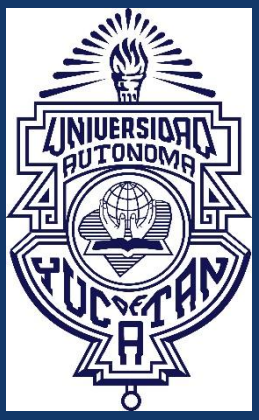


Universidad Autónoma de Yucatán

Facultad de Matemáticas

Análisis de Supervivencia

Proyecto Final

Jorge Miguel Guzmán Arjona

Introducción

En este informe plantearemos un modelo para comprender la mortalidad de las personas en edad avanzada. Esto con el objetivo de permitir a las aseguradoras mejorar el entendimiento y ofrecer mejores productos que vayan acorde a las necesidades financieras de los usuarios y de su perfil de riesgo.

Para este estudio la variable tiempo de vida será el tiempo en semanas que transcurre desde la fecha de emisión de la póliza hasta el momento en el que fallece el asegurado. Asumimos que el tiempo es en semanas puesto que las pólizas de la base se recopilaron entre el año 1997 y 2000, además el valor máximo para la variable tiempo es 1000, lo cual si se trata de semanas corresponde a 19 años. El evento de interés es la muerte del asegurado.

Existen muchos factores que afectan la mortalidad de las personas, para este estudio intentaremos crear un modelo considerando las siguientes covariables:

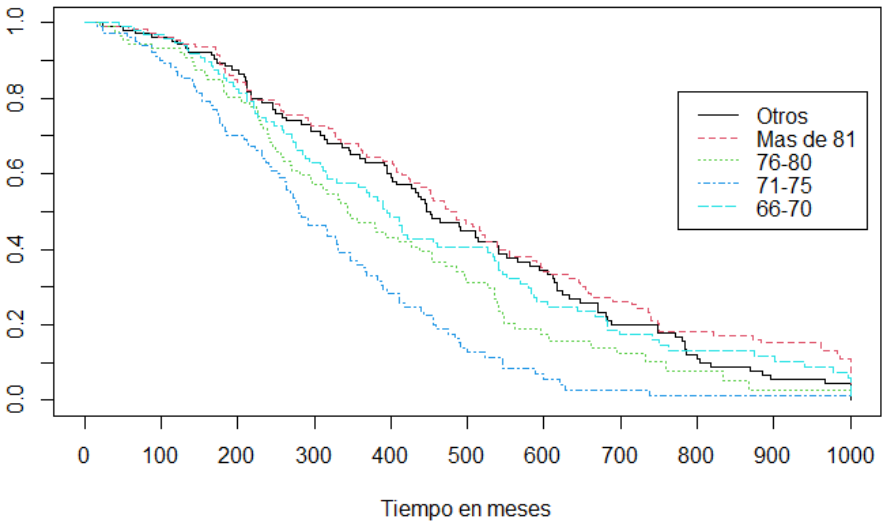
- **Edad**: Cinco categorías (66-70 años, 71-75 años, 76-80 años, más de 81 años y Otros)
- **Sexo**: Dos categorías (hombre y mujer)
- **Fuma**: Dos categorías (fumador y no fumador)
- **Producto**: Cuatro categorías (Term, Whole Life, Universal Life, Other)

Debido a que en nuestro estudio existen personas que aún se encuentran con vida los datos del tiempo de vida presentan censura. También puede ocurrir que la persona no fallezca, pero se cancele su contrato. Por estas razones optaremos por utilizar un análisis de supervivencia.

Análisis exploratorio

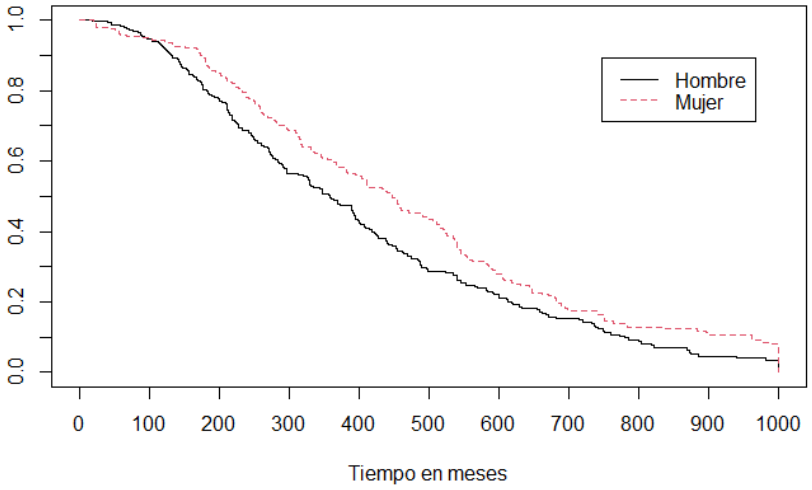
Graficaremos las funciones de supervivencia estimadas por Kaplan-Meier por cada covariable y sus respectivas subpoblaciones.

Edad



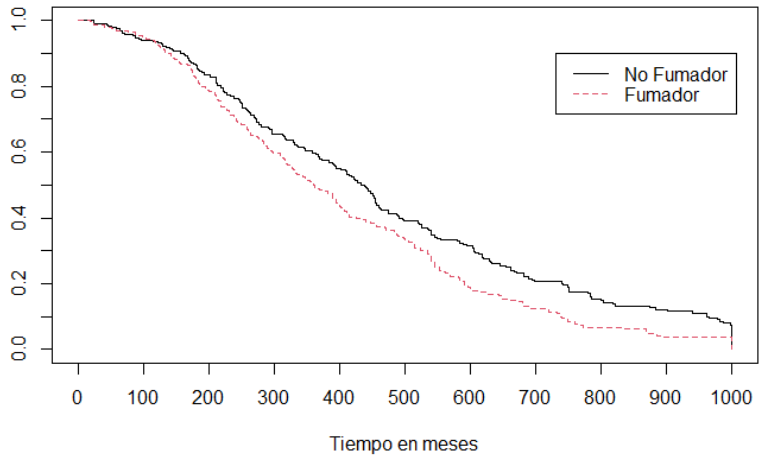
La función de supervivencia que parece ser diferente al resto es la de las personas de 71 a 75 años, pues todas las demás funciones quedan por encima de ella. La función que parece quedar por encima de todas en la mayoría de los puntos es la de las personas mayores de 81 años. Lo cual significaría que para cualquier tiempo t es más probable que una persona mayor de 81 años alcance dicho tiempo de supervivencia en comparación a cualquier otra subpoblación, en este caso de menor edad. Lo último tiene sentido pues si la persona tiene mayor edad se esperaría que su póliza haya acumulado un mayor tiempo en vigencia.

Sexo



La función de supervivencia de las mujeres se encuentra por encima de la función de supervivencia de los hombres. Es decir, dado un tiempo t, la probabilidad de supervivencia para una mujer es mayor a la de un hombre, en términos del problema se espera que la póliza de una mujer se encuentre un mayor período de tiempo en vigencia que la de un hombre.

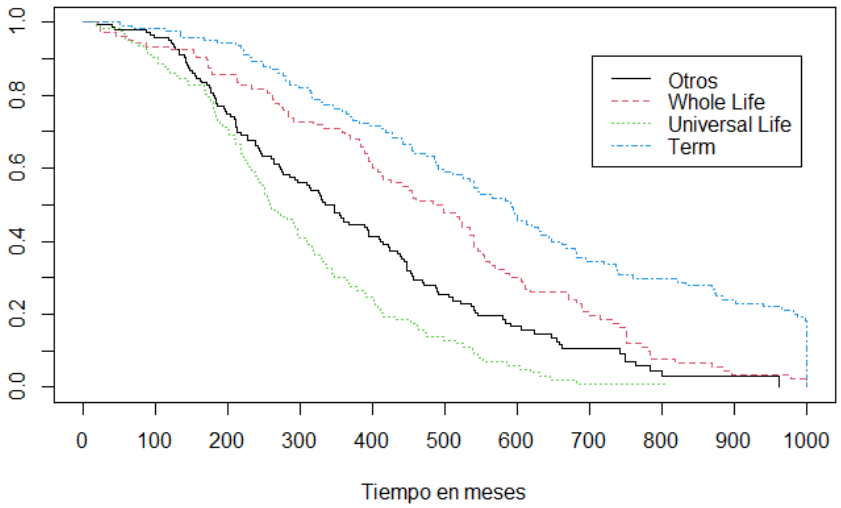
Fuma



En este caso observamos claramente que la función de supervivencia de los no fumadores se encuentra por encima de la función de supervivencia de los fumadores, lo cual significa que dado un tiempo t la probabilidad de

supervivencia para un no fumador es mayor a la de un fumador, en términos del problema se espera que la póliza de un fumador se encuentre un mayor período de tiempo en vigencia, lo cual está directamente relacionado a la supervivencia del asegurado.

Producto



Para esta covariable podemos observar que las funciones de supervivencia de las subpoblaciones no parecen cruzarse mucho. También podemos ver con claridad que la función de supervivencia para el tipo de producto “Term” se encuentra por encima de todas las demás. Asimismo, observamos que esta última curva cae a probabilidad cero en los mil meses de tiempo, esto puede deberse a que este tipo de producto tiene por definición una cota para la duración de la póliza.

Como la curva de supervivencia de la subpoblación “Term” se encuentra por encima de las demás, podemos decir que dado un tiempo t, la probabilidad de supervivencia de una persona con un producto “Term” es mayor a la probabilidad de las personas con otros productos.

Modelo de Cox

Paso 1

Estableceremos modelos simples con cada covariable para encontrar cuales son significativas. Para aquellas con más de 2 factores intentaremos encontrar las combinaciones que tengan más componentes significativos.

> **Edad:** Al tener 5 categorías utilizaremos 4 variables dummy. Al utilizar como variables dummy las categorías de las personas correspondientes a las edades 66-70, 76-80, más de 81 y otros, obtuvimos que todas fueron significativas, por lo que nuestro riesgo base corresponde a la edad de 71-75.

```
                  coef exp(coef) se(coef)      z        p
Edad66.70     -0.6642    0.5147   0.1529 -4.345 1.40e-05
EdadOtros     -0.7388    0.4777   0.1488 -4.965 6.86e-07
Edad76.80     -0.4191    0.6576   0.1576 -2.659  0.00784
EdadMas.de.81 -0.8852    0.4126   0.1489 -5.947 2.73e-09

Likelihood ratio test=38.4  on 4 df, p=9.264e-08
```

Todos los p valores tanto individuales como el global son menores a 0.05, las variables son significativas.

Sexo:

```
             coef exp(coef) se(coef)      z       p
SexoMujer -0.2588    0.7720   0.0947 -2.733 0.00628

Likelihood ratio test=7.47  on 1 df, p=0.006265
```

El p valor resultó ser 0.00628<0.05 por lo que la covariable es significativa.

Fumar:

```
           coef exp(coef) se(coef)     z       p
FumarSi 0.26840   1.30787  0.09538 2.814 0.00489

Likelihood ratio test=7.93  on 1 df, p=0.004851
```

El p valor resultó ser 0.00489<0.05 por lo que la covariable es significativa.

Producto: Esta covariable tiene 4 categorías por lo que se utilizaran 3 variables dummy. Al utilizar como variables dummy las categorías de los productos Term, Universal Life y Whole Life, obtuvimos que todas fueron significativas, por lo que nuestro riesgo base corresponde al producto Otros.

```
                     coef exp(coef) se(coef)      z        p
ProductoTERM       -0.9883    0.3722   0.1411 -7.004 2.49e-12
ProductoU.L.        0.4640    1.5905   0.1305  3.556 0.000377
ProductoWHOLE.LIFE -0.4441    0.6414   0.1364 -3.255 0.001135

Likelihood ratio test=106.3  on 3 df, p=< 2.2e-16
```

Podemos observar que todas las variables obtuvieron un valor p menor a 0.05, por lo que son significativas.

Paso 2

Del paso 1 obtuvimos que todas las covariables fueron significativas así que plantearemos un modelo que las contenga a todas para observar cuales mantienen su significancia en presencia de las otras.

	coef	exp(coef)	se(coef)	z	p
Edad66.70	-0.94531	0.38856	0.15775	-5.992	2.07e-09
EdadOtros	-1.17085	0.31010	0.15730	-7.443	9.81e-14
Edad76.80	-0.73240	0.48075	0.16071	-4.557	5.18e-06
EdadMas.de.81	-1.17337	0.30932	0.15281	-7.679	1.61e-14
ProductoTERM	-1.20580	0.29945	0.14565	-8.279	< 2e-16
ProductoU.L.	0.57704	1.78075	0.13155	4.387	1.15e-05
ProductoWHOLE.LIFE	-0.51532	0.59731	0.13836	-3.724	0.000196
SexoMujer	-0.29009	0.74819	0.09631	-3.012	0.002593
FumarSi	0.29932	1.34894	0.09643	3.104	0.001910

Likelihood ratio test=195.1 on 9 df, p=< 2.2e-16

Al generar el modelo observamos que todos los valores p son menores a 0.05 por lo que todas las covariables mantuvieron su significancia en presencia de las otras.

Paso 3
Como ninguna variable dejó de ser significativa en el paso 2 entonces se omite este paso. También omitiremos el uso de interacciones para mantener simple nuestro modelo.

Paso 4
Nuestro modelo final es el mencionado en el paso 2, todas las covariables fueron significativas.

Validación

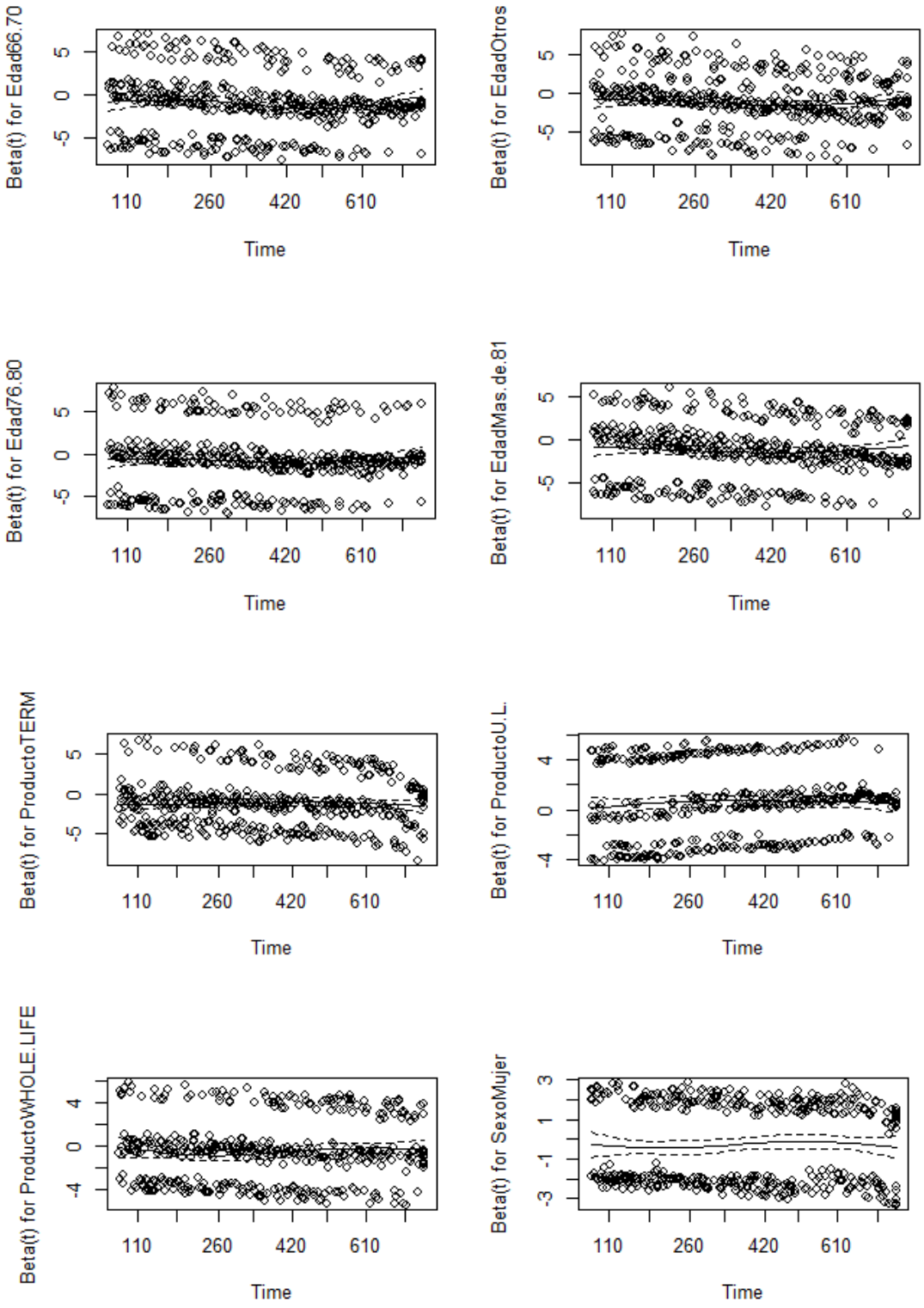
Riesgos proporcionales
Verificaremos el supuesto de riesgos proporcionales para cada una de nuestras covariables, eso con el fin de validar el modelo. Se tienen las siguientes hipótesis:

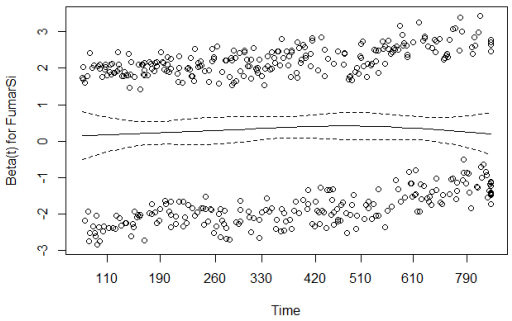
Hipótesis nula: Existe correlación 0 entre los residuales y el tiempo (existen riesgos proporcionales).
Hipótesis alternativa: La correlación entre los residuales y el tiempo es diferente de 0.

Realizamos la prueba de hipótesis en R y obtenemos los siguientes resultados:

	chisq	df	p
Edad66.70	0.46767	1	0.49
EdadOtros	0.00531	1	0.94
Edad76.80	0.04131	1	0.84
EdadMas.de.81	0.01003	1	0.92
ProductoTERM	0.48392	1	0.49
ProductoU.L.	0.10098	1	0.75
ProductoWHOLE.LIFE	1.39652	1	0.24
SexoMujer	0.24857	1	0.62
FumarSi	0.17677	1	0.67
GLOBAL	2.76758	9	0.97

Los p valores son mayores a 0.05 para todas las pruebas individuales y también para la prueba global, por lo tanto, no hay evidencia suficiente para rechazar la hipótesis nula, es decir, existen riesgos proporcionales.
La gráfica de dispersión de residuos nos arroja el siguiente resultado:





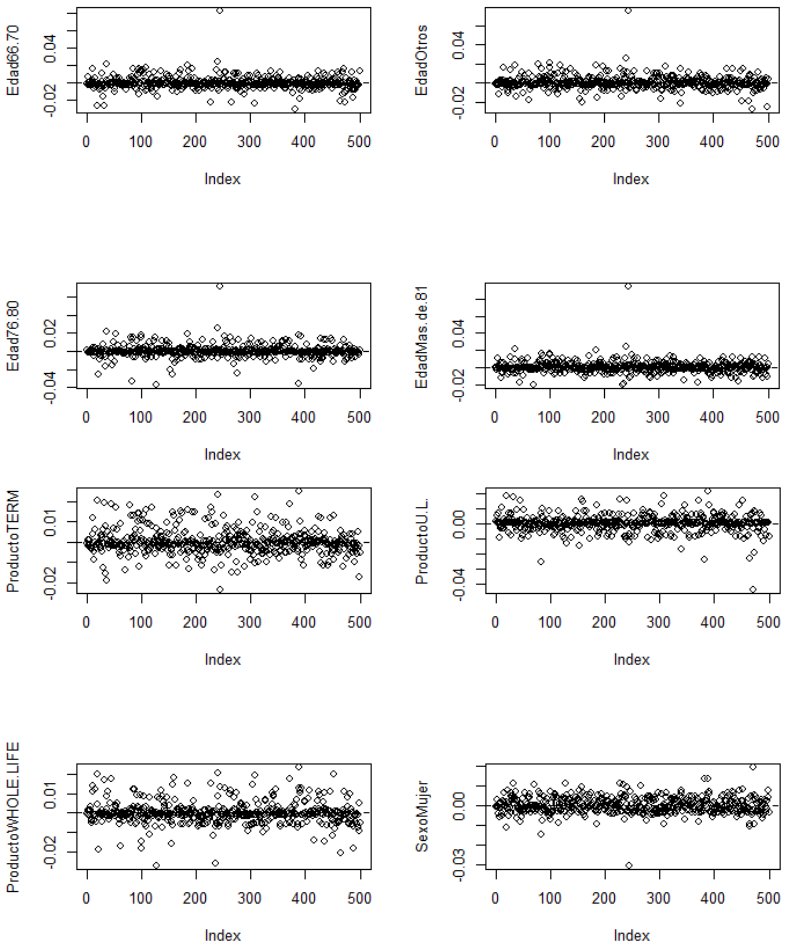
Las curvas observadas se aproximan a líneas rectas horizontales, lo cual coincide con lo obtenido con las pruebas de hipótesis.

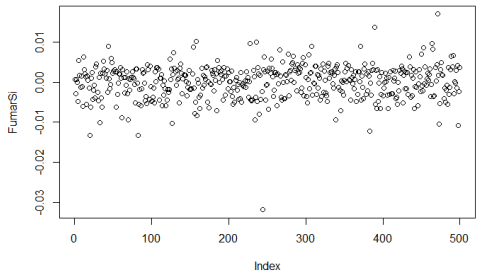
Linealidad

Como no contamos con covariables continuas omitimos esta verificación

.

Datos influyentes





Los datos parecen comportarse de manera estable y no parece haber datos outliers que puedan arruinar el modelo significativamente. Para mayores detalles habría que hablar con un experto.

Interpretación del modelo

Denotaremos las covariables de la siguiente manera:

- W_1: Edad 66-70, W_2: Edad 76-80, W_3: Edad mayor a 81, W_4: Otros (Diferente de 71-75)
- X**:** Hombre (0), Mujer (1)
- Y: No fuma (0), Fuma (1)
- Z_1**:** Producto Term, Z_2**:** Producto Universal Life, Z_3**:** Producto Whole Life

Por lo tanto, la ecuación resultante de nuestro modelo es la siguiente:

$$\boldsymbol{h(t;G) = h_0(t)exp\{-0.9453W_1 - 0.7324W_2 - 1.1734W_3 - 1.1709W_4 - 0.2901X + 0.2993Y - 1.2058Z_1 + 0.5770Z_2 - 0.5153Z_3\}}$$

$$\boldsymbol{G} = (W_1, W_2, W_3, W_4, X, Y, Z_1, Z_2, Z_3)$$

A continuación, compararemos algunas subpoblaciones:

Ejemplo 1

$G_1 = (1, 0, 0, 0, 1, 1, 0, 1, 0)$
Grupo 1: Mujer de edad entre 66 y 70 años, fumadora, el producto es Universal Life.

$G_2 = (1, 0, 0, 0, 1, \mathbf{0}, 0, 1, 0)$
Grupo 2: Mujer de edad entre 66 y 70 años, **no fumadora**, el producto es Universal Life.

$$\frac{\boldsymbol{h(t;G_1)}}{\boldsymbol{h(t;G_2)}} = \boldsymbol{exp\{0.2993\} = 1.3489 > 1}$$

El riesgo para el grupo 1 es mayor que para el grupo 2. Dado un tiempo t, el riesgo para una persona del grupo 1 es 1.3489 veces mayor que para una persona del grupo 2. La probabilidad de supervivencia después de un tiempo t, es menor para el grupo 1 que para el grupo 2.

Estos resultados son los esperados pues las personas del grupo 1 presentan la característica de ser fumadoras.

Ejemplo 2

$G_3 = (0, 0, 1, 0, 0, 1, 0, 1, 0)$
Grupo 3: Hombre de edad mayor a 81 años, fumador, el producto es Universal Life.

$G_4 = (\mathbf{1}, 0, \mathbf{0}, 0, \mathbf{1}, \mathbf{0}, 0, 1, 0)$
Grupo 4: Mujer de edad entre 66 y 70 años, no fumadora, el producto es Universal Life.

$$\frac{\boldsymbol{h(t; G_3)}}{\boldsymbol{h(t; G_4)}} = \boldsymbol{exp}\{\mathbf{0.9453 - 1.1734 + 0.2901 + 0.2993}\} = \mathbf{1.4352 > 1}$$

El riesgo para el grupo 3 es mayor que para el grupo 4. Dado un tiempo t, el riesgo para una persona del grupo 3 es 1.4352 veces mayor que para una persona del grupo 4. La probabilidad de supervivencia después de un tiempo t, es menor para el grupo 3 que para el grupo 4.

Conclusión

Las variables utilizadas para el modelo nos permitieron distinguir entre subpoblaciones con diferentes riesgos asociados. Ahora que conocemos cuales son los grupos más riesgosos y en qué magnitud lo son, ya podemos establecer mejores productos para cada uno de ellos.

Hemos mejorado el entendimiento de los factores de riesgo con base a las covariables analizadas y con ello podemos clasificar a las personas según su perfil de riesgo.