el tema del trabajo tiene una aplicación real importante, el de ser el primer paso para la obtención del precio del seguro, no podíamos quedarnos en el plano teórico sino que ha sido necesario comprobar las dificultades o peculiaridades de los cálculos reales con las diferentes técnicas. Para ello hemos utilizado cuatro conjuntos de datos distintos que se desarrollan en las cuatro aplicaciones en que se divide el capítulo.

La primera aplicación utiliza unos datos sobre las cuantías de los siniestros de los impagos de préstamos de una entidad financiera, cuya característica principal es su simplicidad. La tercera aplicación utiliza unos datos sobre las cuantías de siniestros para la cobertura de daños propios del seguro del automóvil, también simples pero agregados y por lo tanto ponderados. Estas dos aplicaciones nos han servido principalmente para comprobar el correcto funcionamiento en la práctica de la metodología de selección propuesta para la regresión basada en distancias.

Las aplicaciones segunda y cuarta utilizan datos reales de España sobre carteras del seguro del automóvil. La segunda se refiere a la cuantía por siniestro para daños personales y la cuarta al número de siniestros para daños materiales. Al ser datos reales, se trata de ficheros complicados que reflejan la realidad de los estudios de tarificación que deben realizar las compañías de seguros. Analizar estos dos conjuntos de datos ha supuesto aplicar e ilustrar las metodologías con unos datos reales y por lo tanto complejos, teniendo en cuenta las características peculiares que en España tienen los seguros de automóviles.

Cabe destacar que las aplicaciones segunda, tercera y cuarta hacen referencia al seguro del automóvil, y que con cada una de ellas hemos tenido la posibilidad de analizar el comportamiento de las diferentes coberturas de dicho seguro (daños materiales, daños personales y daños propios). Corroborando que las variables de tarifa a emplear en la tarificación deben ser seleccionadas por separado para cada cobertura.

Aportaciones

Recogemos a continuación, a modo de resumen, las principales aportaciones del trabajo:

- En el capítulo 2, exponemos el estado actual del seguro del automóvil en España en aplicación a la tarificación *a priori*.
- En el capítulo 3, recopilamos y revisamos las metodologías de selección de variables de tarifa utilizadas históricamente. Describimos especialmente el funcionamiento teórico, en aplicación a la selección, del análisis de segmentación y del modelo lineal generalizado.
- En el capítulo 4, planteamos un método de selección de predictores para la regresión basada en distancias con el objetivo de constituir una herramienta alternativa al resto de metodologías de selección de variables de tarifa. Para ello:
  - Definimos las medidas y tests estadísticos apropiados para la regresión basada en distancias.
  - Realizamos la estimación de los p-valores con la metodología *bootstrap*, a partir de estimaciones por simulación de las distribuciones de probabilidad de los estadísticos de test.
  - Puesto que es usual trabajar con datos agregados y por lo tanto ponderados, en especial para la frecuencia de siniestralidad en estudio del número de siniestros, construimos la versión ponderada de la regresión basada en distancias, mostrando su consistencia con la regresión basada en distancias usual. Este modelo permite tratar también el caso heteroscedástico.
  - Para la aplicación de la metodología de selección propuesta implementamos (con `octave`) los programas informáticos que nos permiten realizar los cálculos de las aplicaciones del capítulo 5.
- Finalmente, en el capítulo 5, ilustramos la aplicabilidad de las tres metodologías estudiadas con detalle en el trabajo (análisis de segmentación, modelo lineal generalizado y regresión basada en distancias) con datos reales de carteras de seguros no vida, analizando las distintas dificultades empíricas y discutiendo las ventajas e inconvenientes de todas ellas.

Bibliografía

Agresti, A. (1984). *Analysis of ordinal categorical data*. John Wiley & Sons. New York.

Agresti, A. (1990). *Categorical data analysis*. John Wiley & Sons. New York.

Agsaa (1977). Exploitation du sondage automobile 1971 en France par une méthode d'analyse multidimensionelle. *ASTIN*[29] *Bulletin* **9:C**, 10-25.

Ajne, B. (1975). A note on the multiplicative ratemaking model. *ASTIN Bulletin* **8:2**, 144-153.

Ajne, B. (1980). Hipothesis testing in the multiplicative ratemaking model. *Transactions of the 21st International Congress of Actuaries* **2**, 1-9.

Ajne, B. (1986). Comparison of Some Methods to Fit a Multiplicative Tariff Structure to Observed Risk Data. *ASTIN Bulletin* **16:1**, 63-68.

Albrecht, P. (1983a). Parametric multiple regression risk models: Theory and statistical analysis. *Insurance: Mathematics and Economics* **2**, 49-66.

Albrecht, P. (1983b). Parametric multiple regression risk models: Connections with tariffication, especially in motor insurance. *Insurance: Mathematics and Economics* **2**, 113-117.

Almer, B. (1957). Risk analysis in theory and practical statistics. *Transactions of the 15-th International Congress of Actuaries* **2**, 314-349.

Andenberg, M. R. (1973). *Cluster analysis for applications*. Academic Press. New York.

[29] ***Actuarial STudies In Non-life insurance***

Andrade y Silva, J. M. (1989). An application of GLM to Portuguese motor insurance. *Proceedings of the XXI ASTIN Colloquium.* New York. pp. 633-649.

Ávila, C. A. (1996). *Una alternativa al análisis de segmentación basada en el análisis de hipótesis de independencia condicionada.* Tesis Doctoral. Universidad de Salamanca.

Bailey, R. A. y L. J. Simon (1960). Two studies in automobile insurance ratemaking. *ASTIN Bulletin* **1:4**, 192-217.

Bailey, R. A. (1963). Insurance rates with minimum bias. *Proceedings of the Casualty Actuarial Society* **50**, 4-11.

Baxter, L. A., S. M. Coutts y G. A. F. Ross (1980). Applications of linear models in motor insurance. *Transactions of the 21 st International Congress of Actuaries* **2**, 11-29.

Beirlant, J., Derveaux, V., de Meyer, A. M., Goovaerts, M. J., Labie, E. y B. Maenhoudt (1991). Statistical risk evaluation applied to (Belgian) car insurance. *Insurance: Mathematics and Economics* **10**, 289-302.

Bennet, M. (1978). Models in motor insurance. *Journal of the Institute of Actuaries Students' Society* **22**, 87-148.

Beomha, J. (1989). A comparative analysis of alternative pure premium models in the automobile risk classification system. *Journal of Risk and Insurance* **56:3**, 434-459.

Berg, P. (1980). On the loglinear Poisson and Gamma model. *ASTIN Bulletin* **1**, 35-40.

Bermúdez, Ll. y M. A. Pons (1997). Determinación del riesgo de impago en una cartera de préstamos según el tipo de cliente. *Matemática de las Operaciones Financieras 97'*. Publicaciones de la Universidad de Barcelona, pp. 291-308.

Beuthe, M. y Ph. van Namen (1975). La seléction des assurés et la détermination des primes d'assurances par l'analyse discriminante. *Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker* **75:2**, 137-156.

Boj, E., Claramunt, M. M. y J. Fortiana (2000). Una alternativa en la selección de los factores de riesgo a utilizar en el cálculo de primas. *Anales del Instituto de Actuarios Españoles*, Tercera Época **6**, 11-35.

Boj, E., Claramunt, M. M. y J. Fortiana (2001). Herramientas estadísticas para el estudio de perfiles de riesgo. *Anales del Instituto de Actuarios Españoles*, Tercera Época **7**, 59-89.

Boj, E., Claramunt, M. M., Fortiana, J. y A. Vidiella (2002). The use of distance-based regression and generalised linear models in the rate making process. An empirical study. *Mathematics Preprint Series, Institut de Matemàtica de la Universitat de Barcelona* **305**, 1-22.

Booth, P., Chadburn, R., Cooper, D., Haberman, S. y D. James (1999). *Modern Actuarial Theory and Practice.* Chapman & Hall. Boca Raton (California).

Borg, I. y P. Groenen (1997). *Modern multidimensional scaling: theory and applications.* Springer. New York.

Boskow, M. y R. J. Verrall (1994). Premium Rating by Geographical Area using Spatial Models. *ASTIN Bulletin* **24**, 131-143.

Breiman, L., Friedman, J. H., Olshen R. A. y J. S. Stone (1993). *Classification and Regression Trees.* Chapman & Hall. New York.

Brockman, M. J. y T. S. Wright (1992). Statistical Motor Rating: Making Effective Use of your Data. *Journal of the Institute of Actuaries* **119:3**, 457-543.

Bühlmann, H. (1967). Experience rating and credibility. *ASTIN Bulletin* **4**, 119-207.

Bühlmann, H. (1974). A comparison of three credibility formulae using multidimensional techniques. *ASTIN Bulletin* 7:**3**, 203-207.

Bühlmann, P. y H. Bühlmann (1999). Selection of credibility regression models. *ASTIN Bulletin* **29:2**, 245-270.

Byron, J. T., Morgan, T. J. y A. P. G. Ray (1995). Non-Uniqueness and Inversions in Cluster Analysis. *Applied Statistics* **44:1**, 117-134.

Cabral, M. A. y J. A. Garcia (1977). Study of factors influencing the risk and their relation to credibility theory. *ASTIN Bulletin* **9:C**, 84-104.

Cailliez, F. (1983). The analytical solution to the additive constant problem. *Psychometrika* **48**, 305-308.

Calatayud, J. y I. Martínez (1997). Pricing No Vida. En: *Manual de banca, finanzas y seguros*. Ediciones gestión 2000, S. A. Colección Universitaria Eserp, pp. 433-444.

Campbel, M. (1986). An integrated system for estimating the risk premium of individual car models in motor insurance. *ASTIN Bulletin* **16:2**, 165-184.

Carrasco, J. L. y M. A. Hernán (1993). *Estadística Multivariante en las Ciencias de la Vida. Fundamentos, Métodos y Aplicación*. Editorial Ciencia 3, S. L. Madrid.

Chang, L. y W. B. Fairley (1979). Pricing automobile insurance under multivariate classification. *Journal of Risk and Insurance* **46:1**, 75-93.

Cheng, S. H. y N. J. Higham (1998). A modified Cholesky Algorithm Based on a Symmetric Indefinite Factorization. *SIAM Journal of Matrix Analysis and Applications* **19:4**, 1097-1110.

Christensen, R. (1990). *Log-Linear Models*. Springer-Verlag. New York.

Cohen, A., Dupin, G. y Ch. Levy (1986). Tarification de l'incendie des risques industriels français par la méthode de la crédibilité. *ASTIN Bulletin* **16:2**, 149-164.

Conger, R. F. (1987). The construction of automobile rating territories in Massachusetts. *Proceedings of the Casualty Actuarial Society* **74:141**, 1-74.

Coutts, S. M. (1984a). Motor Insurance Rating - an Actuarial Approach. *Journal of the Institute of Actuaries* **111**, 87-148.

Coutts, S. M. (1984b). Motor premium rating. *Insurance: Mathematics and Economics* **3**, 73-96.

Cramér, H. (1946*). The elements of probability theory and some of its applications.* John Wiley. New York.

Cramer, J. S. (1964). Efficient Grouping Regression and Correlation in Engel Curve Analisys. *Journal of the American Statistical Association* **59**, 233-250.

Cuadras, C. M. (1989a). Distancias estadísticas. *Estadística Española* **30:119**, 295-378.

Cuadras, C. M. (1989b). Distance Analysis in discrimination and classification using both continuous and categorical variables. In: *Statistical Data Analysis and Inference* (Y. Dodge ed.), Elsiever Science Publisher. North-Holland. Amsterdam, pp. 459-474.

Cuadras, C. M. y C. Arenas (1990). A distance-based model for prediction with mixed data. *Communications in Statistics: Theory and Methods* **19**, 2261-2279.

Cuadras, C. M. (1991). *Problemas de probabilidades y estadística. Vol. 2: Inferencia estadística.* Colección: Estadística y Análisis de Datos. Publicaciones de la Universidad de Barcelona (PPU). Barcelona.

Cuadras, C. M. (1992). Some examples of distance based discrimination. *Biometrical Letters* **29**:1, 3-20.

Cuadras, C. M. y J. Fortiana (1993a). Aplicaciones de las distancias en estadística. *Qüestiió* **17:1**, 39-74.

Cuadras, C. M. y J. Fortiana (1993b). Continuous metric scaling and prediction. In: *Multivariate Analysis. Future Directions 2.* Cuadras and Rao, Eds. Elsevier. North Holland. Amsterdam, pp. 47-66.

Cuadras, C. M. (1996). *Métodos de Análisis Multivariante*. Ediciones de la Universidad de Barcelona (EUB). Barcelona.

Cuadras, C. M., Arenas, C. y J. Fortiana (1996). Some Computational aspects of a distance-based model for prediction. *Communications in Statistics: Simulation and Computation* **25:3**, 593-609.

Cuadras, C. M. y J. Fortiana (1998). Regresión basada en distancias generalizada. *XXIV Congreso Nacional de Estadística e Investigación Operativa*. Almería 20-23 de octubre de 1998.

Cuadras, C. M. (2003). *Models Estadístics Multivariants.* Colección de publicaciones del Departamento de Estadística de la Universidad de Barcelona.

Delicado, P. y M. del Río (1994). Bootstrapping the general linear hypothesis test. *Computational Statistics and Data Analysis* **18**, 305-316.

Delicado, P. y I. Placencia (2001). Comparing empirical distributions of p-values from simulations. *Communications in Statistics: Simulation and Computation* **30:2**, 403-422.

de Wit, G. W. (1986). Risk Theory, a Tool for Management. In: M. Goovaerts et al. eds., *Insurance and Risk Theory*. Reidel. Dordrecht-Boston, MA, pp. 7-17.

Derrig, R. A. y K. M. Ostaszewski (1995). Fuzzy techniques of pattern recognition in risk and claim classification. *Journal of Risk and Insurance* **62:3**, 447-482.

Diaconis, P. y B. Efron (1983). Métodos estadísticos intensivos por ordenador. *Investigación y Ciencia*, Julio 1983, 70-83.

Dionne, G. y Ch. Vanasse (1989). A generalization of automobile insurance rating models: The negative binomial distribution with a regression component. *ASTIN Bulletin* **19:2**, 199-212.

Dobson, A. J. (2001). *An Introduction to Generalized Linear Models*. Second Edition. Chapman & Hall. London.

Domènech, J. M. (1977). Bioestadística. Métodos estadísticos para investigadores. Editorial Herder. Barcelona, pp. 231-275.

Dorado, A. (1998). *Métodos de búsqueda de variables relevantes en análisis de segmentación: aportaciones desde una perspectiva multivariante*. Tesis Doctoral. Universidad de Salamanca.

Draper, N. R. y H. Smith (1981). *Applied Regression Analysis*. Second edition. John Wiley & Sons. New York, pp. 519-520.

Efron, B. y R. Tibshirani (1993). *An introduction to the bootstrap*. Chapman & Hall. London.

Escobar, M. (1992). *El Análisis de Segmentación: Concepto y Aplicaciones*. Estudios/Working Papers 31. Instituto Juan March de Estudios e Investigaciones. Madrid.

Estabrook, G. F. y D. J. Rogers (1966). A general method of taxonomic description for a computed similarity measure. *BioScience* **16**, 789-793.

Esteve, A. (2003). *Distancias estadísticas y relaciones de dependencia entre conjuntos de variables*. Tesis Doctoral. Universidad de Barcelona.

Fairley, W. B. y T. J. Tomberling (1981). Pricing automobile insurance under cross-classification of risks: evidence from New Yersey. *Journal of Risk and Insurance* **48:3**, 505-514.

Font, M. (1999). Tiempos difíciles para el ramo del automóvil. *Actuarios* **17**, Mayo/Junio.

Fortiana, J. y C. M. Cuadras (1994). Estimación bootstrap de la distancia entre poblaciones. *XXI Congreso Nacional de Estadística e Investigación Operativa.* Calella 18-21 de abril de 1994.

Fortiana, J. y C. M. Cuadras (1998). Generalized distance-based regression. *Classification & Psychometric Society Joint Meeting,* June 1998.

Fortiana, J. y A. Esteve (1999a). Términos de interacción en el modelo lineal basado en distancias. *VII Conferencia Española de Biometría.* Palma de Mallorca, marzo de 1999.

Fortiana, J. y A. Esteve (1999b). Interaction terms in the distance-based regression model. *11th European Meeting of the Psychometric Society.* Lüneburg 19-22, july 1999.

Franckx, E. (1974). Considérations sur les modèles d'avant project pour classes de tarif. *ASTIN Bulletin* **7:3**, 208-214.

Garrido y Comas J. J. (1987). Teoría general y derecho español de los seguros privados En: *Tratado general de seguros: teoría y práctica de los seguros privados* **I:II**. Editado por el Consejo General de los Colegios de Agentes y Corredores de Seguros de España.

Gogol, D. F. (1993). The value of information in insurance pricing. *Journal of Risk and Insurance* **60:1**, 119-128.

Goodman, L. A. (1979). Simple models for the analysis of association in cross-classifications having ordered categories. *Journal of the American Statistical Association* **74**, 537-552.

Goovaerts, M. J. y W. J. Hoogstad (1987). *Credibility theory.* Survey of Actuarial Science 4, Nationale-Nederlanden N. V. Rotterdam.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika* **53**, 325-338.

Gower, J. C. (1968). Adding a point to a vector diagrams in multivariate analysis. *Biometrika* **55**, 582-585.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics* **27**, 857-874.

Gower, J. C. (1982). Euclidean distance geometry. *Math. Scientist* **7**, 1-14.

Gower, J. C. y P. Legendre (1986). Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification* **3**, 5-48.

Gower, J. C. y S. A. Harding (1988). Nonlinear biplots. *Biometrika* **75**, 445-455.

Gower, J. C. (1992). Generalized biplots. *Biometrika* **79**, 475-493.

Greene, W. H. (1999). *Análisis Econométrico.* Tercera Edición. Prentice Hall Iberia S.R.L. Madrid.

Grünig, R. (1975). How to find the right subdivision into tariff classes. A numerical example. *ASTIN Bulletin* **8:2**, 261-263.

Gutiérrez, J., Rodríguez, V. y P. Santos (1995). *Técnicas cuantitativas (Estadística básica).* Oikos-Tau, S.L. Barcelona, pp. 22-35.

Haberman, S. y A. E. Renshaw (1996). Generalized Linear Models and Actuarial Science. *The Statistician* **45:4**, 407-436.

Haberman, S. y A. E. Renshaw (1998). Actuarial applications of Generalized Linear Models. En: *Statistics in Finance.* Hand, D. J. and S. D. Jacka (eds). Arnold Applications of Statistics. London-Sydney-Auckland, pp. 42-65.

Hallin, P. M. (1977). Méthodes statistiques de construction de tariff. *Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker* **77:2**, 160-175.

Hallin, P. M. y J. F. Ingenbleek (1981). Étude statistique de la probabilité de sinistre en assurance automobile. *ASTIN Bulletin* **12:1**, 40-56.

Harrington, S. E. y H. I. Doerpinghaus (1993). The economics and politics of automobile insurance rate classification. *Journal of Risk and Insurance* **60:1**, 59-84.

Hartigan, J. A. (1975). *Clustering Algorithms*. John Wiley and Sons. New York.

Hastie, T. J. y R. J. Tibshirani (1991). *Generalized Additive Models*. Second Edition. Chapman & Hall. London.

Haughton, D. y S. Oulabi (1983). Direct Marketing Modelling with CART and CHAID. *Journal of Direct Marketing* **7:3**, 16-26.

Hawkins, D. M. y G. V. Kass (1982a). *Topics in Applied Multivariate Analysis*. Ed. D M Hawkins. Cambridge University Press.

Hawkins, D. M. y G. V. Kass (1982b). Automatic Interaction Detection. In: *Topics in Applied Multivariate Analysis*. Ed. D M Hawkins. Cambridge University Press, pp. 269-302.

Hawkins, D. M. (1997). *FIRM: Formal Inference-based Recursive Modelling*. Technical Report Number 546, School of Statistics. University of Minnesota.

Heilmann, W. R. (1988). *Fundamentals of Risk Theory*. Verlag Versicherungswirtschaft e. V. Karlsruhe.

Hey, G. B. (1970). Statistics in non-life insurance. *Journal of the Royal Statistical Society, Series A* **133**, 56-85.

Hipp, Ch. (2000). *Least Squares, Generalized Linear Models and Credibility Theory with applications to Tariffication*. Working Paper. University of Karlsruhe.

Holler, K. D., Sommer, D. y G. Trahair (1999). Something old, something new in classificaiton ratemaking with a novel use of GLMs for credit insurance. *Casualty Actuarial Society Winter Forum.*

Hooge, L. D. (1974). Détermination de la fonction de structure d'une classe de tarif. *ASTIN Bulletin* **7:3**, 223-236.

Hossack, I. B., Pollard, J. H. y B. Zehnwirth (2001). *Introducción a la estadística con aplicaciones a los seguros generales*. Traducción de A. Vegas Montaner. Editorial MAPFRE, S. A. Madrid.

Hutchinson, T. P. y S. Rowell (1986). Points system for car insurance. *Insurance: Mathematics and Economics* **5**, 255-259.

Ingenbleek, J. F. y J. Lemaire (1988). What is a sports car. *ASTIN Bulletin* **18:2**, 175-188.

Jewell, W. S. (1975). Isotonic optimisation in tariff construction. *ASTIN Bulletin* **8:2**, 175-203.

Johnson, P. D. y G. B. Hey (1972). Statistical review of a motor insurance portfolio. *ASTIN Bulletin* **6:3**, 222-232.

Jørgensen, B. (1987). Exponential dispersion models (with discussion). *Journal of the Royal Statistical Society, Series B* **49**, 127-162.

Jørgensen, B. y M. C. Paes de Souza (1994). Fitting Tweedie's compound Poisson model to insurance claim data. *Scandinavian Actuarial Journal* **1**, 69-93.

Jung, J. (1968). On automobile insurance ratemaking. *ASTIN Bulletin* **5**, 41-48.

Kass, G. V. (1980). An explanatory technique for investigating large quantities of categorical data. *Applied Statistics* **29**, 119-127.

Kim, Ch. y S. Hwang (2000). Influential subsets on the variable selection. *Communications in Statistics: Theory and Methods*, **29:2**, 335-347.

Krzanowski, W. J. (1994). Ordination in the presence of group structure, for general multivariate data. *Journal of Classification* **11**, 195-207.

Lanteli, G. (1962). Novelties in Swedish Automobile insurance rating. *ASTIN Bulletin* **2:1**, 96-101.

Lebart, L., Morineau, A. y J. A. Fénelon (1985). *Tratamiento Estadístico de Datos (métodos y programas)*. Marcombo, Boixareu Editores. Barcelona-México.

Legendre, P. y A. Chodorowski (1977). A generalization of Jaccard's association coefficient for Q analysis of multi-state ecological data matrices. *Ekologia Polska* **25**, 297-308.

Lemaire, J. (1977). Selection procedures of regression analysis applied to automobile insurance. *Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker* **77:2**, 143-160.

Lemaire, J. (1979). Selection procedures of regression analysis applied to automobile insurance. Part II: Sample Inquiry and Underwriting Applications. *Mitteilungen der Vereinigung Schweizerischer Versicherungs-Mathematiker* **79:1**, 65-72.

Lemaire, J. (1985). *Automobile Insurance: Actuarial Models*. Kluwer-Nijhof Publishing. Boston, MA.

Lemaire, J. (1988). Construction of the new Belgian motor third party tariff structure. *ASTIN Bulletin* **18:1**, 99-112.

Lemaire, J. (1995). *Bonus-malus system in automobile insurance*. Kluwer-Nijhof Publishing. Boston, MA.

Lingoes, J. C. (1971). Some boundary conditions for a monotone analysis of symmetric matrices. *Psychometrika* **36**, 195-203.

Loimaranta, K., Jacobsson, J. y H. Lonka (1980). On the Use of Mixture Models in Clustering Multivariate Frequency Data. *Transactions of the 21 st International Congress of Actuaries* **2**, 147-161.

López, M. y J. López de la Manzanara (1996). *Estadística para actuarios*. Editorial MAPFRE, S.A. Madrid.

Mack, T. (1991). A simple parametric model for rating automobile insurance or estimating IBNR claims reserves. *ASTIN Bulletin* **21:1**, 93-109.

Magidson, J. (1992). Chi-squared analysis of a scalable dependent variable. In: *Proceedings of the 1992 Annual Meeting of the American Statistical Association, Educational Statistics Section.*

Magidson, J. (1993a). *SPSS for Windows: Chi-square Automatic Interaction Detection CHAID. Release 6.0*. SPSS Inc. Chicago.

Magidson, J. (1993b). The use of the new ordinal algorithm in CHAID to target profitable segments. *Journal of Database Marketing* **1:1**.

Mardia, K. V. (1978). Some properties of classical multidimensional scaling. *Communications in Statistics* **A7:13**, 1233-1241.

Mardia, K. V., Kent, J. T. y J. M. Bibby (1979). *Multivariate Analysis*. Academic Press. London.

Masure, L. (1978). Les méthodes de l'analyse discriminante appliquées aux problèmes de l'assurance automobile. *Bulletin de l'Association Royale des Actuaires Belgues*, 29-51.

McCullagh, P. y J. A. Nelder (1989). *Generalized Linear Models*. Second Edition. Chapman & Hall. London.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons. New York.

Messenger, R. y L. Mandell (1972). A model search technique for predictive nominal scale multivariate analysis. *Journal of the American Statistical Association* **67**, 768-772.

Mielke, P. W. y K. J. Berry (1985). Non-asymptotic inferences based on the chi-square statistic for r by c contingency tables. *Journal of Statistical Planning and Inference* **12**, 41-45.

Millenhall, S. J. (1999). A systematic relationship between minimum bias and generalized linear models. *Proceedings of the Casualty Actuarial Society* **86**, 393-487.

Miller, A. (2000). Another look at subset selection using linear least squares. *Communications in Statistics: Theory and Methods*, **29:(9&10)**, 2005-2017.

Morgan, J. y R. Messenger (1973). *THAID – A Sequential Analysis Program for the Analysis of Nominal Scale Dependent Variables*. Survey Research Centre, Institute for Social Research. University of Michigan.

Morgan, J. y J. A. Sonquist (1963). Problems in the analysis of survey data a proposal. *Journal of the American Statistical Association* **58**, 415-434.

Morillo, I. (2001). *Sistemas de Bonus-Malus: Comparaciones y alternativas*. Tesis Doctoral. Universidad de Barcelona.

Murphy, K. P., Brockman, M. J. y P. K. W. Lee (2000). Using generalized linear models to build dynamic pricing systems. *Casualty Actuarial Society Winter Forum.*

Nelder, J. A. y R. J. Verrall (1997). Credibility theory and generalised linear models. *ASTIN Bulletin* **27:1**, 71-82.

Neter, J. y W. Wasserman. (1973). *Fundamentos de estadística*. Compañía Editorial Continental, SA. México, pp. 207-213.

Nieto, U. y J. Vegas (1993). *Matemática Actuarial*. Editorial MAPFRE, S.A. Madrid.

Niggemeyer, B., Radtke, M. y A. Reich (1995). Applications of risk theory and multivariate analysis in insurance practice. *Applied Stochastic Models and Data Analysis* **11:3**, 231-244.

Oliva, F. (1995). *Aportacions a l'anàlisi discriminant basada en distancies. Estudi comparatiu de mètodes d'anàlisi discriminant amb dades mixtes.* Tesis Doctoral. Universidad de Barcelona.

Panjer, H. H. y G. E. Willmot (1992). *Insurance risk models*. The Society of Actuaries.

Pearson, W. H. (1966). Estimation of correlation measure from an uncertainty measure. *Psychometrika* **31:3**, 421-433.

Peduzzi, P. N., Hardy, R. J. y T. R. Holford (1980). A stepwise variable selection procedure for nonlinear regression models. *Biometrics* **36**, 511-516.

Peña, D. (1990). *Estadística: Modelos y Métodos*. Segunda Edición. Alianza Editorial. Madrid.

Pérez, J. L. (1986). Seguros de personas. En: *Tratado general de seguros: teoría y práctica de los seguros privados* **II**. Editado por el Consejo General de los Colegios de Agentes y Corredores de Seguros de España.

Pérez, J. L. (2001). *Conociendo el seguro*. Segunda Edición. Editorial UMESER, S.A. Barcelona.

Picech, L. y R. Pelessoni (1998). Some applications of unsupervised neural networks in Rate Making Procedure. *Proceedings of the 1998 General Insurance Convention and ASTIN Colloquium.* Glasgow, pp. 549-567.

Pitkänen, P. (1975). Tariff Theory. *ASTIN Bulletin* **8:2**, 204-228.

Pons, M. A. (1995). *Introducción a la teoría de la credibilidad.* Colección de publicaciones del Departamento de Matemática Económica, Financiera y Actuarial de la Universidad de Barcelona, nº 30.

Prokkola, E. y Y. Romppainen (1992a). Rate making for credit insurance in theory. *Transactions of the 24 th International Congress of Actuaries*, pp. 223-228.

Prokkola, E. y Y. Romppainen (1992b). Rate making for credit insurance in practice. *Transactions of the 24 th International Congress of Actuaries*, pp. 229-233.

Ramachandran, G. (1975). Factors affecting fire loss. Multiple regression models with extreme values. *ASTIN Bulletin* **8:2**, 229-241.

Ramírez, G. (1995). *Contribuciones al análisis de segmentación*. Tesis Doctoral. Universidad de Salamanca.

Rao, C. R. (1964). The use and interpretation of principal component analysis in applied research. Sankhä. *The Indian Journal of Statistics, Series A* **26**, 329-358.

Rao, C. R. y Y. Wu (1989). A strongly consistent procedure for model selection in a regression problem. *Biometrika* **76:2**, 369-74.

Reid, D. H. (1975). Representations of claims arising from a risk portfolio. *ASTIN Bulletin* **8:2**, 242-256.

Reid, D. H. (1984). The limits of resolution of a statistical rating system. *Transactions of the 22 nd International Congress of Actuaries* **4**, 281-287.

Renshaw. A. E. (1993). An application of exponential dispersion models in premium rating. *ASTIN Bulletin* **23**, 145-147.

Renshaw, A. E. (1994). Modelling the claims process in the presence of covariates. *ASTIN Bulletin* **24**, 265-286.

Ross, Sheldon M. (1989). *Introduction to Probability Models*. Fourth Edition. Academic Press, Inc. London.

Sánchez, M. (1992). Factores de riesgo y tarificación. *VIII Jornadas Comunitarias del Seguro del Automóvil.*

Sánchez, M. (1997). Segmentación de carteras. Aplicación al seguro de responsabilidad civil del automóvil. *Actuarios*, diciembre96/enero-febrero97, 62-65.

Schmitter, H. y E. Straub (1975). How to find the right subdivision into tariff classes. *ASTIN Bulletin* **8:2**, 257-261.

Schoenberg, I. J. (1935). Remarks to Maurice Fréchet 's article « Sur la définition axiomatique d'une classe d'espaces distanciés vectoriellement applicables sur l'espace de Hilbert ». *The Annals of Mathematics* **36**, 724-732.

Seber, G. A. F. (1984). *Multivariate observations*. John Wiley & Sons. New York.

Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Society* **88:422**, 486-494.

Sierra, M. A. (1986). *Análisis multivariante: teoría y aplicaciones en economía*. Ediser, DL. Barcelona.

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Royal Statistical Association* **B:13**, 238-241.

Smyth, G. K., y B. Jørgensen (To appear). Fitting Tweedie's compound Poisson model to insurance claims data: dispersion modelling. *ASTIN Bulletin.*

Stroinski, K. (1986). Modelling motor insurance claim frequencies. In: M. Goovaerts et al. eds., *Insurance and Risk Theory*. Reidel, Dordrecht-Boston, MA, pp. 453-458.

Stroinski, K. J. y I. D. Currie (1989). Selection of Variables for Motor Insurance Rating. *Insurance: Mathematics and Economics* **8**, 35-46.

Suits, D. (1984). Dummy variables: mechanics vs. interpretation. *Review of Economics and Statistics* **66**, 177-180.

Sundt, B. (1987). Two credibility regression approaches for the classification of passenger cars in a multiplicative tariff. *ASTIN Bulletin* **17:1**, 41-70.

Taylor, G. C. (1989). Use of spline functions for premium rating by geographic area. *ASTIN Bulletin* **19:1**, 91-122.

Toniolo, D. y H. Schmitter (1998). The handlign of continuous tariff variables: Tips and experience. *Proceedings of the 1998 General Insurance Convention and ASTIN Colloquium.* Glasgow, pp. 651-660.

Thrasher, R. P. (1991). CART: A Recent Advance in Tree-Structured Segmentation Methodology. *Journal of Direct Marketing* **5:1**, 36-47.

UNESPA (1995). *Estudio de la siniestralidad en la cartera de turismos. Modalidad Responsabilidad Civil. Segmentación año 1994*. Dirección de Estudios.

van der Laan, B. S. (1988). *Modelling Total Cost of Claims for Non-life Insurances*. Eburon Publisher. Delft.

van Eeghen, J., E. K. Greup y J. A. Nijssen (1983). *Rate Making*. Survey of Actuarial Science 2, Nationale-Nederlanden N. V. Rotterdam.

Vegas, A. (1992a). Fundamentos técnicos del sistema Bonus-Malus. *VIII Jornadas Comunitarias del Seguro del Automóvil.*

Vegas, A. (1992b). Fundamentos técnicos del sistema Bonus-Malus. *Previsión y Seguro* **22**, Enero 1993, 141-167.

Vegas, A. (1993). Aplicación del sistema Bonus-Malus a la tarificación de los jóvenes conductores en el seguro de automóvil. *Previsión y Seguro* **29**, Septiembre 1993, 55-78.

Ward, J. H. (1963). Hierarchical grouping to optimise an objective function. *Journal of the American Statistical Association*, 236-244.

Ward, J. H. y M. E. Hook (1963). Application of an Hierarchical Grouping Procedure to a problem of Grouping Profiles. *Educ. and Psychol. Measurement* **23:1**,69-82.

Worsley, K. J. (1977). A non parametric extension of a cluster analysis method by Scott and knott. *Biometrics* **33**, 532-535.

Zehnwirth, B. (1994). Ratemaking: from Bailey and Simon (1960) to Generalised Linear Regression Models. *Casualty Actuarial Society Winter Forum*, pp. 615-659.