APLICACIÓN DEL MODELO DE RIESGOS PROPORCIONALES DE COX A UN GRUPO DE PERSONAS EN EDAD AVANZADA ASEGURADAS

Verónica Bañuelos Martínez*, Julio A. de la Cruz Canul Medina*, Cristhian A. Díaz Chablé*.

* *Universidad Autónoma de Yucatán. Facultad de Matemáticas.*

Resumen

La mortalidadd de las personas en edad avanzada está influenciada por diversos factores de riesgo como son la edad, sexo, tabaquismo, entre otros. En el presente reporte se aplicó el modelo de riesgos proporcionales de Cox para estimar no sólo la mortalidad de una población de edad avanzada asegurada, sino también el efecto que tienen ciertos factores en relación con dicha mortalidad. La base de datos utilizada fue proporcionada por el maestro en el curso de Análisis de Supervivencia.

Palabras clave: *Seguros de vida, Seguro para fumadores,Tiempo de vida, Modelo de Cox.*

1. Introducción

Tanto para sistemas de pensiones como para seguros de vida, la mayor inquietud es el poder estimar los tiempos de vida de las personas inscritas en estos sistemas, ya que dicha estimación es la pieza clave de su costeo.

No es un tema nuevo el hablar de la tendencia positiva de la esperanza de vida alrededor del globo, debida a diversos factores como los avances en la tecnología, mayor acceso a la salud, mejor calidad de vida, etc. Por supuesto, conforme la esperanza de vida se modifica también lo hace la industria aseguradora.

En el campo de seguros de vida, por ejemplo, dicha tendencia podría llevar a un aumento del número de clientes potenciales, ya que al disminuir la mortalidad en edades avanzadas, se incrementa el límite de edad en que es factible brindar un seguro de vida a un segmento de la población.

Sin embargo, antes de aceptar emitir estas pólizas, se deben realizar los estudios pertinentes para entender mejor la mortalidad de la población de edad avanzada.

En el siguiente reporte, pretendemos tener un mejor acercamiento de dicha mortalidad al aplicar un modelo de riesgos proporcionales de Cox. Ya que ésta se ve influenciada por diversos factores como cualquier otra población, tales como el género, edad, tabaquismo, entre otros. Mediante el uso de software estadístico R, no sólo buscamos estimar la mortalidad si no también el efecto que tienen algunos factores en el tiempo de vida de los asegurados. Definiendo como tiempo de vida el número de meses desde la fecha de emisión de una póliza hasta la fecha de fallecimiento del titular de la póliza.

Los productos (seguros) que fueron parte del estudio consisten en lo siguiente:

- Term: seguro de vida que garantiza el pago de un beneficio por fallecimiento declarado durante un plazo especificado. Conocido en español como seguro de vida temporal. [1]

- Whole life: seguro de vida que proporciona cobertura para el titular del contrato durante toda su vida. Ante la muerte inevitable del titular, el pago del seguro se realiza a los beneficiarios del contrato. Conocido en español como ordinario de vida. [2]

- Universal life: seguro de vida permanente con un

elemento de ahorro de inversión y primas bajas como el seguro de vida temporal. La mayoría de las pólizas de seguro de vida universal contienen una opción de prima flexible. Sin embargo, algunos requieren una prima única o primas fijas programadas. Traducido al español como seguro de vida universal. [3]

- Other: cualquier otro tipo de seguro de vida.

Se cuenta con una base de datos proporcionada por el profesor del curso de Análisis de Supervivencia basada en la utilizada en el documento titulado "Estimating Mortality of Insured Advanced-age Population With Cox Regression Model". Dicha base de datos esta compuesta por las siguientes variables:

- `Fumar`: Variable indicadora que denota si el asegurado fuma (valor 0) o no fuma (valor 1)
- `Producto`: Indica el tipo de seguro que el asegurado tiene, toma valores de 1,2,3 y 4, para Term, Whole life, Universal life y Other respectivamente.
- `Sexo`: Indica si el asegurado es hombre (valor 1) o mujer (valor 0)
- `Edad`: Indica el número de años cumplidos del asegurado en el momento que se emitió la póliza.
- `Censura`: Indica si hubo censura por la derecha (valor 0) al recopilar los datos o si ocurrió el evento (valor 1).
- `Tiempo`: número de meses desde la fecha de emisión de la póliza hasta la fecha de fallecimiento del titular de la póliza o de la que se tiene último registro.

Definimos como tiempo de vida al número de meses desde la fecha de emisión de una póliza hasta la fecha de fallecimiento del titular de la póliza. El evento de interés entonces es la reclamación de la póliza.

Cabe mencionar que al haberse realizado el estudio en una ventana de tiempo de 123 meses, posiblemente nos encontraremos con datos censurados. Ya que el Análisis de supervivencia es una herramienta estadística ideal para modelar variables del tipo "tiempo hasta que ocurre un evento de interés", más aún, cuando hay censura en los datos, es una herramienta adecuada para llegar al objetivo de este estudio.

1.1. Metodología y cuestiones técnicas

A continuación se hablará acerca de ciertas definiciones como son curva de supervivencia, función de riesgo, modelo de Cox, riesgo proporcionales y otras cuestiones técnicas a usar.
Primero definimos el concepto de función de supervivencia,

Definición 1.1 *Sea $t > 0$ y sea T una v.a. positiva, entonces se define $S_T(t)$ como la función de supervivencia de la variable T tal que:*

$$S_T(t) = \int_t^\infty f_T(t)dt$$

de esta función podemos deducir la función de riesgo (Hazard Risk),

Definición 1.2 *Sea $t, T, S_T(t)$ como en (1.1), entonces se verifica que,*

$$h_T(t) = \frac{d[\ln(S_T(t))]}{dt}$$

donde $h_T(t)$ es la función de riesgo de la v.a. T

Por otro lado, el modelo de Cox se define de la siguiente manera,

Definición 1.3 *Sea $X = (X_1, X_2, ..., X_n)'$ el vector con las covariables de interés y $\beta = (\beta_1, \beta_2, ..., \beta_p)$. Entonces el modelo de riesgos proporcionales de Cox, se define como:*

$$\begin{aligned} h_T(t; X) &= h_0(t) \exp[\beta * X] \\ &= h_0(t) \exp\left[\sum_{i=1}^{n} \beta_i * X_i\right] \end{aligned}$$

La función h_0 se interpreta como la función de riesgo de los individuos cuando el valor de todas las covariables es cero.

Cabe mencionar que $\beta = (\beta_1, \beta_2, ..., \beta_p)$ es el vector de coeficientes a ser estimado.

De lo anterior podemos deducir que se cumple lo siguiente,

$$\begin{aligned} h_T(t; X) &= -\frac{d[\ln[S_T(t)]]}{dt} \\ &= h_0(t) \exp[\beta * X] \\ &= -\frac{d[\ln[S_0(t)]]}{dt} \exp[\beta * X] \end{aligned}$$

Integrando de 0 a t y al notar que $\ln(1) = 0$ obtenemos que,

$$\begin{aligned}\ln[S_T(t)] &= \ln[S_0(t)]\exp[\beta * X] \\ &= \ln[S_0(t)^{\exp[\beta * X]}]\end{aligned}$$

Por lo que

$$S_T(t) = S_0(t)^{\exp[\beta * X]} \text{ , t}\geq 0$$

La función de supervivencia de una muestra con censura por la derecha puede ser estimada vía Máxima Verosimilitud (Kaplan Meir), y por el resultado anterior para riesgos proporcionales, se puede estimar usando dicha curva de supervivencia, una nueva curva de supervivencia que contemple los nuevos factores de riesgo.

2. Análisis Exploratorio

Como parte de un preprocesamiento de datos, se realizó una búsqueda de errores en los registros, así como posibles incoherencias entre las variables. No se encontró ninguno de estos. Para cada una de las variables se tiene un total de 145 datos, es decir, tampoco se hallaron datos faltantes.

Tabla 1: Resumen de las variables

Fumar	**Producto**	**Sexo**	**Edad**	**Censura**	**Tiempo**
Fuma: 65	Term: 33	Mujer: 80	Min. : 65.00	0	Min. : 3.00
No Fuma: 80	Whole Life: 40	Hombre: 65	1st Qu.: 71.00	1	1st Qu.: 12.00
	Universal Life: 32		Median : 79.00		Median : 32.00
	Other: 40		Mean : 77.57		Mean : 38.11
			3rd Qu.: 83.00		3rd Qu.: 54.00
			Max. : 90.00		Max. :123.00

De la Tabla 1 podemos observar la distribución de los asegurados correspondiente para cada variable. Del total de 145 registros, el 41 % está censurado. La mayoría de la población no fuma, tenemos que 65 asegurados fuman y 80 no tienen este hábito. En cuanto a la elección del producto, los más populares son Whole Life y los pertenecientes a la categoría Otros. Sin embargo la elección es más o menos uniforme, hay aproximadamente un cuarto de la población en cada producto. La mayoría de la población es femenina. La edad mínima de contratación de la póliza es 65 años, la máxima 90 años y en promedio el asegurado tiene 77.57 años al contratar. En cuanto al tiempo, se tiene que el mínimo en que se reclamó la póliza a partir de que fue contratada son de 3 meses, el máximo fue 123 meses, más de 10 años, y en promedio, se tomaron 38.11 meses antes de ser reclamada la póliza.

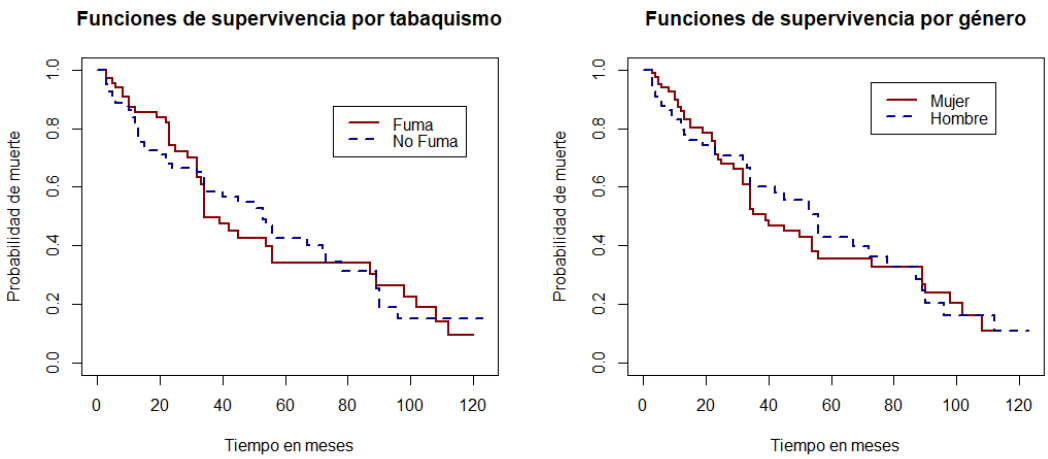
2.1. Curvas de supervivencia

La figura 1a parece sugerir que no hay diferencia entre la curva de supervivencia de los fumadores y no fumadores, y por lo tanto, no existe diferencia entre el tiempo en que un no fumador reclama su póliza al de un no fumador.

Mientras que analizando la figura 2b se puede observar que sí parece haber diferencias entre las curvas de supervivencia de las personas que contrataron Whole Life y Otros con respecto a los que contrataron Universal Life, sobretodo en meses avanzados.

Por otro lado, La figura 1b parece sugerir que no hay diferencia entre la curva de supervivencia de hombres y mujeres, y por lo tanto, no existe diferencia entre el tiempo en que un hombre reclama su póliza con respecto al de las mujeres.

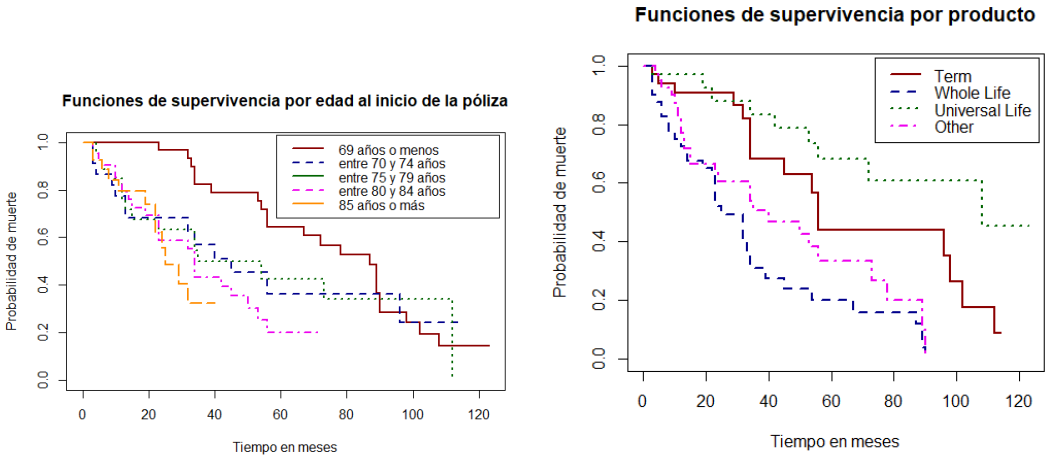
Por último, en la figura 2a se puede observar que sí parece haber diferencias entre las curvas de supervivencia de las personas que al contratar su póliza tenían entre 80 y 84 años con respecto a las personas de 69 años o menos. Mismo es el caso para el grupo de personas de 85 años o más con respecto a las personas de 69 años o menos. Lo cuál sugiere que existe diferencia entre el tiempo de vida para diferentes rangos de edad al momento de contratar la póliza.



(a) Estimación de las curvas de supervivencia de la variable tiempo en meses por estatus de fumador.

(b) Estimación de las curvas de supervivencia de la variable tiempo en meses por género.

Figura 1: En las gráficas se puede observar las curvas de supervivencia de la variable tiempo por factores género y estatus de tabaquismo.



(a) Estimación de las curvas de supervivencia de la variable tiempo en meses por Edad.

(b) Estimación de las curvas de supervivencia de la variable tiempo en meses por tipo de producto.

Figura 2: En las gráficas se puede observar las curvas de supervivencia de la variable tiempo por factores edad y tipo de producto contratado.

A continuación se plantean las hipótesis a juzgar para corroborar lo visto en las gráficas de curvas de supervivencia. Para todas Se considerará un nivel de significancia del 5 %.

2.2. Clasificación por tabaquismo

Como se ha mencionado anteriormente, se quiere comprobar si puede existir algún efecto por el tabaquismo del asegurado en el tiempo, en meses, en que fallece a partir de que contrató la póliza por lo que las hipótesis son:

H_0: La curva de supervivencia para asegurados que fuman es igual a la de los que no fuman para todo tiempo $t \geq 0$
vs
H_1: La curva de supervivencia para los asegurados que fuman difiere a la de los que no fuman para alguna $t \geq 0$

En notación matemática se escribe como

$H_0 : S(t)_H = S(t)_M, \quad \forall\, t, t \geq 0$
vs
$H_1 : S(t_0)_H \neq S(t_0)_M, \quad$ para algún $t_0 \geq 0$.

Al realizarse la prueba se obtuvo un valor p=0.9 por lo que con un nivel de confianza del 95 % no se rechaza la hipótesis nula, es decir, no hay evidencia suficiente para decir que la curva de supervivencia para asegurados que fuman es estadísticamente diferente a la de los asegurados que no fuman.

2.3. Clasificación por producto

Realizando la prueba de hipótesis para la clasificación por producto contratado, se obtiene que las hipótesis a plantar son:

H_0:Las curvas de supervivencia de las personas agrupadas por el producto elegido, son iguales para todo tiempo $t \geq 0$
vs
H_1:Las curvas de supervivencia de las personas agrupadas por el producto elegido no son iguales para algún tiempo $t \geq 0$.

En notación matemática se escribe como
$H_0 : S(t)_i = S(t)_j, \quad \forall\, i, j, t$ con $i \neq j, t \geq 0$
vs
$H_1 : S(t_0)_i \neq S(t_0)_j,$
para algún par $i, j,$ y tiempo $t_0 \geq 0$ con $i \neq j$ donde, i, j toman valores en $\{1, 2, 3, 4\}$.

Al realizar la prueba de hipótesis planteada, se obtuvo que con un nivel de confianza del 95 % y un valor p=0.000005 se rechaza la hipótesis nula, es decir, existe evidencia suficiente para decir al menos una curva de supervivencia para el tipo de producto contratado es estadísticamente diferente con respecto a otra.
Ya que se rechazó la hipótesis nula, se prosigió a realizar comparaciones múltiples por la prueba de Peto & Peto con el p-valor ajustado por el método de Bonferroni para determinar cuáles son los niveles de la variable `Producto` que podrían combinarse para crear subpoblaciones. Se obtuvo que en todas no se podía rechazar la hipótesis nula a excepción de las siguientes

- Con un valor p=0.0022 se rechaza la hipótesis nula, es decir, existe evidencia suficiente para decir que la curva de supervivencia para asegurados que contrataron Whole Life es estadísticamente diferente a la de los asegurados que contrataron Term.
- Con un valor p=000063 se rechaza la hipótesis nula, es decir, existe evidencia suficiente para decir que la curva de supervivencia para asegurados que contrataron Universal Life es estadísticamente diferente a la de los asegurados que contrataron Whole Life.
- Con un valor p=0.0063 se rechaza la hipótesis nula, es decir, existe evidencia suficiente para decir que la curva de supervivencia para asegurados que contrataron Universal Life es estadísticamente diferente a la de los asegurados que contrataron un seguro perteneciente a la categoría de Otros.

2.4. Clasificación por género

Realizando la prueba de hipótesis para la clasificación por género, se obtiene que las hipótesis a plantar son: H_0: La curva de supervivencia para los asegurados hombres es igual a la de las mujeres para todo tiempo $t \geq 0$
vs
H_1: La curva de supervivencia para los asegurados hombres difiere a la de las mujeres para alguna $t \geq 0$

En notación matemática se escribe como
$H_0 : S(t)_H = S(t)_M, \quad \forall\, t, t \geq 0$
vs
$H_1 : S(t_0)_H \neq S(t_0)_M, \quad$ para algún $t_0 \geq 0$. Al realizarse la prueba se obtuvo un valor p=0.9 por lo que con un nivel de confianza del 95 % no se rechaza la hipótesis nula, es decir, no hay evidencia suficiente para decir que la curva de supervivencia para asegurados hombres es estadísticamente diferente a la de los asegurados mujeres.

2.5. Clasificación por rango de edad

Realizando la prueba de hipótesis para la clasificación por rango de edad al momento de contratar la póliza, se obtiene que las hipótesis a plantar son:

H_0:Las curvas de supervivencia de las personas agrupadas por el rango de edad, son iguales para todo tiempo $t \geq 0$
vs
H_1:Las curvas de supervivencia de las personas agrupadas por el rango de edad no son iguales para algún tiempo $t \geq 0$.

En notación matemática se escribe como
$H_0 : S(t)_i = S(t)_j, \quad \forall\, i, j, t$ con $i \neq j, t \geq 0$
vs
$H_1 : S(t_0)_i \neq S(t_0)_j,$
para algún par $i, j,$ y tiempo $t_0 \geq 0$ con $i \neq j$ donde, i, j toman valores en $\{1, 2, 3, 4, 5\}$. Al realizar la prueba de hipótesis planteada, se obtuvo que con un nivel de confianza del 95 % y un valor p=0.008 se rechaza la hipótesis nula, es decir, existe evidencia suficiente para decir al menos una curva de supervivencia para el rango de edad es estadísticamente diferente con respecto a otra.
Ya que se rechazó la hipótesis nula, se prosigió a realizar comparaciones múltiples por la prueba de Peto & Peto con el p-valor ajustado por el método de Bonferroni para determinar cuáles son los niveles de la variable `Rango de Edad` que podrían combinarse para crear subpoblaciones. Se obtuvo que en todas no se podía rechazar la hipótesis nula a excepción de las siguientes:

- Con un valor p=0.00071 se rechaza la hipótesis nula, es decir, existe evidencia suficiente para decir que la curva de supervivencia para asegurados dentro del rango de edad entre 80 y 84 años es estadísticamente diferente a la de los asegurados que al contratar tenían 69 años o menos.
- Con un valor p=0.00022 se rechaza la hipótesis nula, es decir, existe evidencia suficiente para decir que la curva de supervivencia para asegurados que al contratar contaban con 85 años o más es estadísticamente diferente a la de los asegurados que al contratar tenían 69 años o menos.

3. Ajuste del modelo de Cox

De los factores mencionados en la sección anterior se procederá a crear un modelo con estos factores que busque estimar la función de supervivencia del tiempo de vida de una persona de edad mayor a 64 años. Por el numero de factores que se tiene y por el principio de economizar al recopilar datos, a continuación realizaremos una serie de iteraciones a la salida y entrada de factores a un modelo simple hasta un modelo más complejo con el fin de ver que combinación de factores aumenta sus significancias, y de esta manera tener el mejor modelo posible.

3.1. Hipótesis

Para la construcción del modelo se requiere saber cuáles variables y factores se agregarían a éste con el fin de detectar aquellas que aporten información relevante. Esto es equivalente a que, en un sentido estricto, el coeficiente acompañante de dicha variable sea diferente de cero. Es muy importante recalcar que, dada la muestra obtenida, algunas variables quizás no sean consideradas en el modelo (en el sentido de que haya evidencia que $\beta_i = 0$) aunque por la experiencia y literatura, sea demostrable que a pesar de que estadísticamente no debe considerarse dicha variable, sí se tenga que agregar por el efecto que ésta aporta pero que quizás no se refleja en la muestra utilizada.

Las hipótesis son:

$$H_0 : \beta_i = 0 \text{ vs } H_1 : \beta_i \neq 0,$$

que se puede interpretar como

H_0 : se debe considerar la i-ésima variable/factor en el modelo dado que las otras variables/factores están consideradas en éste,

H_1 : no se debe considerar la i-ésima variable/factor en el modelo dado que las otras variables/factores están consideradas en éste.

El estadístico de prueba que se utiliza es el siguiente:

$$\hat{z}_i = \frac{\hat{\beta}_i}{se(\hat{\beta}_i)}$$

el cual sigue una distribución asintótica normal estándar. Se rechaza la hipótesis nula a un nivel de significación $\alpha = 0.05$.

3.2. Construcción del modelo

Primero partimos de un modelo donde solo se incluyen las variables de forma individual.

Para la variable `Producto` se tienen 4 niveles de los cuales, `Term` se va a riesgo base y `Whole life` y `Other` resultan candidatos al modelo; por otro lado, `Universal life` no sería candidato para el modelo al nivel de significación dado.

Para las variable `Edad` resulta ser candidato para el modelo.

Luego los modelos tanto para `Fumador` (no fuma) como `Sexo` (hombre) no resultaron candidatos para los respectivos modelos.

Posteriormente hicimos modelos con dos variables: considerando todas las combinaciones de las variables `Edad`, `Sexo` y `Fumar`; solamente `Edad` resultó ser considerada en los modelos (donde se incluía) mientras que `Sexo` y `Fumar` no. Luego, se formuló un modelo considerando las tres anteriores pero de nuevo resultó apropiado considerar solamente `Edad`.

Lo siguiente fue considerar el modelo anterior agregando las interacciones entre `Fumar`, `Sexo` y `Edad`. También fueron desechados del modelo excepto `Edad`.

Luego, se consideraron modelos de dos variables considerando las interacciones, siendo `Fumar*Edad`, `Sexo*Edad` y `Fumar*Sexo`, sin embargo, de nuevo se desechan todas interacciones quedando solamente `Edad` como posible candidato al modelo.

El paso siguiente fue agregar las 4 variables (sin interacciones), lo cual resultó en que `Edad`, `Whole life` y `Other` resultaron posibles candidatos para el modelo mientras que `Fumar`, `Sexo` y `Universal life` no.

Como el factor `Producto` resultó un posible candidato, el siguiente modelo consideró `Edad`, `Term`, `Whole life` y `Universallife`. Se mandó a riesgo base `Other` en lugar de `Term` ya que ese tipo de seguro es más general en el sentido de que no se tiene una definición específica del producto (es cualquier otro tipo de seguro) por lo que se considera más apropiado que el ese seguro sea mandado al riesgo base por su "generalidad". Los resultados mostraron que todas las variables excepto `Whole life`[1] fueron posibles candidatos.

Luego al modelo anterior se le agregó la variable `Sexo` pero siguió sin ser un candidato al modelo. También se agregó la variable `Fumar` y ocurrió que tanto `Fumar` y `Whole life` deberían salir del modelo. Entonces, se procedió a quitar `Whole life` del modelo dándole una mejor aceptación a la variable `Fumar` en el sentido que su p-valor disminuyó de un 0.47 a 0.093 con respecto al modelo anterior (el que incluía a `Wholelife`). Esto es conveniente

[1] Esa variable tuvo un p-valor de 0.053531, por lo que puede ser discutible su inclusión en el modelo.

para el estudio ya que se tiene evidencia de que fumar tiene un efecto negativo en la calidad de vida en los ciudadanos de Estados Unidos[4].

Luego, del modelo anterior, se intercambió la variable `Fumar` por `Sexo` ya que también se sabe es importante el efecto de este factor (sino no habrían tablas de vida para hombres y mujeres). Sin embargo, al final no mostró evidencia de que aporte un efecto estadísticamente considerable en el modelo.

Posteriormente se analizaron las interacciones entre las variables del modelo previo con `Fumar`. Al final las interacciones no aportaron un efecto estadísticamente significativo en el modelo por lo que el modelo final a considerar fue con las variables `Edad`, `Fumar`, `Term` y `Universal life`.

3.3. Modelo de Cox

$$h_T(t;X) = h_0(t)\exp\{0.05191\text{Edad} - 1.2145\text{Term} \quad (1)$$
$$-1.80304\text{Universal life} - 0.40683\text{Fumar}\}$$

Tabla 2: En la tabla se puede observar las variables a considerar en el modelo final. Nótese que la variable Fumar se consideró en el modelo a pesar que su p-valor no fue menor al nivel de significación por su importancia que se tiene con base en la literatura.

	coef	**exp(coef)**	**se(coef)**	**z**	$Pr(>\|z\|)$
Edad	0.05191	1.05328	0.01642	3.162	0.001569
I(Producto == "Term")TRUE	-1.2145	0.29686	0.32784	-3.705	0.000212
I(Producto == "Universal Life")TRUE	-1.80304	0.1648	0.38649	-4.665	3.08E-06
FumarNo Fuma	-0.40683	0.66575	0.24215	-1.68	0.092936

Tabla 3: Resultados de la comprobación de riesgos proporcionales.

	ρ	χ^2	p
Edad	-0.0913	0.572	0.45
I(Producto == "Term")TRUE	-0.0418	0.144	0.705
I(Producto == Üniversal Life")TRUE	-0.1126	1.201	0.273
FumarNo Fuma	-0.0395	0.132	0.716
GLOBAL		1.815	0.77

4. Validación de modelo

Para la validación del modelo se requiere que se cumplan los supuestos de riesgos proporcionales y linealidad.

4.1. Supuesto de riesgos proporcionales

La prueba que se utilizó para corroborar este supuesto está basada en los residuos de Schoenfeld, en el que se definen a los coeficientes el modelo como una función del tiempo. Las hipótesis son las siguientes:

$$H_0 : \beta_i(t) = \beta_i \text{ vs } H_1 : \beta_i(t) \neq \beta_i,$$

que se puede interpretar como

H_0: Los coeficientes del modelo no varían con el tiempo,

H_1: Los coeficientes del modelo varían con el tiempo.

El estadístico de prueba puede consultarse en [5] y se sabe sigue una distribución asintóticamente χ^2_p.

De acuerdo con la Tabla 3, a un nivel de confianza del 95 % hay evidencia suficiente de que el modelo propuesto en (1) cumple el supuesto de riesgos proporcionales, es decir, los coeficientes no varían en el tiempo.

4.2. Linealidad

Para verificar la linealidad, se muestra el gráfico 3 con los residuos martingalas en la que se puede notar que de manera conjunta no muestra una tendencia lineal, lo que sugeriría que se necesita una transformación para cada variable.

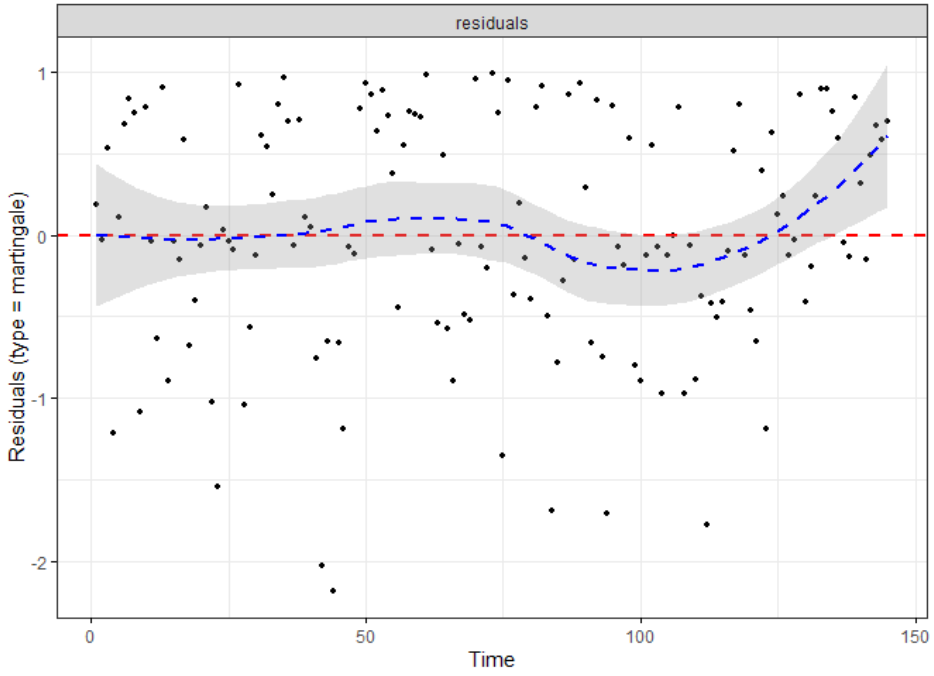


Figura 3: Gráfico de residuos martingala.

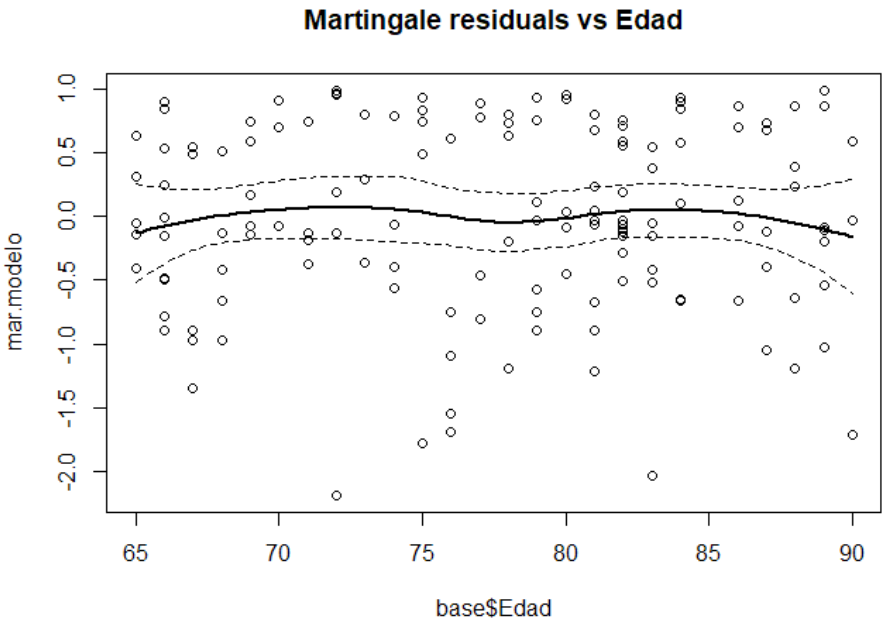


Figura 4: Gráfico de residuos martingala.

respecto a la linealidad de las variables, en este caso la única a la que se le puede realizar una validación es a la variable edad, de acuerdo con (4) se sugiere que no hay problema con el supuesto.

5. Interpretación

Ya que se cumple el supuesto de riesgos proporcionales y el de linealidad se garantiza que el modelo se ajusta adecuadamente a los datos trabajados.

Una forma equivalente del modelo (1) es la siguiente

$$\ln\left[\frac{h_T(t;X)}{h_0(t)}\right] = 0.05191\text{Edad} - 1.2145\text{Term} - 1.80304\text{Universal life} - 0.40683\text{Fumar} \tag{2}$$

el cuál se interpreta de la siguiente forma:

- 0.05191 es el logaritmo del riesgo relativo cuando la edad aumenta una unidad, manteniéndose constantes las demás variables, y por tanto, $e^{0.05191} = 1.53$ es el riesgo relativo cuando la edad aumenta una unidad, manteniéndose constantes las demás.
- -1.2145 es el logaritmo del riesgo relativo cuando el asegurado contrata un seguro de de vida Temp, manteniéndose constantes las demás variables, y por tanto, $e^{-1.2145} = 0.2968$ es el riesgo relativo cuando se cuenta con un riesgo Temp con respecto a los asegurados que tiene un seguro dentro de la categoría Other, manteniéndose constantes las demás. Por lo que las personas con un term son 70.32 % menos probable que tenga reclamaciones que las personas con un producto Other.
- -1.80304 es el logaritmo del riesgo relativo cuando el asegurado contrata un seguro de de vida Universal Life, manteniéndose constantes las demás variables, y por tanto, $e^{-1.80304} = 0.16479$ es el riesgo relativo cuando se cuenta con un riesgo Whole Life con respecto a los asegurados que tiene un seguro dentro de la categoría Other, manteniéndose constantes las demás.Por lo que las personas con un Universal life son 83.52 % menos probable que tenga reclamaciones que las personas con un producto Other.
- -0.40683 es el logaritmo del riesgo relativo cuando el asegurado no fuma manteniéndose constantes las demás variables, y por tanto, $e^{-0.40683} = 0.6657$ es el riesgo relativo cuando el asegurado fuma, manteniéndose constantes las demás.Por lo que las personas que no fuman no 0.3343 % menos probable que tenga reclamaciones que las personas que si fuman.

6. Referencias

[1] Kagan, J. (n.d.). Term Life Insurance. Retrieved December 8, 2019, from https://www.investopedia.com/terms/t/termlife.asp

[2] Kagan, J. (n.d.). Traditional Whole Life Policy. Retrieved December 8, 2019, from https://www.investopedia.com/terms/t/traditionalwholelifepolicy.asp

[3] Kagan, J. (2019). Universal Life Insurance. 08/12/2019, de investopedia.com Sitio web: https://www.investopedia.com/terms/u/universallife.asp

[4] Goldenberg, M., Danovitch, I., & IsHak, W. W. (2014). Quality of life and smoking. The American Journal on Addictions, 23(6), 540?562. https://doi.org/10.1111/j.1521-0391.2014.12148.x

[5] Grambsch, P. M., & Therneau, T. M. (1994). Proportional Hazards Tests and Diagnostics Based on Weighted Residuals. Biometrika, 81(3), 515?526. https://doi.org/10.2307/2337123

7. Anexos

Tabla 4: Valores de los p-valores, en orden según la formula ingresada en R

factor(Producto)	**Fumar + Sexo + Edad**	**Fumar * Sexo**
0.000417 ***	0.86415	0.488
0.054917 .	0.77112	0.315
0.03285 *	0.00123 **	0.271

Podemos ver como el modelo con los factores de producto(Term, Whole life, Universal life y Other) en la tabla (4) muestra mayor cantidad de variables con p-valor menor a 0.05, es decir, no hay evidencia significativa de que los coeficientes de estos factores sean cero.

Tabla 5: Valores de los p-valores, en orden según la formula ingresada en R

Fumar + Sexo	**Sexo + Edad**	**Fumar + Edad**
0.97	0.7847	0.88957
0.78	0.00125 **	0.00123 **

Podemos ver como los modelos con los factores edad en la tabla (5) muestra un p-valor menor a 0.05, es decir, no hay evidencia significativa de que los coeficientes de estos factores sean cero.

Tabla 6: Valores de los p-valores, en orden según la formula ingresada en R

Fumar	**Edad**	**Sexo**
0.996	0.00125 **	0.782

En la tabla (6) se muestra como en un modelo sencillo donde solo se considera un solo factor de una lista de 3, cuales resulta significativos, solo edad muestra un p-valor menor a 0.05.

Tabla 7: Valores de los p-valores, en orden según la formula ingresada en R

Sexo * Edad	**Fumar * Edad**	**Edad + Term + Whole.life + Universal.life + Sexo**
0.112155	0.8262	0.000988 ***
0.000597 ***	0.0178 *	0.022631 *
0.105021	0.836	0.047381 *
		0.000440 ***
		0.628868

Por otro lado tenemos que para modelos con interacciones, podemos usar en R la formula denotada por (factor*factor2) la cual es equivalente a (factor+factor2+factor:factor2), tendremos que ninguna interacción de la tabla (7) muestra tener un p-valor menor a 0.05, solo el modelo con Edad, Term, Whole life, Universal life y sexo tiene p-valores menores a 0.05, a excepción de la variable sexo, con un p-valor de 0.628868.

Tabla 8: Valores de los p-valores, en orden según la formula ingresada en R

Fumar* Sexo *Edad	**Fumar+Sexo+Edad+Whole.life + Universal.life+Other**	**Edad + Term + Whole.life + Universal.life**
0.8907	0.43642	0.000942 ***
0.2903	0.57132	0.024307 *
0.0169 *	0.00112 **	0.053531 .
0.2457	6.74e-05 ***	0.000423 ***
0.8243	0.1476	
	0.02019 *	

Y por último tenemos un análisis de los modelos más complejos, aquellos que tienen mayor numero de variables, en la tabla (8) podemos ver como la suma de todas las variables sin interacciones tiene una salida de p-valores donde la mitad de las variables resultan no significativas, mientras que con una $\alpha = 0.06$ se puede aceptar todas las variables del modelo 3 de dicha tabla (la numero 8) ya que el menor de ellos tiene un p-valor de 0.053.