CUADERNOS DE TRABAJO

FACULTAD DE ESTUDIOS ESTADÍSTICOS

Provisión de siniestros de Incapacidad Temporal utilizando Análisis de Supervivencia

Ana Crespo Palacios

Magdalena Ferrán Aranaz

Cuaderno de Trabajo número 03/2013

UCM

UNIVERSIDAD COMPLUTENSE MADRID

Los Cuadernos de Trabajo de la Facultad de Estudios Estadísticos constituyen una apuesta por la publicación de los trabajos en curso y de los informes técnicos desarrollados desde la Facultad para servir de apoyo tanto a la docencia como a la investigación.

Los Cuadernos de Trabajo se pueden descargar de la página de la Biblioteca de la Facultad www.ucm.es/BUCM/est/ y en la sección de investigación de la página del centro www.ucm.es/centros/webs/eest/

CONTACTO: Biblioteca de la Facultad de Estudios Estadísticos

Universidad Complutense de Madrid

Av. Puerta de Hierro, S/N

28040 Madrid

Tlf. 913944035

buc_est@buc.ucm.es

Los trabajos publicados en la serie Cuadernos de Trabajo de la Facultad de Estudios Estadísticos no están sujetos a ninguna evaluación previa. Las opiniones y análisis que aparecen publicados en los Cuadernos de Trabajo son responsabilidad exclusiva de sus autores.

ISSN: 1989-0567

Provisión de siniestros de Incapacidad Temporal utilizando Análisis de Supervivencia

Ana Crespo Palacios
Magdalena Ferrán Aranaz

1. Introducción

1.1. Antecedentes

En una compañía de seguros es fundamental conocer el capital que debe estar disponible para utilizar en caso, por ejemplo, de una gran catástrofe o en caso que cierto número de clientes soliciten alguna prestación al mismo tiempo. Es muy importante una estimación correcta de este concepto, tanto para no tener capital inmovilizado sin necesidad, como para evitar el incumplimiento de las obligaciones contractuales.

Dentro de la actividad de una aseguradora, las pólizas que garantizan el cobro de una indemnización diaria en caso de incapacidad temporal forman una de las líneas de negocio.

Esta indemnización se fija en el momento de la contratación y, junto con las tasas calculadas en la tarifa técnica, determinan lo que paga el cliente en cada cuota.

El objeto de este trabajo es poder calcular las provisiones que debe tener la compañía en caso que tuviera que hacer frente a estos siniestros de incapacidad laboral, determinando la reserva necesaria para los siniestros abiertos.

Llamamos *siniestro* en este caso, a la ocurrencia de un suceso contemplado en las coberturas de la póliza y que desencadena el comienzo de las obligaciones del asegurador, en este caso, el pago de una indemnización diaria hasta el cierre del siniestro, que coincidirá con el alta médica.

Para poder provisionar, es necesario saber el número de días que va a ser necesario pagar la indemnización.

Habitualmente se estima en función de los días de indemnización contratados por el cliente: hasta 30 días, hasta 90 días, hasta 365 días, hasta 547 días y hasta 730 días (máximo período indemnizable), pero resulta una medida muy grosera, ya que sabemos que depende de la enfermedad asociada a la baja médica, a la edad del cliente, al sexo (hay enfermedades asociadas a un único sexo, por ejemplo) y se piensa, aunque será el modelo el que nos lo confirme o no, que también intervienen el grupo de tarifa (asociado a las profesiones del cliente, que no tienen todas el mismo riesgo. Notar que estos siniestros están ligados a bajas por incapacidad temporal o accidentes de trabajo), la franquicia (supone que la empresa te va a pagar a partir de un día de baja diferente, según la franquicia contratada, o desde el principio), la provincia del cliente y la indemnización que ha contratado.

1.2. Metodología

A la vista de esta situación, se propone probar con la utilización del análisis de supervivencia para calcular la duración de los siniestros abiertos de una forma diferente, y a su vez, que permita ahondar más en qué variables y de qué forma inciden en este hecho.

1.2.1. Teoría del análisis de supervivencia[1]

Supongamos que nos enfrentamos a una variable del tipo T = “tiempo transcurrido hasta que se verifica un determinado suceso”, a la que de un modo general llamaremos *tiempo de*

[1] A. Martín Andrés, J.D. Luna del Castillo, 2004

supervivencia (por ser este tipo de problemas los que generaron el concepto). Tal variable aleatoria tiene la característica de ser continua y positiva (T>=0) y vendrá definida por:

$S(t) = P(T>t) = 1-F(t) =$ *Función de supervivencia*

Es claro el interés de conocer S(t) en un determinado problema. Con tal fin, habrá que tomar una muestra aleatoria de datos (t_1,t_2, …., t_n) a través de la cual se estimará S(t). El modo usual de determinar los valores t_i consiste en realizar un *estudio de seguimiento*: tomar n individuos y seguirlos en el tiempo hasta que ocurra el suceso que se haya fijado como objetivo. Ello exige matizar dos aspectos:

a. Es necesario fijar un margen temporal para el experimento. Es decir, habrá que indicar el punto de inicio y el punto final del estudio.
b. Debe fijarse claramente el punto de entrada y el de salida del estudio. La diferencia entre ambos proporcionará el valor de T en ese individuo.

Se observa pues que los individuos entran y salen al azar del estudio, de modo que el tiempo T entre ambos sucesos -*tiempo de supervivencia*- es una variable aleatoria.

Pero la medición de T presenta problemas. Cuando un individuo "muere" por la causa prevista y dentro del margen temporal del estudio, T puede ser medida si ninguna duda y diremos que su valor es un *dato no censurado*. Pero puede ocurrir que llegado el momento en que el estudio debe concluir y el individuo aún esté "vivo". Entonces diremos que el valor de T es un *dato censurado*. Un fenómeno similar puede ocurrir por otras causas: si un individuo entra en el estudio y se le pierde la pista o "fallece" por otras causas, el valor de T es también un *dato censurado*. Vemos pues que en este tipo de estudios coexisten datos ciertos (o no censurados) con los datos inciertos (o censurados).

1.2.2. Aplicación al estudio

Calcularemos una curva de supervivencia, con el objetivo de determinar lo que llamaremos "esperanza de vida". Esta esperanza de vida indica el número de días que permanecerán abiertos los siniestros.

Definimos el *tiempo de supervivencia* de un siniestro como el número de días transcurrido hasta que se cierra (la persona obtiene el "alta" médica), la *mortalidad* se asemeja al cierre y la *censura* es equivalente a los siniestros que, en el momento de recogida de datos, no han finalizado o han presentado una duración superior al período de indemnización contratado en la póliza.

Con la ayuda del procedimiento PHREG de SAS, obtendremos una función de supervivencia que se aplicará a cada siniestro abierto, valiéndonos del método de Kaplan-Meier y la regresión de Cox.

De acuerdo con los estudios y experiencia previos, se han seleccionado las variables que se consideran más significativas para realizar el cálculo:

- Franquicia (franq)
- Código de enfermedad (sicenf2)
- Grupo de tarifa (gtar)
- Edad (edad2)
- Sexo (sexo)
- Sucursal de la póliza (sdvdel)
- Indemnización diaria (s03ien),

y sus parámetros. Con ellas, vamos a determinar el modelo a aplicar.

Debido al gran volumen de datos manejados, hemos seleccionado una muestra de ellos para realizar este estudio, aunque para nuestra actividad empresarial utilizaremos todos los registros en caso que se adopte el modelo propuesto. La selección se ha hecho

considerando únicamente los cuatro códigos de enfermedad más frecuentes, representando 1/5 de la población total de siniestros en el momento de recoger los datos.

Todos los cálculos, tablas y gráficos que aparecen en este documento se han hecho utilizando el paquete estadístico SAS 9.3. Parte de la depuración de la base de datos, buscando los posibles errores e incoherencias de los mismos, se ha realizado con el paquete estadístico SPSS.

2. Datos

2.1. Obtención de datos

Extracción de la muestra: pólizas en vigor, pólizas canceladas (o en "baja") y sus correspondientes siniestros del producto denominado "Renta" desde 1 de enero 2005.

Fecha de recogida de datos: 25 de marzo de 2013.

El proceso, una vez definido, se lanzará el día 25 de cada mes (después de los procesos de generación y liquidación de recibos y de cancelaciones administrativas de pólizas). La fecha de recogida de datos se utiliza para comparar entre la duración teórica del siniestro (duración) y la real (diferencia entre fecha alta médica y dicha fecha), marcando como censurados aquellos en los que la duración real es superior. Si el siniestro está abierto sin fecha de alta médica, marcaremos estos siniestros también como censurados.

2.2. Definición de las variables del modelo

La variable en estudio o dependiente, es la variable DURSIN, mientras que las variables independientes son FRANQ (franquicia), S03DEL(provincia), SICENF2 (código de enfermedad), GTAR (grupo de tarifa), EDAD2 (edad), SEXO (sexo) y S03IEN (Indemnización por incapacidad temporal). Son las variables elegidas por su posible relación con la duración de siniestro y con las cuales vamos a establecer nuestro modelo inicial.

A continuación, pasamos a definir y describir las variables incluidas en el modelo:

Dursin: La variable dependiente: duración en días del siniestro.

Franq (franquicia): En caso de siniestro, es la parte del "daño" que corre a cargo del asegurado, como consecuencia de haberlo estipulado así en el contrato, lo que implica un menor coste del seguro. La parte que corresponde soportar al asegurado puede ser una cifra exacta o un porcentaje del daño producido por el siniestro. En el primer caso, si el importe del siniestro es inferior a la cantidad estipulada como franquicia, el coste de los daños correrá completamente a cargo del asegurado, si el siniestro supone una cantidad mayor, la entidad aseguradora indemnizará al asegurado por la parte del coste que exceda de la franquicia. Es una variable categórica. Sus categorías (expresadas en días) son: 0, 3, 7, 10, 15 y 30.

Sicenf2: Código de enfermedad. Es una variable categórica. Sus categorías corresponden con cuatro códigos distintos.

Censurado: Cuando los tiempos de supervivencia no se conocen con exactitud, los datos se consideran censurados. En este caso, se consideran censurados los siniestros sin fecha de alta médica o aquellos en que la duración del siniestro es mayor que el periodo contratado. Los valores de **dursin** se consideran censurados si el valor de **censurado** es 1.

Gtar: Se establecen 5 grupos de riesgo que dan lugar a las siguientes tarifas:

Tarifa S: Profesionales liberales altamente cualificados que no utilizan vehículo para desarrollar su profesión.

Tarifa A: Profesiones sin trabajo manual que para su desarrollo no hacen uso habitual de vehículos.

Tarifa B: Profesiones con trabajo manual sin uso de maquinaria o herramientas. Y tarifas S o A que hagan uso habitual de vehículo.

Tarifa C: Profesiones con trabajo manual y utilización de maquinaria o herramientas ligeras y profesionales de la conducción.

Tarifa D: Profesiones con trabajo manual y utilización de maquinaria pesada, andamios,...

En nuestra base de datos, este valor es el asignado según la explicación anterior y la profesión indicada por el cliente. Es una variable categórica, correspondiendo cada categoría con cada una de las tarifas: A, B, C, D y S.

Edad2: Es la edad del cliente cuando se abre el siniestro. No se recalcula conforme pasa el tiempo. Variable numérica (edad en años).

Sexo: Variable categórica que recoge el sexo del asegurado según la siguiente codificación (corresponde a los códigos utilizados en la grabación de la póliza asociada al siniestro): 1 Masculino, 2 Femenino.

Sdvdel: Provincia de contratación de la póliza. La consideramos como variable numérica porque no categoriza en el producto en estudio, al menos hasta el momento.

S03ien: Indemnización diaria por enfermedad o accidente. Elegida por el cliente en el momento de contratación de la póliza. El importe siempre estará dentro de los límites marcados por el departamento técnico en función de los riesgos detectados. Será la cantidad que se pagará al asegurado cada día mientras dure el siniestro. Es una variable numérica.

Como hemos explicado al principio, hemos optado, para presentar este trabajo, limitar los códigos de enfermedad a los cuatro más frecuentes en la base de datos global. Estos son:

Código 487000: Gripe y sus complicaciones.

Código 724000: Lumbago, lumbalgia, o ciática sin especificar.

Código 726000: Tendinitis periférica.

Código 845000: Esguince de tobillo sin especificar.

Hemos escogido estos códigos aleatoriamente, entre todas las posibilidades presentes en el fichero de códigos de enfermedad. Posteriormente a su elección, comprobamos que la muestra representaba 1/5 del total de la población en estudio, lo que nos pareció representativo para el estudio (10773 individuos)

3. Depuración de datos

3.1 Depuración de datos en el estudio

Los datos han sido extraídos de ficheros que forman parte de la base de datos de la compañía. Estos ficheros se graban a través de aplicaciones que forman parte del sistema informático de la misma, por medio de personal formado específicamente para ello.

A su vez, las aplicaciones mencionadas contienen todo un conjunto de chequeos y campos de entrada definidos por el departamento técnico del producto. Por tanto, gran parte de la depuración de datos se realiza en el momento de cumplimentarlos. No obstante, y para poder detectar posibles errores cometidos en alguna carga masiva puntual fuera de las aplicaciones nombradas, hemos realizado las siguientes tareas de depuración:

- Quitar asegurados repetidos

- Corregir los períodos mal grabados (tienen que ser consecutivos, si falta alguno en medio, consideramos que el válido es el superior, ya que es lo que tendrá el cliente garantizado por contrato)

- Comprobar que los datos sexo y edad del campo s03tar coincide con los de las variables sexo y edad2 del fichero. En caso que no sea así, prevalece la del campo s03tar, que es con el que se ha tarificado la póliza.

- Comprobaciones usuales de integridad de datos de todas las variables incluidas en el estudio mediante estadística descriptiva[2]:
  - ✓ Valores fuera de rango o no permitidos
  - ✓ Incoherencia de datos
  - ✓ Valores extremos/outliers

4. Modelo con todas las variables

4.1. Método de regresión de Cox

Comenzamos con un modelo incluyendo todas las variables que pensamos pueden tener algún impacto en la duración de los siniestros, tal y como hemos detallado en el punto anterior.

Para ello, utilizamos proc PHREG de SAS.

Este procedimiento implementa el método de regresión[2] propuesto por primera vez en 1972 por el estadístico británico Sir David Cox. El método de Cox no requiere escoger una distribución de probabilidad determinada para representar tiempos de supervivencia. Así mismo, utilizando este método, es relativamente sencillo incorporar covariables que varíen durante el periodo de observación.

Cox hizo dos innovaciones significativas: propuso un modelo, habitualmente denominado el *modelo de riesgos proporcionales* (proportional hazards model) y un nuevo método de estimación, que fue posteriormente denominado de máxima verosimilitud parcial (máximum partial likelihood). El término *regresión Cox* se refiere a la combinación del modelo y el método de estimación.

A continuación, describimos brevemente dicho modelo:

[2] Paul D. Allison, 2010

4.2. Modelo de riesgos proporcionales

Comenzaremos con el modelo básico que no incluye covariables dependientes del tiempo o riesgos no proporcionales. Habitualmente, se escribe de la siguiente manera:

$$h_i = \lambda_0 (t)\exp(\beta_1 x_{i1}+\ldots+\beta_k x_{ik}) \qquad (4.2.1)$$

Esta ecuación nos dice que el riesgo para cada i en el tiempo t es el producto de dos factores:

- Una función $\lambda_0(t)$ que no puede ser negativa
- Una función linear de un conjunto de k covariables fijas, que es exponencial.

La función $\lambda_0(t)$ se puede considerar como la función de riesgo de un individuo cuyas covariables tienen todas valores 0. Se suele denominar la función *baseline hazard*.

Aplicando el logaritmo a los dos lados de la ecuación, reescribimos el modelo como:

$$\log h_i(t) = \alpha(t)+\beta_1 x_{i1}+\ldots+\beta_k x_{ik} \qquad (4.2.2)$$

donde $\alpha(t) = \log \lambda_0(t)$. Si además especificamos $\alpha(t)=\alpha$, obtenemos el modelo exponencial. Si especificamos $\alpha(t)=\alpha t$, obtenemos el modelo de Gompertz. Por último, si especificamos $\alpha(t)=\alpha \log t$, tenemos el modelo Weibull. Como veremos, uno de los grandes atractivos de la regresión de Cox es que estas elecciones no son necesarias. La función $\alpha(t)$ puede tomar cualquier forma, incluso la de una función *step*.

A la ecuación 4.2.1 se le llama modelo de riesgos proporcionales porque el riesgo para cada individuo es una proporción fija del riesgo de cualquier otro individuo. Si calculamos el cociente de los riesgos de dos individuos i y j, y aplicamos la ecuación 4.2.1, tenemos:

$$\frac{h_i(t)}{h_j(t)} = \exp\{\beta_1(x_{i1}-x_j)+\ldots+\beta_k(x_{ik}-x_{jk})\}$$

Lo fundamental de esta ecuación es que $\lambda_0(t)$ se cancela en numerador y denominador. Como consecuencia de ello, el cociente de los riesgos es constante a lo largo del tiempo.

4.3. Verosimilitud parcial

Lo más destacable de la verosimilitud parcial es que puedes estimar los coeficientes β del modelo de riesgos proporcionales sin tener que especificar la función $\lambda_0(t)$.

La función de verosimilitud para el modelo de riesgos proporcionales de la ecuación 4.2.1 se puede factorizar en dos partes:

- Una parte depende de $\lambda_0(t)$ y de $\beta = [\beta_1, \beta_2\ldots\beta_k]$, el vector de los coeficientes.
- La otra parte depende únicamente de β.

Lo que la verosimilitud parcial hace es descartar la primera parte y tratar la segunda - la función de verosimilitud parcial - como si fuera una función de verosimilitud ordinaria. Obtienes las estimaciones encontrando valores de β que maximicen la verosimilitud parcial. Debido a que hay algo de la información sobre β en la parte que descartamos, la estimación resultante no es completamente eficiente, los errores estándar son mayores que si hubiéramos utilizado la función entera. En la mayoría de los casos, de todas formas, la pérdida de eficiencia es bastante pequeña (Efron, 1977). Lo que obtienes a cambio es robustez porque lo estimado tiene buenas propiedades: es consistente y asintóticamente normal. Dicho de otra forma, para muestras grandes las estimaciones son insesgadas y la distribución de la muestra es aproximadamente normal.

Otra propiedad interesante es que estas estimaciones dependen del rango de los tiempos del evento, y no de los valores numéricos. Cualquier transformación deja los coeficientes estimados inalterados.

A continuación presentaremos los resultados de la aplicación de la regresión de Cox al modelo con todas las variables. Se asimilará el tiempo de supervivencia de un siniestro con su duración en días, la mortalidad se asimila al cierre o fin del siniestro y la censura, como ya hemos explicado en apartados anteriores, con los siniestros sin fecha de alta médica o con una duración superior al tiempo máximo indemnizable.

4.4. Aplicación al estudio

Comentamos a continuación las tablas de salida de SAS 9.3, que son las siguientes:

Tabla 4.4.1

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
10773	10693	80	0.74

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Tabla 4.4.2

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	94858.146	89816.081
AIC	94858.146	89848.081
SBC	94858.146	89964.519

Tabla 4.4.3

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	5042.0645	16	<.0001
Score	7542.7313	16	<.0001
Wald	5559.3923	16	<.0001

Los grados de libertad corresponden al número de coeficientes del modelo. Con el método DISCRETE en el modelo, en la opción TIES, estamos asumiendo que el tiempo es una variable discreta.

Tabla 4.4.4

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
edad2	1	229.3189	<.0001
gtar	4	14.8108	0.0051
franq	5	476.4986	<.0001
sexo	1	10.2044	0.0014
sicenf2	3	4411.3142	<.0001
SDVDEL	1	10.7535	0.0010
S03IEN	1	0.3267	0.5676

Tabla 4.4.5

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
edad2		1	-0.01670	0.00110	229.3189	<.0001	0.983	
gtar	A	1	-0.15637	0.08055	3.7686	0.0522	0.855	gtar A
gtar	B	1	-0.24110	0.07782	9.5984	0.0019	0.786	gtar B
gtar	C	1	-0.18613	0.07789	5.7104	0.0169	0.830	gtar C
gtar	D	1	-0.16564	0.08073	4.2101	0.0402	0.847	gtar D
franq	0	1	0.90203	0.08533	111.7435	<.0001	2.465	franq 0
franq	3	1	0.64829	0.08695	55.5883	<.0001	1.912	franq 3
franq	7	1	0.47631	0.08612	30.5905	<.0001	1.610	franq 7
franq	10	1	0.56149	0.12389	20.5393	<.0001	1.753	franq 10
franq	15	1	0.24673	0.09346	6.9700	0.0083	1.280	franq 15
sexo	1	1	0.09517	0.02979	10.2044	0.0014	1.100	sexo 1
sicenf2	487000	1	1.98684	0.03853	2659.2531	<.0001	7.292	sicenf2 487000
sicenf2	724000	1	-0.24085	0.02816	73.1747	<.0001	0.786	sicenf2 724000
sicenf2	726000	1	-0.40827	0.03060	178.0102	<.0001	0.665	sicenf2 726000
SDVDEL		1	-0.00210	0.0006397	10.7535	0.0010	0.998	
S03IEN		1	-0.0003367	0.0005891	0.3267	0.5676	1.000	

En la tabla 4.4.3 observamos que los tres test rechazan la hipótesis nula de que los coeficientes $\beta = 0$, con p-valor menor que 0.0001. Es decir, hay al menos uno de los parámetros asociados a las variables explicativas que es distinto de cero.

Según el resultado obtenido en el test de la tabla 4.4.4, decimos que las variables edad2, franq y sicenf2 son significativas, al igual que se aprecia claramente que la variable s03ien no lo es, con un p-valor de 0.5676. Es decir, la duración del siniestro depende de la edad, franquicia y código de enfermedad, pero no de la indemnización diaria recibida por el cliente. El test para el resto de las variables introducidas en el modelo también es sinificativo.

En la variable gtar (grupo de tarifa), observamos diferentes comportamientos en sus distintas clases, sobre todo entre el grupo de tarifa B y el resto de grupos, ya que el Hazard

Ratio nos indica una probabilidad de supervivencia menor, es decir, la duración de siniestros parecen ser ligeramente inferior.

En la variable franq (franquicia) es destacable la diferencia entre el Hazard Ratio de la franquicia 0 (sin franquicia) y los demás grupos. Es un resultado lógico por definición. En los censurados con franquicia distinta de cero, puede que encontremos siniestros todavía en período de franquicia, que no interviene en la duración del siniestro.

Las variables edad2, sdvdel y s03ien (edad, provincia e indemnización diaria) apenas muestran diferencia entre los diferentes grupos.

El código de enfermedad 487000 es estadísticamente significativo y tiene un Hazard Ratio elevado. La duración de estos siniestros es inferior que para los otros códigos de enfermedad (menos probabilidad de supervivencia).

Aunque el modelo general con todas las variables ya nos da indicaciones claras de qué variables eliminaríamos del modelo (haciéndolo más sencillo sin perder poder explicativo), vamos a utilizar selección stepwise para encontrar el modelo definitivo (ver siguiente apdo).

5. Selección de variables

Para la selección definitiva de las variables presentes en el modelo utilizamos el método STEPWISE añadiéndolo como parámetro en PHREG.

Describimos brevemente las características de este método[3]:

Este método es muy similar al FORWARD. Partiremos de que ninguna variable está en el modelo. Se calcula el estadístico chi-cuadrado para cada una de ellas del Score test que verifica que la hipótesis nula de que el efecto correspondiente es cero. Si el mayor de estos estadísticos es significativo al nivel fijado en slentry es significativo, se añade la variable correspondiente al modelo. En cada paso se verifica a continuación si alguna de las variables presentes en el modelo debería salir, utilizando el nivel de significación definido en slstay (esta es la diferencia con el método FORWARD). El proceso se repite hasta que ningún estadístico es significativo al nivel del slentry o hasta que la variable que debería entrar en el modelo es la misma que debería salir en el subsiguiente proceso de eliminación.

A continuación detallaremos cada uno de los pasos del proceso y la conclusión final:

Los “score tests” individuales se usan para determinar cuál de todas las variables explicatorias va a ser la primera en entrar en el modelo. En nuestro caso particular, la salida obtenida de la ejecución del programa muestra los estadísticos chi-cuadrado y los correspondientes p-valores. La variable sicenf2 (código de enfermedad) tiene el valor de chi-cuadrado más alto y es significativa ($p < 0.0001$) con slentry=0.01. Por tanto, la variable sicenf2 se introduce en el modelo.

[3] SAS Help and Documentation

Tabla 5.1

Analysis of Effects Eligible for Entry			
Effect	**DF**	**Score Chi-Square**	**Pr > ChiSq**
edad2	1	447.0085	<.0001
gtar	4	59.6672	<.0001
franq	5	828.0424	<.0001
sexo	1	0.1488	0.6997
sicenf2	3	6507.5023	<.0001
SDVDEL	1	3.1694	0.0750
S03IEN	1	0.9225	0.3368

Residual Chi-Square Test		
Chi-Square	**DF**	**Pr > ChiSq**
7287.0353	16	<.0001

Step 1. Effect sicenf2 is entered. The model contains the following effects:

sicenf2

Tabla 5.2

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Without Covariates**	**With Covariates**
-2 LOG L	178125.90	174241.36
AIC	178125.90	174247.36
SBC	178125.90	174269.20

Tabla 5.3

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	3884.5308	3	<.0001
Score	6507.5023	3	<.0001
Wald	4883.4275	3	<.0001

En estas tablas vemos los resultados del modelo. Como el estadístico de Wald es significativo (p<0.0001) con slstay=0.01, la variable sicenf2 continúa en el modelo.

Tabla 5.4

Analysis of Effects Eligible for Entry			
Effect	**DF**	**Score Chi-Square**	**Pr > ChiSq**
edad2	1	320.4312	<.0001
gtar	4	56.9034	<.0001
franq	5	554.8948	<.0001
sexo	1	8.9876	0.0027
SDVDEL	1	11.6346	0.0006
S03IEN	1	0.6518	0.4195

Residual Chi-Square Test		
Chi-Square	**DF**	**Pr > ChiSq**
826.9010	13	<.0001

En el siguiente paso, seleccionamos otra variable que añadir al modelo. En la tabla superior, se muestran los estadísticos chi-cuadrado y los p-valores de los “score tests” individuales (ajustados por sicenf2) para las restantes variables del modelo general. En este caso se está comprobando si cada variable es significativa o no contando con la presencia de sicenf2 en el modelo. En este caso, la variable franq (franquicia) va a ser la seleccionada, porque tiene el valor de chi-cuadrado más alto (554.8948) y es significativa (p< 0.0001).

Step 2. Effect franq is entered. The model contains the following effects:

franq sicenf2

Tabla 5.5

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Tabla 5.6

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	178125.90	173668.23
AIC	178125.90	173684.23
SBC	178125.90	173742.45

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4457.6692	8	<.0001
Score	7015.6041	8	<.0001
Wald	5318.4875	8	<.0001

En las tablas anteriores vemos los resultados del modelo ajustado conteniendo las variables sicenf2 (código de enfermedad) y franq (franquicia). Basándonos en el estadístico de Wald, ninguna de las dos saldría del modelo.

Continuamos con un paso más:

Tabla 5.7

Analysis of Effects Eligible for Entry			
Effect	DF	Score Chi-Square	Pr > ChiSq
edad2	1	227.1884	<.0001
gtar	4	40.4079	<.0001
sexo	1	16.2686	<.0001
SDVDEL	1	9.4456	0.0021
S03IEN	1	1.3443	0.2463

Tabla 5.8

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
271.4032	8	<.0001

Como hemos hecho en pasos anteriores, basándonos en la variable con el estadístico de chi-cuadrado más alto y que sea significativa según los resultados de los "score tests" individuales, sería la variable edad2 (edad) la que entraría en el modelo.

Step 3. Effect edad2 is entered. The model contains the following effects:

edad2 franq sicenf2

Tabla 5.9

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Without Covariates**	**With Covariates**
-2 LOG L	178125.90	173438.88
AIC	178125.90	173456.88
SBC	178125.90	173522.38

Tabla 5.10

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	4687.0123	9	<.0001
Score	7242.2931	9	<.0001
Wald	5540.6381	9	<.0001

Tal y como hemos hecho en pasos anteriores, nos basamos en los resultados del test de Wald para afirmar que las tres variables continuarían en el modelo.

Avanzamos un paso más:

Tabla 5.11

Analysis of Effects Eligible for Entry			
Effect	**DF**	**Score Chi-Square**	**Pr > ChiSq**
gtar	4	23.5831	<.0001
sexo	1	18.6048	<.0001
SDVDEL	1	11.0719	0.0009
S03IEN	1	1.0537	0.3047

Residual Chi-Square Test		
Chi-Square	**DF**	**Pr > ChiSq**
43.7221	7	<.0001

Atendiendo a los mismos criterios que venimos describiendo, tanto la variable gtar (grupo de tarifa) como la variable sexo (sexo) son candidatas a entrar en el modelo. La variable sexo tiene 1 grado de libertad, es la elegida.

A continuación veremos los efectos en el modelo de la variable que se ha introducido en este paso.

Reseñar que, de momento, no hay ninguna variable que haya entrado en el modelo y sea candidata a salir en el siguiente *step*, circunstancia que nos obligaría a detener el proceso.

A su vez, tampoco hemos recibido ningún aviso indicando que ninguna variable cumple ya el criterio de entrada. Hasta que nos enfrentemos a alguna de estas situaciones, el proceso de selección de variables continúa.

Step 4. Effect sexo is entered. The model contains the following effects:

edad2 franq sexo sicenf2

Tabla 5.12

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	**Without Covariates**	**With Covariates**
-2 LOG L	178125.90	173419.88
AIC	178125.90	173439.88
SBC	178125.90	173512.66

Tabla 5.13

Testing Global Null Hypothesis: BETA=0			
Test	**Chi-Square**	**DF**	**Pr > ChiSq**
Likelihood Ratio	4706.0114	10	<.0001
Score	7262.3190	10	<.0001
Wald	5557.5247	10	<.0001

A la vista de los resultados del modelo ajustado, las cuatro variables permanecerían en el modelo con slstay=0.01. Llegamos al quinto paso de nuestro proceso de selección de variables.

Tabla 5.14

Analysis of Effects Eligible for Entry			
Effect	DF	Score Chi-Square	Pr > ChiSq
gtar	4	14.9128	0.0049
SDVDEL	1	10.9813	0.0009
S03IEN	1	0.3521	0.5530

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
25.3451	6	0.0003

Observamos que las dos posibles candidatas son las variables gtar (grupo de tarifa) y sdvdel (provincia). Aunque gtar presenta un "score chi-square" más alto, el p-valor de sdvdel es inferior e introduciríamos menos grados de libertad al modelo. Esa variable va a ser la elegida, como vemos a continuación:

Step 5. Effect SDVDEL is entered. The model contains the following effects:

edad2 franq sexo sicenf2 SDVDEL

Tabla 5.15

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	178125.90	173408.95
AIC	178125.90	173430.95
SBC	178125.90	173511.00

Tabla 5.16

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4716.9487	11	<.0001
Score	7273.6824	11	<.0001
Wald	5564.8746	11	<.0001

A tenor de los resultados de los tests con el modelo ajustado, comprobamos que las cinco variables permanecerían en el modelo.

SAS todavía nos indica que una variable más puede formar parte del modelo:

Tabla 5.17

Analysis of Effects Eligible for Entry			
Effect	DF	Score Chi-Square	Pr > ChiSq
gtar	4	14.0090	0.0073
S03IEN	1	0.1414	0.7069

Residual Chi-Square Test		
Chi-Square	DF	Pr > ChiSq
14.2631	5	0.0140

La variable gtar (grupo de tarifa) es la variable que entra en el modelo en este paso, a tenor del resultado del "score Chi-square" y de su p-valor. Estamos todavía ante una variable significativa. Notar que en ningún momento alguna variable ha salido del modelo.

Step 6. Effect gtar is entered. The model contains the following effects:
edad2 gtar franq sexo sicenf2 SDVDEL

Tabla 5.18

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	178125.90	173395.23
AIC	178125.90	173425.23
SBC	178125.90	173534.40

Tabla 5.19

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4730.6607	15	<.0001
Score	7286.4388	15	<.0001
Wald	5575.8186	15	<.0001

Note: No (additional) effects met the 0.01 level for entry into the model.

Los resultados de los tests con el modelo ajustado nos indica que, nuevamente, todas las variables permanecen en el modelo. Asi mismo, SAS nos indica que ninguna otra alcanza el nivel (slentry) necesario para entrar en el modelo, con lo que el proceso concluye aquí.

Tabla 5.20

Type 3 Tests			
Effect	**DF**	**Wald Chi-Square**	**Pr > ChiSq**
edad2	1	217.0800	<.0001
gtar	4	13.9790	0.0074
franq	5	460.6269	<.0001
sexo	1	9.2507	0.0024
sicenf2	3	4369.8099	<.0001
SDVDEL	1	10.2024	0.0014

Tabla 5.21

Analysis of Maximum Likelihood Estimates								
Parameter		**DF**	**Parameter Estimate**	**Standard Error**	**Chi-Square**	**Pr > ChiSq**	**Hazard Ratio**	**Label**
edad2		1	-0.01578	0.00107	217.0800	<.0001	0.984	
gtar	**A**	1	-0.13946	0.07543	3.4184	0.0645	0.870	gtar A
gtar	**B**	1	-0.21666	0.07165	9.1446	0.0025	0.805	gtar B
gtar	**C**	1	-0.16401	0.07098	5.3388	0.0209	0.849	gtar C
gtar	**D**	1	-0.14448	0.07435	3.7762	0.0520	0.865	gtar D
franq	**0**	1	0.87860	0.08466	107.6945	<.0001	2.408	franq 0
franq	**3**	1	0.63930	0.08620	55.0020	<.0001	1.895	franq 3
franq	**7**	1	0.47214	0.08544	30.5342	<.0001	1.603	franq 7
franq	**10**	1	0.55300	0.12206	20.5254	<.0001	1.738	franq 10
franq	**15**	1	0.24581	0.09264	7.0412	0.0080	1.279	franq 15
sexo	**1**	1	0.08686	0.02856	9.2507	0.0024	1.091	sexo 1
sicenf2	**487000**	1	1.78925	0.03565	2518.9682	<.0001	5.985	sicenf2 487000
sicenf2	**724000**	1	-0.23311	0.02761	71.2954	<.0001	0.792	sicenf2 724000
sicenf2	**726000**	1	-0.39649	0.03006	174.0262	<.0001	0.673	sicenf2 726000
SDVDEL		1	-0.00198	0.0006199	10.2024	0.0014	0.998	

Tabla 5.22

Summary of Stepwise Selection							
Step	Effect Entered	Effect Removed	DF	Number In	Score Chi-Square	Wald Chi-Square	Pr > ChiSq
1	sicenf2		3	1	6507.5023		<.0001
2	franq		5	2	554.8948		<.0001
3	edad2		1	3	227.1884		<.0001
4	sexo		1	4	18.6048		<.0001
5	SDVDEL		1	5	10.9813		0.0009
6	gtar		4	6	14.0090		0.0073

Para finalizar, SAS nos muestra unas tablas resumen de los pasos del proceso de selección de variables y de cómo quedaría el modelo.

Recapitulando lo visto en este apartado:

- El proceso stepwise de selección de variables nos proporciona como resultado final un modelo con seis variables explicativas de las siete posibles, con lo que, sin perder explicativo a tenor de lo obtenido en los tests, tendríamos un modelo más sencillo.

- A pesar de que el resultado óptimo es este, presentado el modelo al departamento técnico correspondiente, se decide eliminar tambien del modelo la variable sdvdel. La razón es que obligaría a un replanteamiento del modelo comercial del ramo, aspecto no abordable en este momento. Aunque estadísticamente es el modelo obtenido, vamos a seguir abordando el estudio según estas opiniones, buscando la posible aplicación futura del modelo.

- Con estas consideraciones, el modelo reducido por el que se va a optar es aquel en que la duración del siniestro viene explicada por el código de enfermedad (sicenf2), la franquicia (franq), la edad (edad2), el sexo (sexo) y el grupo de tarifa (gtar).

6. Modelo definitivo

Una vez terminado el proceso de selección de variables, vamos a analizar el modelo para comprobar su idoneidad.

La justificación teórica es la mismo que el apartado 4. Estamos utilizando las mismas técnicas estadísticas, pero aplicadas a un modelo con menos variables explicativas, con los mismos fundamentos.

Los resultados obtenidos al ejecutar el programa en SAS son los siguientes:

Tabla 6.1

Summary of the Number of Event and Censored Values			
Total	Event	Censored	Percent Censored
10773	10693	80	0.74

Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Tabla 6.2

Model Fit Statistics		
Criterion	Without Covariates	With Covariates
-2 LOG L	94858.146	89827.003
AIC	94858.146	89855.003
SBC	94858.146	89956.886

Tabla 6.3

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	5031.1428	14	<.0001
Score	7531.1455	14	<.0001
Wald	5550.5405	14	<.0001

Si comparamos con el análisis hecho al modelo con todas las variables, vemos que estamos en los mismos valores para los estadísticos que verifican el ajuste del modelo. Al tener dos variables explicativas menos, estamos trabajando con un modelo más sencillo, no obstante, habiendo eliminado variables que no aportaban nada al modelo.

Rechazamos la hipótesis nula de que todos los coeficientes β son cero (p-valor < 0.0001 en los tests).

Tabla 6.4

Type 3 Tests			
Effect	DF	Wald Chi-Square	Pr > ChiSq
edad2	1	228.1660	<.0001
gtar	4	15.5326	0.0037
franq	5	485.1634	<.0001
sexo	1	10.5923	0.0011
sicenf2	3	4403.9574	<.0001

Tabla 6.5

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
edad2		1	-0.01665	0.00110	228.1660	<.0001	0.983	
gtar	B	1	-0.08516	0.03726	5.2233	0.0223	0.918	gtar B
gtar	C	1	-0.03395	0.03694	0.8446	0.3581	0.967	gtar C
gtar	D	1	-0.01259	0.04257	0.0875	0.7674	0.987	gtar D
gtar	S	1	0.15474	0.07746	3.9909	0.0457	1.167	gtar S
franq	3	1	-0.26044	0.02681	94.3785	<.0001	0.771	franq 3
franq	7	1	-0.43065	0.02548	285.5881	<.0001	0.650	franq 7
franq	10	1	-0.34408	0.09276	13.7583	0.0002	0.709	franq 10
franq	15	1	-0.65507	0.04512	210.7765	<.0001	0.519	franq 15
franq	30	1	-0.90310	0.08533	112.0233	<.0001	0.405	franq 30
sexo	2	1	-0.09580	0.02944	10.5923	0.0011	0.909	sexo 2
sicenf2	724000	1	-2.22599	0.03599	3825.1253	<.0001	0.108	sicenf2 724000
sicenf2	726000	1	-2.38829	0.03807	3934.6740	<.0001	0.092	sicenf2 726000
sicenf2	845000	1	-1.98095	0.03848	2650.5242	<.0001	0.138	sicenf2 845000

En los tests individuales, apreciamos que todas las variables que permanecen en el modelo son significativas.

Analizando la tabla anterior, podemos afirmar que únicamente el grupo de tarifa S incide positivamente en la duración de los siniestros respecto a la referencia. En el resto de los casos, en mayor o menor medida, cada uno de los posibles grupos de cada variable disminuye la probabilidad de que el siniestro haya finalizado en cualquier momento del periodo de seguimiento.

En el caso de la edad y el sexo, prácticamente nos encontramos con un Hazard Ratio de 1, estaríamos ante comportamientos muy similares.

Analizando los valores de HR para la variable SICENF2, parece claro que la duración de los siniestros para el código de enfermedad de referencia es muy inferior a la duración de los siniestros correspondientes a los otros códigos de enfermedad.

A su vez, el análisis de los HR obtenidos para las distintas clases de la variable franquicia nos muestra una probabilidad inferior conforme la franquicia es mayor, excepto para franquicia 7 días y franquicia 10 días. Es un hecho a tener en cuenta cuando analicemos las curvas log-rank de cada una de las variables.

En los tests individuales, apreciamos que todas las variables que permanecen en el modelo son significativas.

Analizando la tabla 6.5, podemos afirmar respecto a la variable *gtar,* que únicamente el grupo de tarifa S incide positivamente en la duración de los siniestros respecto a la referencia. En el resto de los casos, en mayor o menor medida, cada uno de los grupos de la variable disminuye la probabilidad de que el siniestro haya finalizado en cualquier

momento del periodo de seguimiento. Notar, no obstante, que gtar = C y gtar = D no son estadísticamente significativas.

Este análisis es el mismo en el resto de variables, con parámetros negativos (y Hazard Ratio inferior a 1.

En el caso de la variable *sexo*, con un Hazard Ratio de 0.909. Estamos diciendo que ser mujer lleva asociado duraciones de siniestros más largas que ser hombre.

Edad2 (edad) es una variable en la que no se han diferenciado grupos. La interpretación de su Hazard Ratio es global: la probabilidad de que el siniestro haya finalizado en el siguiente seguimiento es menor conforme la edad es más avanzada, sus siniestros tienen una duración mayor.

Analizando los valores de Hazard Ratio para la variable SICENF2, parece claro que la duración de los siniestros para el código de enfermedad de referencia es muy inferior a la duración de los siniestros correspondientes a los otros códigos de enfermedad.

A su vez, el análisis de los Hazard Ratio obtenidos para los distintos grupos de la variable franquicia nos muestra una probabilidad inferior conforme la franquicia es mayor, excepto para franquicia 7 días y franquicia 10 días. Es un hecho a tener en cuenta cuando analicemos las curvas log-rank de cada una de las variables.

Realizada la prueba de riesgos proporcionales, hemos detectado algún problema con la franquicia 0 y 3 (resultados que coinciden con lo que veremos en la curvas log-rank y en el análisis de residuos). El resto de variables no nos dan ninguna evidencia que nos haga rechazar la hipótesis de proporcionalidad. Hemos analizado el modelo incluyendo como *strata* la variable franq y los coeficientes obtenidos para el resto de variables son prácticamente los mismos que para el modelo que estamos describiendo. A la vista de todas estas evidencias, aceptamos la hipótesis de proporcionalidad, sin la cual no estaríamos escogiendo el modelo adecuado.

7. Curva general Kaplan-Meier

El método más conocido para estimar la función de supervivencia S(t) es el *método de Kaplan-Meier*[4]. Este método ya era conocido con anterioridad a 1958, cuando Kaplan y Meier mostraron que era, de hecho, el estimador no paramétrico de máxima verosimilitud. Si no hay datos censurados, el estimador KM es simple e intuitivo. Sería la proporción de observaciones en la muestra cuyo tiempo de fallo es mayor que t (asimilaremos el término “fallo” con “muerte” o “fin de siniestro” en nuestro caso.

Consideramos una muestra de n individuos de los que se conoce su tiempo de fallo o el instante de censura. Supondremos que se han observado s, s<=n, tiempos de fallo que denotamos, una vez ordenados, $t_1,..,t_s$. Es posible que en la muestra se produzcan empates, es decir, observaciones cuyo tiempo de fallo es el mismo y por eso de define d_i, (d_i>= 1), como el número de fallos que se producen en el instante t_i. Las restantes observaciones, $n - \sum d_i$, son los tiempos de seguimiento de los individuos cuyo fallo no ha sido observado.

El estimador Kaplan-Meier de S(t) se define como:

$$\hat{S}(t) = \prod_{t_i \leq t} \frac{n_i - d_i}{n_i}$$

[4]A. Martín Andrés, J.D. Luna del Castillo, 2004
Paul D. Allison, 2010
Jesús Abaurrea, Ana Carmen Cebrián, 2005

donde n_i es el número de individuos en riesgo en el instante t_i; es decir, el número de individuos vivos y no censurados, justo antes de t_i. Si existe alguna observación censurada cuyo valor coincide con un tiempo de fallo, se hace la hipótesis de que la observación censurada ocurre inmediatamente después del tiempo de fallo y, en consecuencia, los individuos censurados en ese instante se contabilizan como individuos en riesgo.

El estimador KM es una función constante entre los tiempos de fallo consecutivos, que vale 1 antes del menor tiempo de fallo, t_1, y cuyo valor decrece según unfactor variable en cada instante de fallo. El estimador no cambia en los tiempos de observación correspondientes a los individuos censurados, aunque esas observaciones influyen en el estimador a través de los valores n_i. Cuando el mayor de los tiempos observados en la muestra, t_M, es un tiempo de fallo, la estimación KM toma el valor 0 a partir de ese instante; si t_M corresponde a una observaión censurada, es habitual considerar que $\hat{S}(t)$ no está definido para $t > t_M$.

A través del procedimiento LIFETEST de SAS, que nos da por defecto el estimador KM, hemos obtenido los siguientes resultados:

Tabla 7.1

Quartile Estimates				
		95% Confidence Interval		
Percent	Point Estimate	Transform	[Lower	Upper)
75	42.000	LOGLOG	41.000	45.000
50	21.000	LOGLOG	21.000	22.000
25	11.000	LOGLOG	.	.

Tabla 7.2

Mean	Standard Error
40.586	0.716

El estimador KM cuando *dursin* =1 es 0.9974. Esto significa que la probabilidad estimada de que la duración de un siniestro sea de un día o más es 0.9974. Veríamos también que cuando *durs*in = 547 y el dato está censurado, el estimador KM no está definido. Para ese mismo valor de *dursin*, y dato no censurado, la probabilidad estimada de que la duración del siniestro sea de 547 días o más es 0.00214.

De forma complementaria al estimador KM obtenemos la probabilidad estimada de que un siniestro haya finalizado antes de 1 día (1- estimador KM). Como hemos visto en el desarrollo teórico, el estimador KM para *dursin* = 0 días es 1 y el que corresponde a *dursin* más alto (567 días) no censurado es 0.

Si seguimos estudiando esa línea, obtenemos también un estimador del error estándar, que se obtiene a través de la fórmula de Greenwood (Collett, 2003), así como el número de casos acumulado que ha tenido fallos hasta ese punto y el número de casos que no han tenido fallos hasta ese punto y no han sido censurados hasta entonces.

En la tabla 7.1 encontramos los percentiles 25, 50 y 75.

El percentil 25, cuyo valor es 11, indica que 11 días es el menor valor de *dursin* tal que la probabilidad de que un siniestro se haya cerrado antes es mayor que 0.25.

El percentil 50 nos da el valor de la mediana (21 días) de cierre de un siniestro, con un intervalo de confianza al 95% de (21,22).

El percentil 75, cuyo valor es 42, indica que 42 días es el menor valor de *dursin* tal que la probabilidad de que un siniestro se haya cerrado antes es mayor que 0.75.

Adicionalmente, la tabla 7.2 nos proporciona una estimación de la media de días hasta que cierra un siniestro (casi 41 días). Este valor se calcula directamente con la función estimada de supervivencia, y en ocasiones puede estar bastante sesgada por los datos censurados. Normalmente, la mediana suele ser mejor medida de la tendencia central en estos casos.

Gráfico 7.1

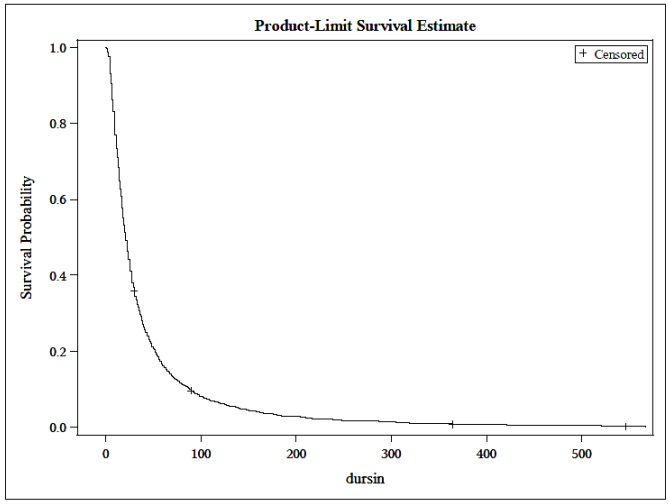


Gráfico 7.2

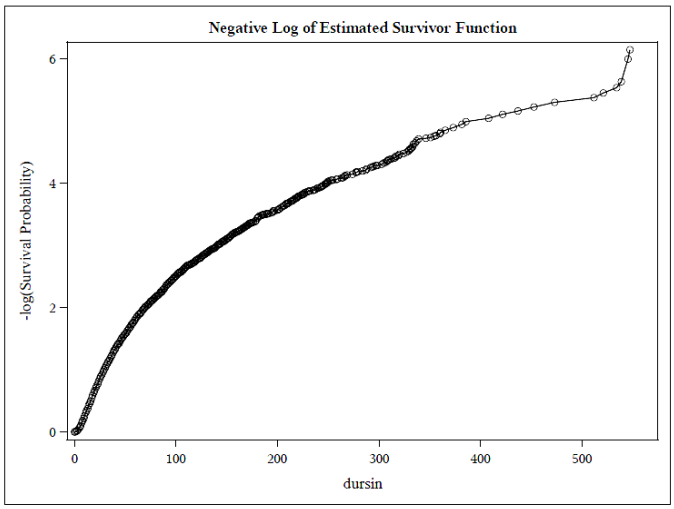


Gráfico 7.3

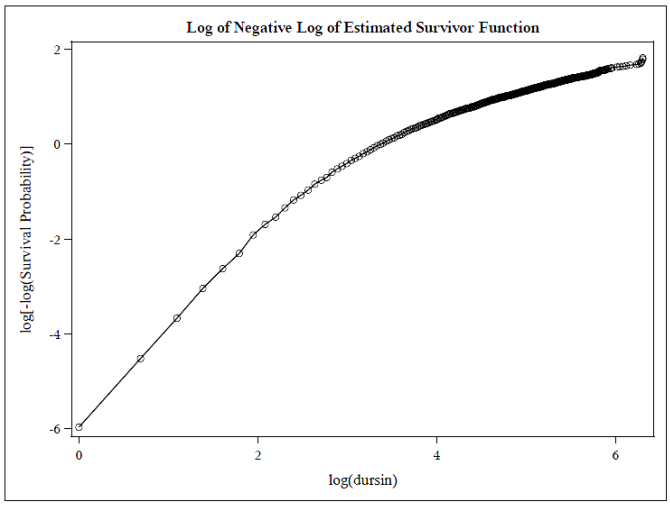
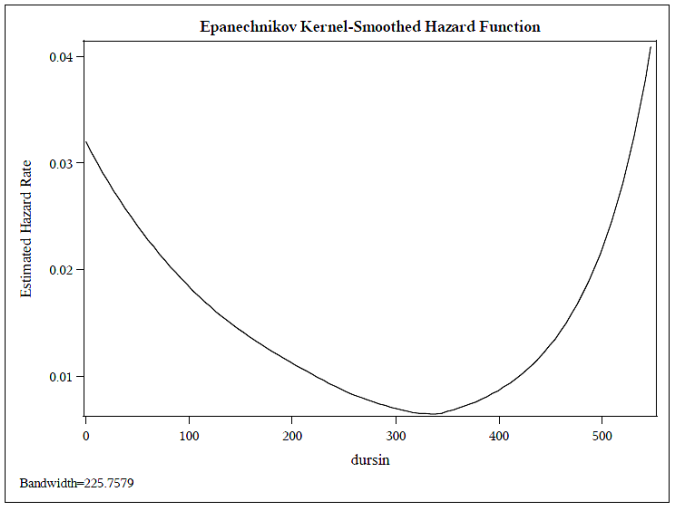


Gráfico 7.4



Analizando los distintos gráficos obtenidos, extraemos las siguientes conclusiones:

- En el gráfico 7.1 observamos que S(t) tiene una caída bastante rápida al principio, pero luego se mantiene constante. Esto quiere decir la probabilidad que en un siniestro permanezca abierto disminuye de forma rápida al principio, con el transcurso de los días, estabilizándose más o menos a partir de los 90 días, coincidiendo con el inicio del tercer período (91 a 365 días, pág. 5).

- Comparando los gráficos 7.2 y 7.3 vemos que se aproxima más a la diagonal la curva log-log, por tanto podemos decir que se aproxima más a una Weibull. Aunque observamos cierta tendencia, no es suficiente como para rechazar dicha distribución.

- Si el Hazard Ratio fuera constante, el gráfico 7.2 mostraría una línea recta. Lo que observamos es una tendencia decreciente desde aproximadamente día 30. Por el contrario, a partir del día 500 aproximadamente, la tendencia es creciente. De todas

formas, son impresiones visuales que deben ser refutadas por medio de su correspondiente test

- En el gráfico 7.4, H(t) tiene forma de "bañera". El primer período (hasta 30 días) con tasa de fallo alto, va suavizándose entre 2º (hasta 90 días) y 3º período (hasta 365 días) para empezar a crecer de nuevo.

8. Test log-rank[5]

8.1. Justificación teórica

A continuación, vamos a determinar si hay diferencias en las distintas curvas de supervivencia de los diferentes grupos que tenemos en cada variable del modelo.

Para ello nos basaremos en los resultados de los tests de las curvas log-rank, que nos permiten hacer una comparación de la supervivencia de dos o más poblaciones que se diferencian en alguna característica, como es nuestro caso (por ejemplo, distinto código de enfermedad en el siniestro).

En este tipo de ensayos es habitual utilizar métodos no paramétricos, como éste, ya que no suele disponer de la información suficiente para formular hipótesis sobre la forma de la función de supervivencia.

Supongamos que se quiere comparar la supervivencia en dos poblaciones de individuos, A y B, y que se dispone de una muestra de cada población de tamaño n_A y n_B respectivamente; sea $n = n_A+n_B$ el tamaño de la muestra combinada y $t_1<\ldots t_j<\ldots t_J$, los J tiempos de fallo distintos observados en la muestra conjunta ordenados en forma creciente. Denotaremos d_{Aj} y d_{Bj} el número de fallos ocurridos en el instante t_j en cada muestra y por d_j el número total de fallos observados en ese instante, es decir, $d_j = d_{Aj} + d_{Bj}$. Análogamente, denotaremos por n_{Aj} y n_{Bj} el número de individuos en riesgo en cada muestra justo antes del instante t_j, y por n_j la suma $n_{Aj} + n_{Bj}$. Toda esta información se puede disponer en un conjunto de J tablas de contingencia 2x2, una para cada tiempo de fallo.

La hipótesis a contrastar es que las funciones de supervivencia en ambas poblaciones coinciden en el intervalo de tiempo observado:

H_0:$S_A(t) = S_B(t)$ para todo t<=T

o, equivalentemente,

H_0: $h_A(t) = h_B(t)$ para todo t<=T

donde, en general, T se toma igual al mayor tiempo de supervivencia observado. Para contrastar esta hipótesis consideramos la discrepancia que se observa entre los individuos que fallan en cada uno de los grupos en cada instante de fallo, y el correspondiente número esperado de fallos bajo H_0 (fallo=fin siniestro).

El siguiente paso es construir un estadístico que combine la información de las J tablas 2x2 y proporcione una medida global de la desviación existente entre los valores observados y esperados bajo H_0. Sumando las desviaciones observadas en los distintos instantes de fallo se obtiene:

$U_L = \sum_{j=1}^{J}(d_{Aj} - e_{Aj}) = \sum_{j=1}^{J} d_{Aj} - \sum_{j=1}^{J} e_{Aj}$, donde $e_{Aj} = n_{Aj}\frac{d_j}{n_j}$

[5] Jesús Abaurrea, Ana Carmen Cebrián, 2005

Estadístico que tiene media cero y, bajo la hipótesi de que las J tablas son independientes, varianza la suma de las varianzas de los sumandos. Aplicando el teorema central del límite se puede justificar, si el número de tiempos de fallo no es demasiado pequeño, que U_L tiene una distribución aproximadamente normal (0,1).

El cuadrado de una variable Normal estándar tiene una distribución chi-cuadrado con 1 grado de libertad, por lo que una forma equivalente de evaluar las diferencias es construir un test basado en $Q_L = U_L^2/Var(U_L)$. Podemos generalizar esta teoría para tres o más poblaciones, con G >=3. Para este caso la hipótesis nula seguiría siendo:

H_0: $h_1(t) = h_2(t) = \ldots = h_G(t)$, para todo t <= T,

frente a la alternativa de que las tasas de fallo de al menos dos de los grupos difieran para algún instante t.

El test basado en U_L es el test log-rank, planteado por Mantel y Haenszel en 1959. De el nos vamos a ayudar para comparar la curvas de supervivencia de los grupos de cada variable categórica.

8.2. Curva log-rank variable franquicia (franq)

Mostramos las tablas de resultados, obtenidas a través de SAS:

Tabla 8.2.1

Summary of the Number of Censored and Uncensored Values					
Stratum	franq	Total	Failed	Censored	Percent Censored
1	0	5248	5221	27	0.51
2	3	2105	2088	17	0.81
3	7	2555	2523	32	1.25
4	10	125	125	0	0.00
5	15	594	590	4	0.67
6	30	146	146	0	0.00
Total		10773	10693	80	0.74

Tabla 8.2.2

Rank Statistics	
franq	Log-Rank
0	1283.4
3	-8.5
7	-690.5
10	-32.6
15	-382.3
30	-169.5

Tabla 8.2.3

Covariance Matrix for the Log-Rank Statistics						
franq	0	3	7	10	15	30
0	2380.96	-748.57	-1132.59	-55.65	-337.09	-107.06
3	-748.57	1630.27	-608.48	-29.75	-183.93	-59.53
7	-1132.59	-608.48	2168.05	-46.18	-286.80	-93.99
10	-55.65	-29.75	-46.18	150.11	-13.95	-4.56
15	-337.09	-183.93	-286.80	-13.95	851.95	-30.18
30	-107.06	-59.53	-93.99	-4.56	-30.18	295.34

Tabla 8.2.4

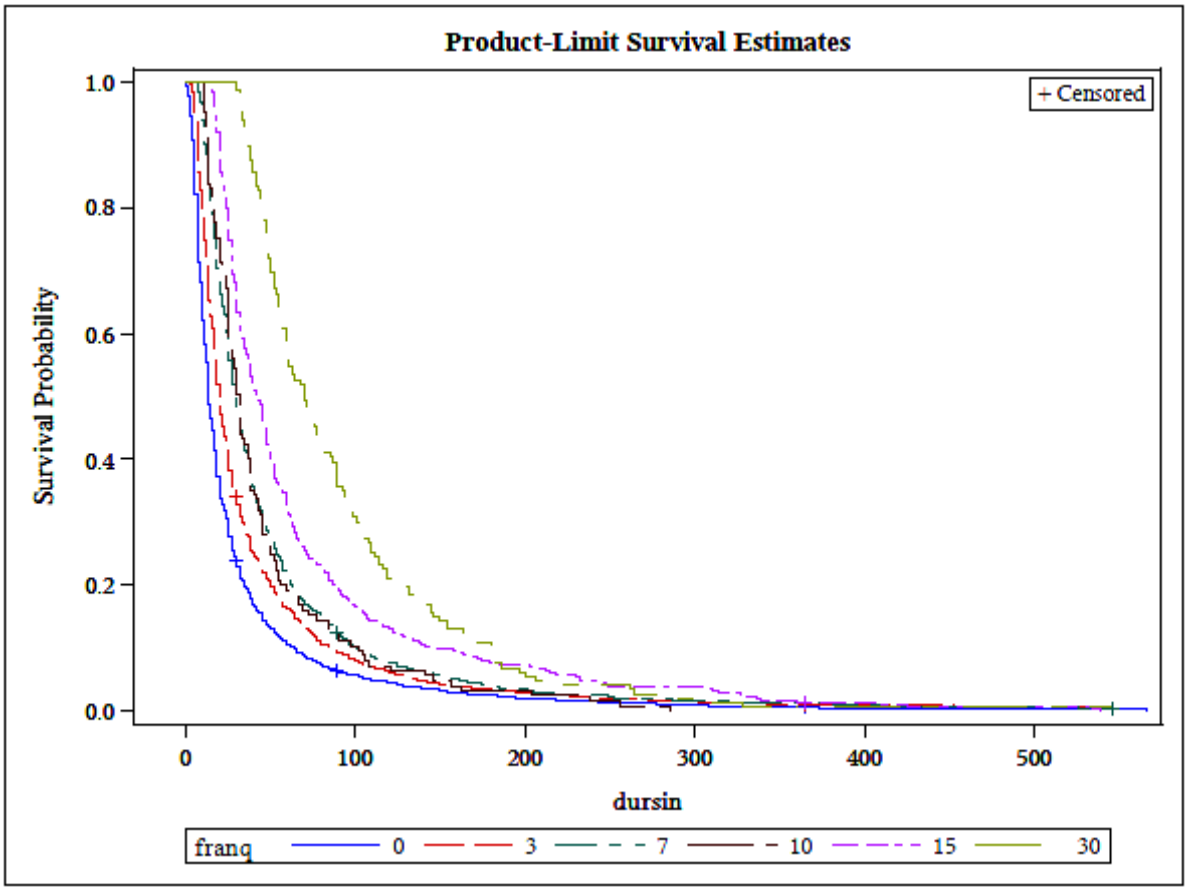
Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	855.4005	5	<.0001

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
franq	franq	Chi-Square	Raw	Sidak
0	3	303.0	<.0001	<.0001
0	7	571.7	<.0001	<.0001
0	10	655.4	<.0001	<.0001
0	15	710.1	<.0001	<.0001
0	30	730.3	<.0001	<.0001

Tabla 8.2.5

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
franq	franq	Chi-Square	Raw	Sidak
3	7	92.7387	<.0001	<.0001
3	10	0.3169	0.5735	1.0000
3	15	49.0394	<.0001	<.0001
3	30	12.6871	0.0004	0.0055
7	10	179.5	<.0001	<.0001
7	15	26.4218	<.0001	<.0001
7	30	102.3	<.0001	<.0001
10	15	118.7	<.0001	<.0001
10	30	41.2382	<.0001	<.0001
15	30	37.4942	<.0001	<.0001

Gráfico 8.2.1



Los resultados de los tests muestran que, dos a dos, son estadísticamente significativas las diferencias entre las curvas de supervivencia de los diferente grupos, excepto en el caso

de la franquicia 3 y la franquicia 10 (ver Tabla 8.5). Si nos fijamos, no obstante, en el test log-rank, el p-valor es < 0.0001, lo que nos indica que rechazamos la hipótesis de igualdad.

Visualmente, la diferencia es más acusada cuanto mayor es la franquicia. La franquicia 30 presenta una característica diferente al resto: al principio, la probabilidad de que un siniestro siga abierto es alta y CONSTANTE hasta cierto número de días. A partir de ahí, presenta una curva similar al resto de las franquicias, pero menos pronunciada. La curva para franquicia 30 y en menor medida para franquicia 15 muestra una tendencia a siniestros más largos. Coincide con su propia definición: hasta el día 15 ó 30 respectivamente, no reciben la indemnización diaria.

A su vez, el descenso en la probabilidad que un siniestro permanezca abierto después de un determinado número de días es mayor cuanto mayor es la franquicia de la póliza.

8.3. Curva log-rank variable código de enfermedad (sicenf2)

Mostramos las tablas de resultados, que comentaremos a continuación:

Tabla 8.3.1

Summary of the Number of Censored and Uncensored Values					
Stratum	sicenf2	Total	Failed	Censored	Percent Censored
1	487000	1874	1874	0	0.00
2	724000	4037	3998	39	0.97
3	726000	2733	2697	36	1.32
4	845000	2129	2124	5	0.23
Total		10773	10693	80	0.74

Tabla 8.3.2

Rank Statistics	
sicenf2	Log-Rank
487000	1482.4
724000	-644.2
726000	-1019.4
845000	181.2

Tabla 8.3.3

Covariance Matrix for the Log-Rank Statistics				
sicenf2	487000	724000	726000	845000
487000	338.29	-151.23	-106.66	-80.40
724000	-151.23	2531.32	-1581.84	-798.25
726000	-106.66	-1581.84	2313.03	-624.53
845000	-80.40	-798.25	-624.53	1503.18

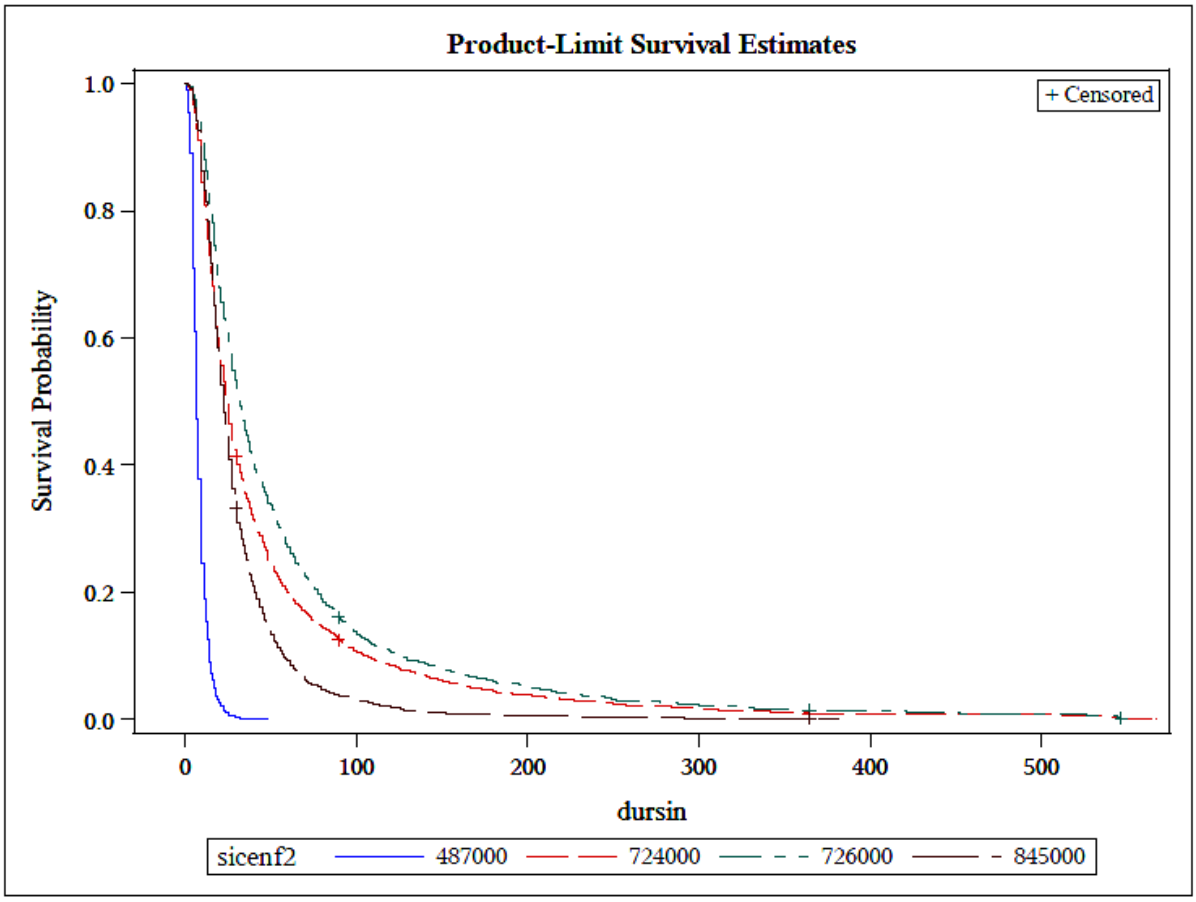
Tabla 8.3.4

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	6738.7539	3	<.0001

Tabla 8.3.5

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
sicenf2	sicenf2	Chi-Square	Raw	Sidak
487000	724000	1425.8	<.0001	<.0001
487000	726000	2185.0	<.0001	<.0001
487000	845000	845.6	<.0001	<.0001
724000	726000	17.5765	<.0001	0.0002
724000	845000	121.0	<.0001	<.0001
726000	845000	284.6	<.0001	<.0001

Gráfico 8.3.1



En este caso, diferencias entre las curvas de supervivencia son estadísticamente significativas, según el test log-rank general y los de múltiples comparaciones, realizados para contrastar la hipótesis nula de igualdad.

Si nos fijamos en el test log Rank, tenemos un p-valor < 0.0001. Nos indica que los tiempos de supervivencia (o duración de siniestros *dursin*) son significativamente distintos para los cuatro grupos de la variable sicenf2 (código de enfermedad)

A su vez, si revisamos los p-valores correspondientes en la tabla de múltiples comparaciones de Sidak, observamos que hay diferencias significativas entre las funciones de supervivencia de todos los grupos.

Visualmente, el resultado es el mismo. Claramente, cada código de enfermedad tiene su propia curva de supervivencia.

Destacar que en caso del código 487000, no hay probabilidad de que el siniestro siga abierto más allá de los 50 días aproximadamente, siendo estable a partir de unos 15-20 días aproximadamente. Este resultado ya nos aparecía, como comentamos, a la hora de comprobar la hipótesis de proporcionalidad. Es un caso irregular, pero no significa que el ajuste sea malo, hay que interpretarlo de forma correcta, tal como hemos explicado.

Análogamente (aunque no de una forma tan claramente diferenciada como en el caso anterior) ocurre con el 845000, no existiendo probabilidad de que ningún siniestro abierto a los 400 días de duración, y estabilizándose la probabilidad a partir de los 90 días aproximadamente.

8.4. Curva log-rank variable grupo de tarifa (gtar)

Los resultados obtenidos y su posterior análisis son los siguientes:

Tabla 8.4.1

Summary of the Number of Censored and Uncensored Values					
Stratum	gtar	Total	Failed	Censored	Percent Censored
1	A	1075	1070	5	0.47
2	B	2881	2863	18	0.62
3	C	4786	4737	49	1.02
4	D	1816	1808	8	0.44
5	S	215	215	0	0.00
Total		10773	10693	80	0.74

Tabla 8.4.2

Rank Statistics	
gtar	Log-Rank
A	94.54
B	-95.86
C	-219.97
D	238.46
S	-17.17

Tabla 8.4.3

Covariance Matrix for the Log-Rank Statistics					
gtar	A	B	C	D	S
A	856.09	-260.76	-435.27	-139.32	-20.75
B	-260.76	2069.80	-1327.45	-419.33	-62.26
C	-435.27	-1327.45	2568.23	-702.06	-103.45
D	-139.32	-419.33	-702.06	1293.68	-32.97
S	-20.75	-62.26	-103.45	-32.97	219.42

Tabla 8.4.4

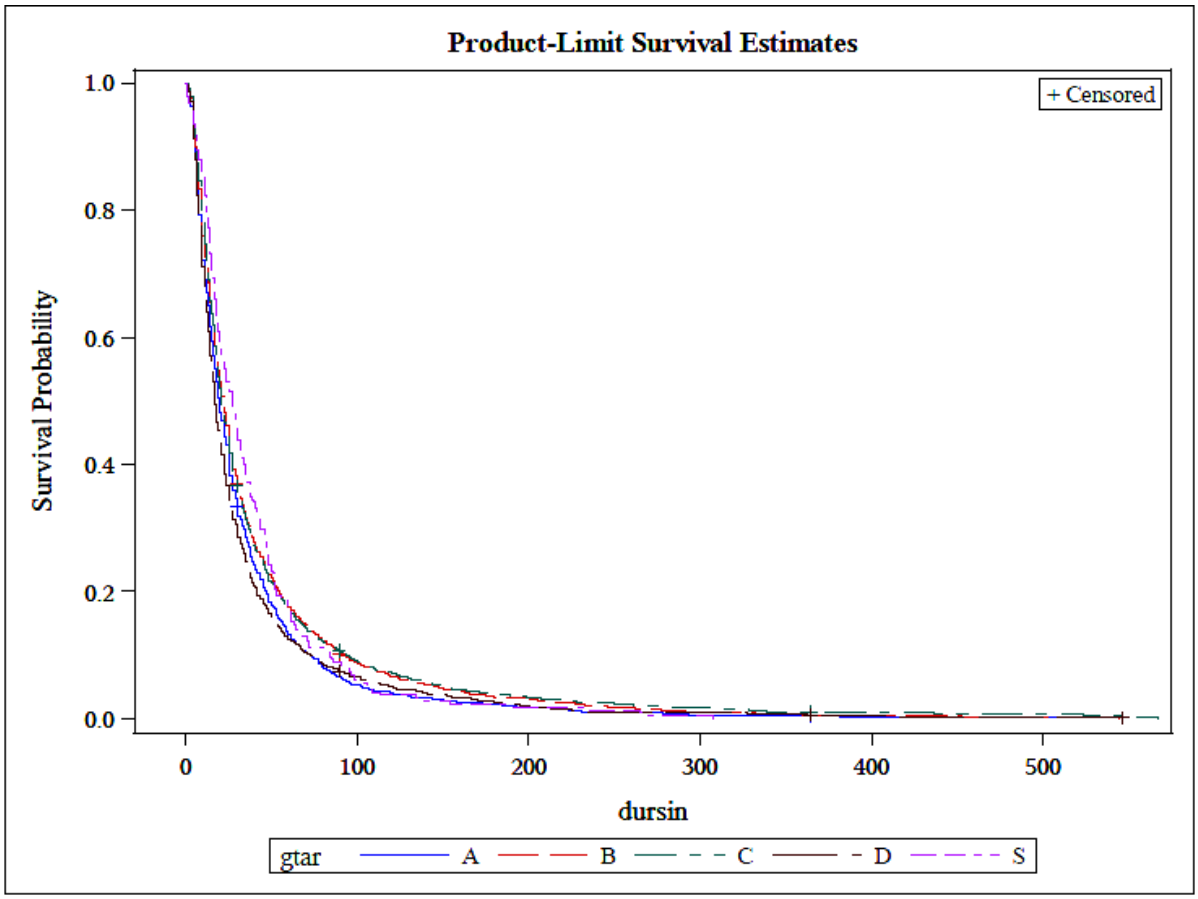
Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	61.6825	4	<.0001

Tabla 8.4.5

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
gtar	gtar	Chi-Square	Raw	Sidak
A	B	10.5153	0.0012	0.0118
A	C	23.0310	<.0001	<.0001
A	D	8.5298	0.0035	0.0344
A	S	11.1726	0.0008	0.0083
B	C	2.1121	0.1461	0.7940
B	D	26.5980	<.0001	<.0001
B	S	2.5649	0.1093	0.6856

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
gtar	gtar	Chi-Square	Raw	Sidak
C	D	39.9083	<.0001	<.0001
C	S	13.7334	0.0002	0.0021
D	S	41.3855	<.0001	<.0001

Gráfico 8.4.1



Atendiendo al test de comparaciones múltiples, para el grupo de tarifa B no podemos rechazar que tiene una curva de supervivencia igual que la de C y al S. La comparación de A y B tiene también un p-valor límite (0.0118), aunque el test general rechaza la hipótesis nula de igualdad. Análogo ocurre con la comparación entre A y D.

8.5. Curva log-rank variable sexo (sexo)

Los resultados obtenidos son los siguientes:

Tabla 8.5.1

Summary of the Number of Censored and Uncensored Values					
Stratum	sexo	Total	Failed	Censored	Percent Censored
1	1	8778	8713	65	0.74
2	2	1995	1980	15	0.75
Total		10773	10693	80	0.74

Tabla 8.5.2

Rank Statistics	
sexo	Log-Rank
1	15.537
2	-15.537

Tabla 8.5.3

Covariance Matrix for the Log-Rank Statistics		
sexo	1	2
1	1570.50	-1570.50
2	-1570.50	1570.50

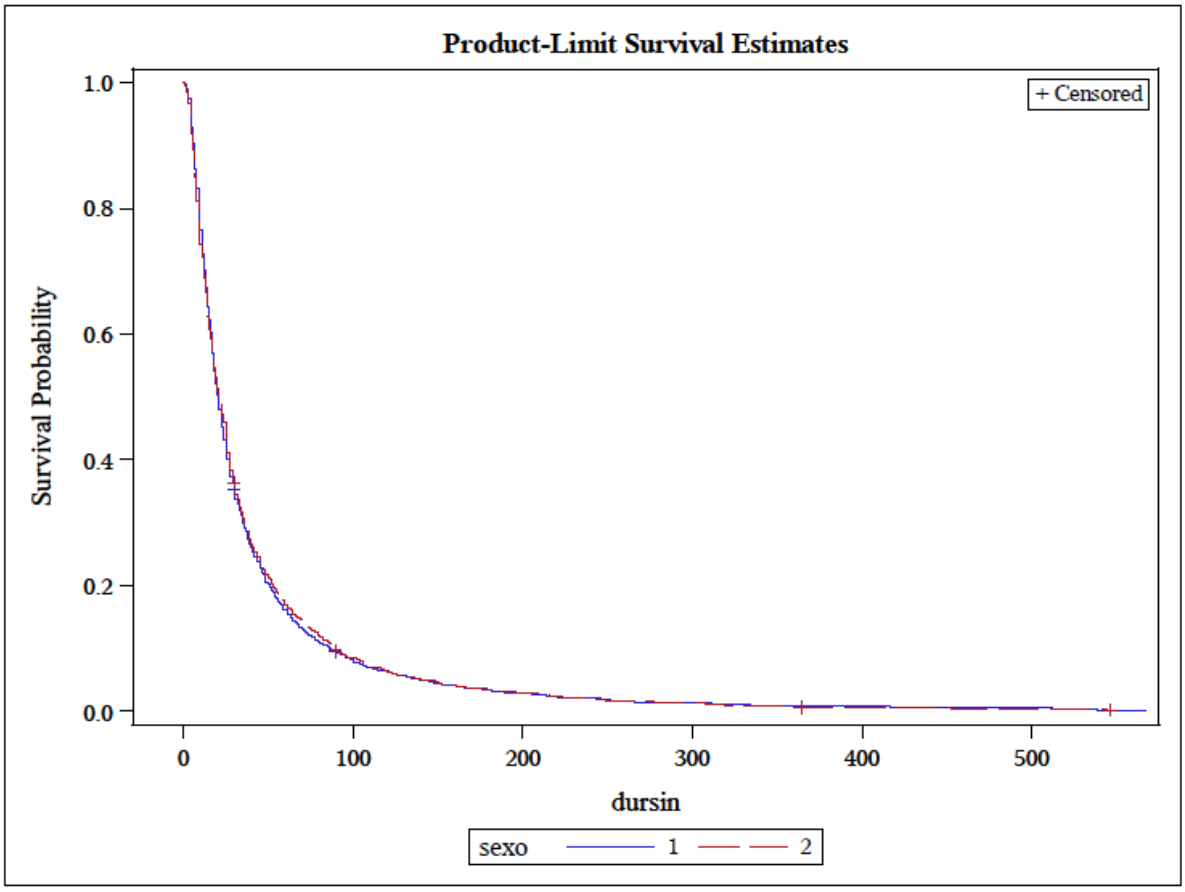
Tabla 8.5.4

Test of Equality over Strata			
Test	Chi-Square	DF	Pr > Chi-Square
Log-Rank	0.1537	1	0.6950

Tabla 8.5.5

Adjustment for Multiple Comparisons for the Logrank Test				
Strata Comparison			p-Values	
sexo	sexo	Chi-Square	Raw	Sidak
1	2	0.1537	0.6950	0.6950

Gráfico 8.5.1



Ambas curvas de supervivencia son iguales y muy similares a la curva general que hemos analizado en el apartado 7 del documento. La tabla 8.5.4 nos indica que no podemos rechazar la hipótesis general de igualdad, al igual que la tabla de múltiples comparaciones (8.5.5), con p-valores igual a 0.695 en ambos casos.

Para finalizar, el estudio de todas las curvas log-rank nos confirma la hipótesis de curvas de supervivencia diferentes para los distintos grupos de cada clase, excepto para el sexo, aunque con matices.

9. Análisis de residuos

Gráfico 9.1

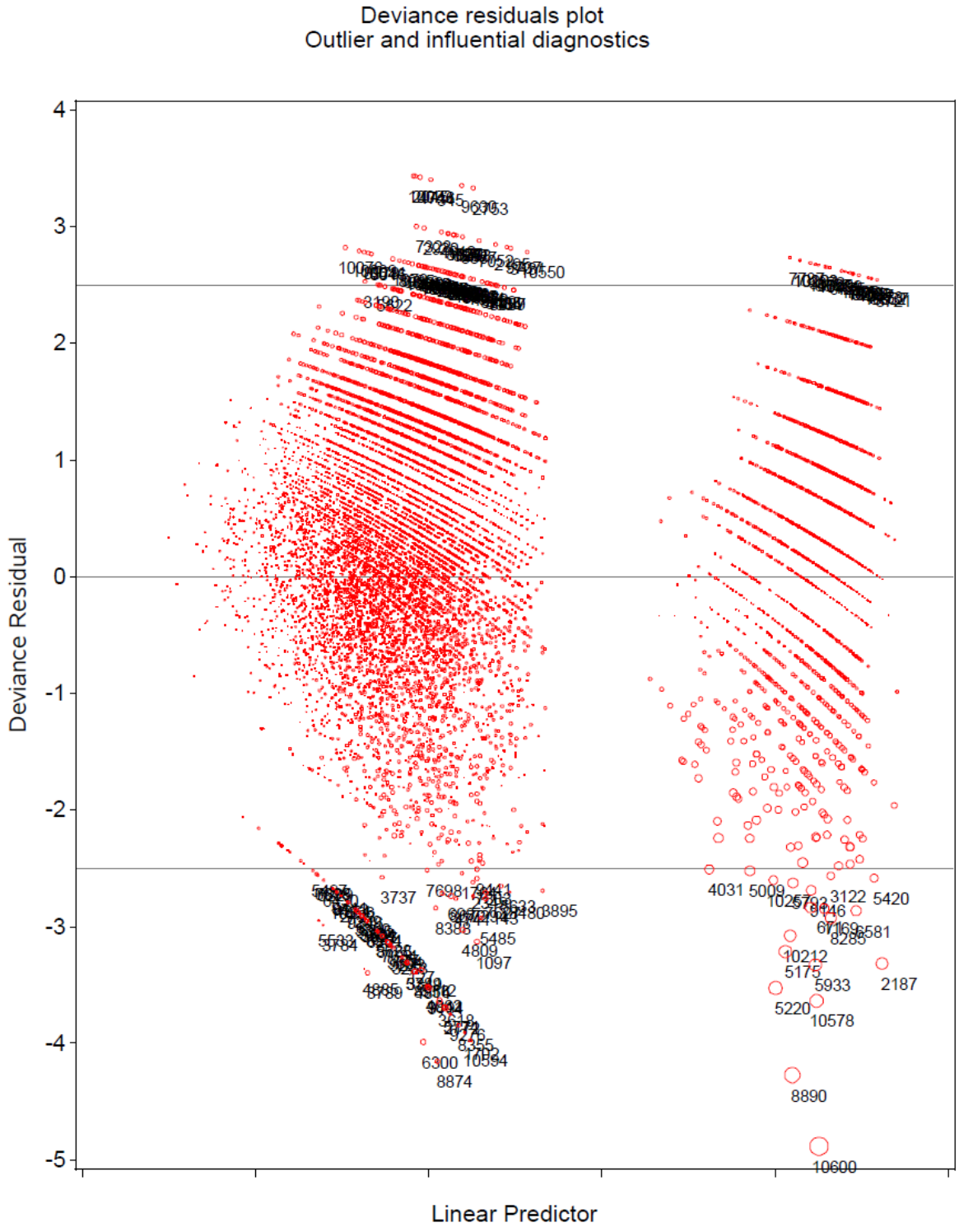
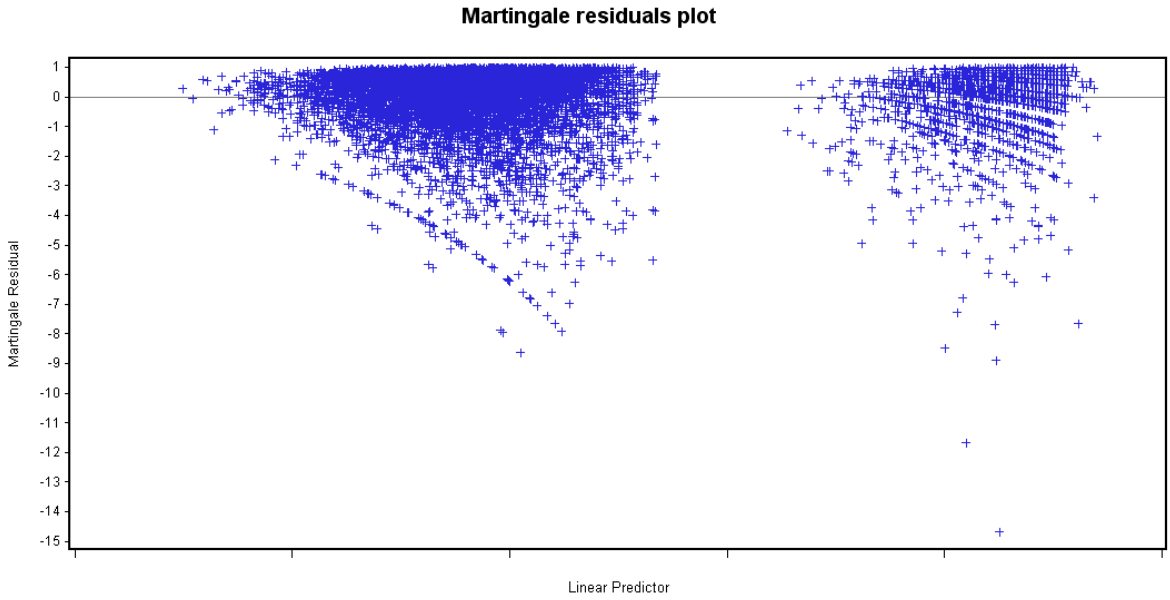


Gráfico 9.2



Estos son los gráficos de residuos obtenidos al ejecutar el procedimiento PHREG en SAS.

La información proporcionada es semejante. Podemos ver los dos grupos diferenciados en función del sexo que tienen un buen comportamiento general, sin apreciarse tendencias.

De una manera más evidente, distinguimos en el gráfico 9.1 las “líneas” que conforman los residuos de las diferentes combinaciones entre las diferentes variables y sus clases.

Sí que podemos apreciar la existencia de “outliers”, tanto por arriba como por abajo. Como el número no es muy elevado, hemos analizado toda la casuística a través de la salida de los betas que nos proporciona SAS. Los datos con valor negativo alto corresponden a siniestros anormalmente largos. Si miramos la duración definida de los siniestros según el código de enfermedad que tienen, deberían oscilar entre los 8 días para el código 487000 y los 51 del 726000. Aunque esto es un valor teórico, ilustra lo que estamos explicando. No olvidar que la mediana obtenida por Kaplan-Meier es de 22 días. Analizando los datos, como hemos dicho, vemos que el 78% de estos casos tienen una duración superior a 125 días. Si miramos Kaplan-Meier, tiene una probabilidad del 5.99% el que un siniestro tenga una duración superior a esos 125 días.

En el otro extremo, no encontramos con un número similar de siniestros con duración entre 1 y 3 días. Con Kaplan-Meier obtuvimos que el percentil 25 estaba en 11 días, y que la probabilidad de que un siniestro fuera mayor que 3 días era de 97.5 %. Inclina a pensar que el modelo no realiza estimaciones muy finas para siniestros cortos, sobre todo en cuanto mayor es la edad de la persona. Recordemos también que, estudiando los Hazard Ratio, la probabilidad de supervivencia de un siniestro hasta el siguiente seguimiento aumentaba con la edad.

10. Conclusiones finales

- El modelo seleccionado es un modelo que explica la duración de los siniestros en función de las variables edad2, gtar, franq, sexo y sicenf2.

- Respecto a la edad, disminuye el riesgo en un 1.17% por incremento de un año. Esto es, cuanto mayor es la edad, mayor será la duración del siniestro.

- El grupo de tarifa S multiplica el riesgo respecto al de referencia en un 16.7%, encontrando ligeras reducciones en los grupos de tarifa B y D. Respecto al grupo de tarifa C, el de referencia tiene una reducción relativa un poco mayor, del 3.3%.

- En las franquicias, destacar la reducción del riesgo relativo de la franquicia 0 respecto a las 15 y 30 del 48.1 % y del 59.5% respectivamente. Franquicia 3 y 10 tienen un comportamiento similar frente a la de referencia, y la franquicia 0 (referencia) tiene un riesgo del 65% respecto a la 7, un valor algo atípico, ya que esperaríamos un comportamiento más alineado con la franquicia 7 y la franquicia 10.

- El riesgo del sexo hombre es un 90.9% que el de sexo mujer. Tiene una reducción relativa del riesgo del 9.1%. En nuestros términos, estamos hablando de siniestros con una duración ligeramente más corta (en realidad, es lo que nos interesa, ya que tendríamos que indemnizar durante menos días)

- Los códigos de enfermedad tienen, respecto al de referencia, un comportamiento muy similar. Tienen reducciones relativas del riesgo bastante significativas. Por tanto, estaremos hablando de duraciones de siniestro bastante más largas que la de referencia.

Bibliografía:

A. Martín Andres, J.D. Luna del Castillo, 2004, “Bioestadística para las Ciencias de la Salud”

Antonio Rial, Jesús Varela, Antonio J. Rojas, 2001, “Depuración y Análisis Preliminares de Datos en SPSS”

Jesús Abaurrea, Ana Carmen Cebrián, 2005, “Análisis de Supervivencia”

David G. Kleinbaum, 1995, “Survival Analysis A Self-Learning Text”

Paul D. Allison, 2010, “Survival Analysis Using SAS”

SAS Help and Documentation

Efron, 1977

Collett, 2003

Cuadernos de Trabajo
Facultad de Estudios Estadísticos

CT02/2013 **Consumer need for touch and Multichannel Purchasing Behaviour.**
R. Manzano, M. Ferrán, D. Gavilán

CT01/2013 **Un método gráfico de comparación de series históricas en el mercado bursátil.**
Magdalena Ferrán Aranaz

CT03/2012 **Calculando la matriz de covarianzas con la estructura de una red Bayesiana Gaussiana** *Miguel A. Gómez-Villegas y Rosario Susi*

CT02/2012 **What's new and useful about chaos in economic science.**
Andrés Fernández Díaz, Lorenzo Escot and Pilar Grau-Carles

CT01/2012 **A social capital index**
Enrique González-Arangüena, Anna Khmelnitskaya, Conrado Manuel, Mónica del Pozo

CT04/2011 **La metodología del haz de rectas para la comparación de series temporales.**
Magdalena Ferrán Aranaz

CT03/2011 **Game Theory and Centrality in Directed Social Networks**
Mónica del Pozo, Conrado Manuel, Enrique González-Arangüena y Guillermo Owen.

CT02/2011 **Sondeo de intención de voto en las elecciones a Rector de la Universidad Complutense de Madrid 2011**
L.Escot, E. Ortega Castelló y L. Fernández Franco (coords)

CT01/2011 **Juegos y Experimentos Didácticos de Estadística y Probabilidad**
G. Cabrera Gómez y Mª.J. Pons Bordería

CT04/2010 **Medio siglo de estadísticas en el sector de la construcción residencial**
M. Ferrán Aranaz

CT03/2010 **Sensitivity to hyperprior parameters in Gaussian Bayesian networks.**
M.A. Gómez-Villegas, P. Main, H. Navarro y R. Susi

CT02/2010 **Las políticas de conciliación de la vida familiar y laboral desde la perspectiva del empleador. Problemas y ventajas para la empresa.**
R. Albert, L. Escot, J.A. Fernández Cornejo y M.T. Palomo

CT01/2010 **Propiedades exóticas de los determinantes**
Venancio Tomeo Perucha

CT05/2009 **La predisposición de las estudiantes universitarias de la Comunidad de Madrid a auto-limitarse profesionalmente en el futuro por razones de conciliación**
R. Albert, L. Escot y J.A. Fernández Cornejo

CT04/2009 **A Probabilistic Position Value**
A. Ghintran, E. González–Arangüena y C. Manuel

CT03/2009 **Didáctica de la Estadística y la Probabilidad en Secundaria: Experimentos motivadores**
A. Pajares García y V. Tomeo Perucha

CT02/2009 **La disposición entre los hombres españoles a tomarse el permiso por nacimiento. ¿Influyen en ello las estrategias de conciliación de las empresas?**
L. Escot, J.A. Fernández-Cornejo, C. Lafuente y C. Poza

CT01/2009 **Perturbing the structure in Gaussian Bayesian networks**
R. Susi, H. Navarro, P. Main y M.A. Gómez-Villegas

CT09/2008 **Un experimento de campo para analizar la discriminación contra la mujer en los procesos de selección de personal**
L. Escot, J.A. Fernández Cornejo, R. Albert y M.O. Samamed

CT08/2008 **Laboratorio de Programación. Manual de Mooshak para el alumno**
D. I. de Basilio y Vildósola, M. González Cuñado y C. Pareja Flores

CT07/2008 **Factores de protección y riesgo de infidelidad en la banca comercial**
J. Mª Santiago Merino

CT06/2008 **Multinationals and foreign direct investment: Main theoretical strands and empirical effects**
María C. Latorre

CT05/2008 **On the Asymptotic Distribution of Cook's distance in Logistic Regression Models**
Nirian Martín y and Leandro Pardo

CT04/2008 **La innovación tecnológica desde el marco del capital intelectual**
Miriam Delgado Verde, José Emilio Navas López, Gregorio Martín de Castro y Pedro López Sáez

CT03/2008 **Análisis del comportamiento de los indecisos en procesos electorales: propuesta de investigación funcional predictivo-normativa**
J. Mª Santiago Merino

CT02/2008 **Inaccurate parameters in Gaussian Bayesian networks**
Miguel A. Gómez-Villegas, Paloma Main and Rosario Susi

CT01/2008 **A Value for Directed Communication Situations.**
E. González–Arangüena, C. Manuel, D. Gómez, R. van den Brink

UNIVERSIDAD COMPLUTENSE
MADRID