Supervivencia de pacientes de bronquitis

Modelización Estadística II

Pablo Suárez Vieites y Alexander Olza Rodriguez

28 de diciembre de 2020

1. Descripción de los datos

Disponemos de una muestra de 483 pacientes con bronquitis: 463 hombres y 20 mujeres, con edades entre los 40 y los 83 años. El seguimiento se realizó durante 2045 unidades de tiempo, durante las cuales murieron 128 personas (proporción de censura 73 %). Los individuos censurados lo fueron en el instante 1825, y más tarde se reportaron 5 muertes.

La función de supervivencia estimada por el método Kaplan-Meier para la población general se muestra en la Figura 1.

A continuación veremos cómo cambia la estimación de Kaplan-Meier para distintos estratos de la población. Hemos clasificado los individuos por sexo, grupos de edad, nivel de disnea (ahogo o dificultad de respiración), si hacen o no ejercicio, su BMI (Índice de Masa Corporal, por sus siglas en inglés), su resultado en FEV (Volumen Espiratorio Forzado en un segundo en porcentaje) y en WDist (test de marcha, capacidad de esfuerzo en metros). Para las variables continuas (edad, BMI, FEV y WDist) se han hecho dos grupos, que contienen a los individuos que están por encima o por debajo de la mediana respectivamente.

Las distintas funciones de supervivencia se muestran en la figura 2. Se estiman tiempos de supervivencia inferiores para personas mayores de 70 años (Fila 1, columna 1), para personas con FEV$<$ 58 (4,1) y para personas con WDist$<$ 414 (3,2). En cuanto a la disnea, hay diferencias entre los distintos grupos pero no hay una tendencia clara de que mayor disnea implique menor tiempo de supervivencia (o viceversa). En el resto de las variables, las estimaciones de Kaplan-Meier no son significativamente diferentes.

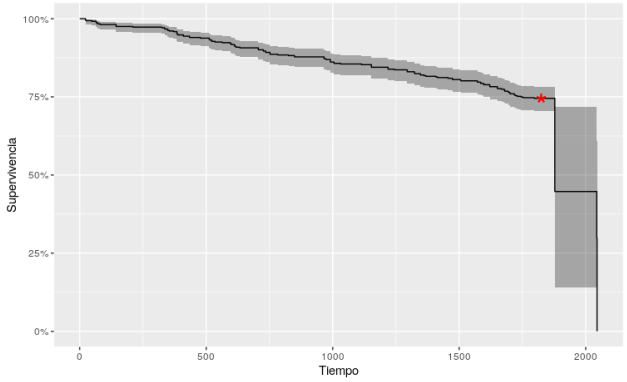

Figura 1: Función de supervivencia estimada para la población general

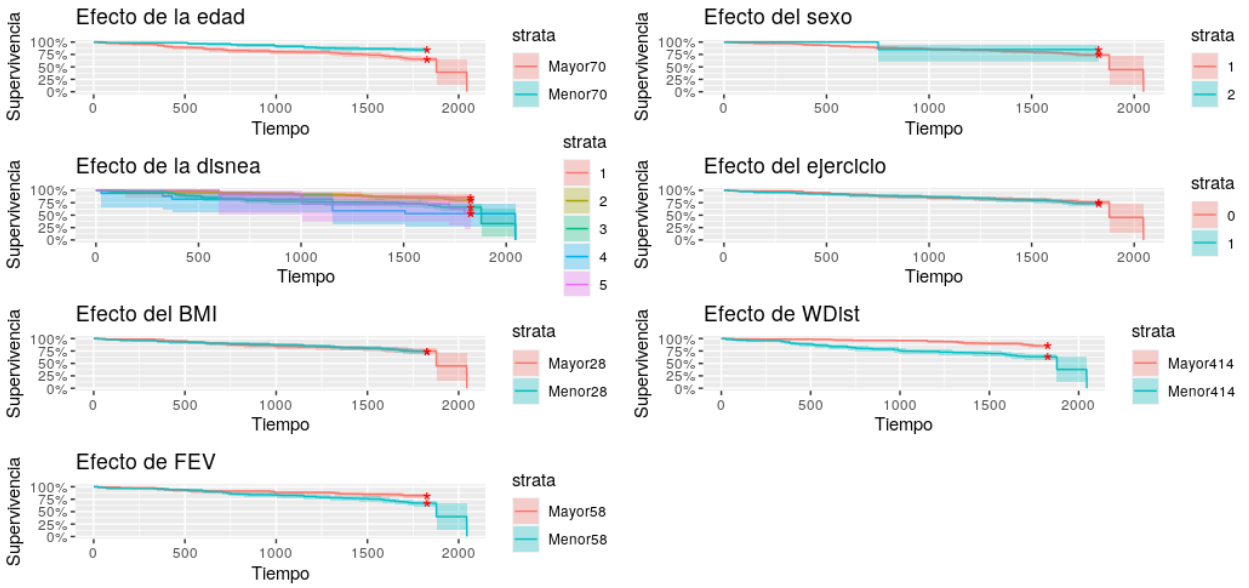

Figura 2: Función de supervivencia estimada para los distintos grupos

2. Modelos univariantes

A continuación se explora el efecto de cada variable aislada sobre el tiempo de supervivencia. Hablaremos sólo de las variables significativas que, en base al test de la razón de verosimilitud, son la edad, WDist, FEV y disnea. Cabe mencionar que esto es coherente con la figura 2. La significancia de las variables se presenta en la siguiente tabla.

Variable	Edad	Sexo	BMI	WDist	FEV	Ejercicio	Disnea
p-valor	5.7e-08	0.2594	0.5698	2.2e-11	1.5e-07	0.5231	0.0003

Presentamos a continuación una tabla con los parámetros más significativos de los análisis univariantes de variables continuas.

	β	e^{β}	$\sigma(\beta)$	p-valor(Wald)	Concordancia
Edad	0.072	1.074	0.014	4.84e-07	0.641
WDist	-6.33e-03	0.994	8.69e-04	3.39e-13	0.683
FEV	-0.037	0.964	0.007	1.27e-07	0.623

En base a los coeficientes de la tabla, podemos hacer las siguientes interpretaciones independientes (ya que se trata de modelos univariantes):

- **Edad:** Por cada año adicional de edad del paciente, el riesgo **aumenta un 7.4 %**.
- **WDist:** Por cada metro adicional en el test de capacidad de esfuerzo, el riesgo **disminuye un 0.6 %**.
- **FEV:** Por cada aumento de un 1 % en el volumen espiratorio forzado en un segundo, el riesgo **disminuye un 3.7 %**

En la siguiente tabla se resume el modelo univariante con la variable Disnea, que tiene 5 categorías. El tiempo de supervivencia no cambia significativamente para pacientes con Disnea 2 respecto a aquellos que tienen Disnea 1, pero a partir de Disnea 3 hay aumentos significativos en el riesgo de fallecer.

	β	e^{β}	$\sigma(\beta)$	p-valor(Wald)	Significativa
Disnea2	0.3520	1.4219	0.3825	0.35741	No
Disnea3	1.0165	2.7636	0.3782	0.00719	Sí
Disnea4	1.4429	4.2329	0.4913	0.00331	Sí
Disnea5	1.2551	3.5084	0.5702	0.02771	Sí

La concordancia del modelo univariante con Disnea es 0.61. Su interpretación es la siguiente:

- Los pacientes con **Disnea 3** tienen un riesgo **2.76 veces mayor** que aquellos con Disnea 1.
- Los pacientes con **Disnea 4** tienen un riesgo **4.23 veces mayor** que aquellos con Disnea 1.
- Los pacientes con **Disnea 5** tienen un riesgo **3.51 veces mayor** que aquellos con Disnea 1.

Teniendo en cuenta la definición de disnea, que el nivel 5 se considere un factor de riesgo más débil que el nivel 4 es contraintuitivo. Sin embargo, debe tenerse en cuenta que los intervalos de confianza para ambos niveles están muy solapados. Con un 95 % de probabilidad, el efecto de Disnea 4 en el riesgo está entre 1.62 y 11.09, mientras que el de Disnea 5 está entre 1.15 y 10.73. La amplitud de estos intervalos está relacionada con que sólo haya 17 pacientes con Disnea 4 y 11 con Disnea 5, mientras que las otras categorías están mucho más pobladas.

3. Modelo multivariante

3.1. Obtención del modelo y análisis

Para la selección del modelo multivariante utlizaremos el criterio de información de Akaike (AIC). El AIC nos calcula la bondad del ajuste con la función de verosimilitud parcial, penalizando el número de parámetros que usamos en el modelo. El modelo elegido será aquel que tenga menor AIC.
Para calcularlo utilizaremos la función step de R. Estableceremos como límite superior el modelo que contiene todas las covariables, y como límite inferior el modelo univariante con WDist, obteniendo que el mejor modelo es el que incluye las variables edad+WDist+FEV, con un $AIC = 1428.72$ Los valores de los coeficientes obtenidos son los que siguen:

	β	e^{β}	$\sigma(\beta)$	p-valor(Wald)
Edad	0.057	1.059	0.015	1.1e-04
WDist	-0.004	0.996	9.8e-04	1.7e-05
FEV	-0.027	0.973	0.007	2.2e-04

Vemos que todos los p-valores están por debajo de 0.05, por lo que todas las variables *son significativas dentro de nuestro modelo*. Los β pueden interpretarse de la siguiente manera:

- **Edad:** Por cada año adicional de edad del paciente, el riesgo **es un 5.9 % mayor, manteniendo el resto de covariables constantes**.
- **WDist:** Por cada metro adicional en el test de capacidad de esfuerzo, el riesgo **es un 0.4 % menor, manteniendo el resto de covariables constantes**.
- **FEV:** Por cada aumento de un 1 % en el volumen espiratorio forzado en un segundo, el riesgo **es un 2.8 % menor, manteniendo el resto de covariables constantes**

El valor de concordancia obtenido es $C = 0.711$ el cuál es a priori un buen valor. Veremos después si estamos en lo cierto aplicando un método de validación al modelo.

Además, para este modelo obtenemos según el test de razón de verosimilitud un p-valor con respecto al modelo constante de $p - valor(Verosimilitud, 1) = 9 \cdot 10^{16}$. Esto nos dice que nuestro modelo es significativo con respecto al modelo constante (que no incluye ninguna covariable). Por lo tanto la inclusión de las variables edad,WDist y FEV aportan información significativa el modelo.

Con este mismo razonamiento podemos ver si cada una de las covariables que tenemos en el modelo aporta información nueva. Esto se hará con un test de razón de verosimilitud, comparando el modelo reducido con el completo. Por ejemplo para la variable edad el modelo reducido será el que incluya las variables WDist y FEV. Este test se puede hacer con la función anova de R, obteniendo los siguientes resultados:

	$I(\beta_{reducido})$	p-valor(Verosimilitud)
Edad	-719.72	4.32e-05
WDist	-719.69	4.46e-05
FEV	-718.21	2.15e-04

Además $I(\beta_{completo}) = -711.36$. Se puede ver por los p-valores que todas las variables aportan información relevante al modelo. Cabe notar que estos p-valores no tienen el mismo significado que los obtenidos con el test de Wald para el modelo concreto. Una variable puede no aportar información al modelo y sin embargo ser significativa dentro de él.

3.2. Suposiciones de linealidad y proporcionalidad.

Una vez obtenido el modelo, es hora de comprobar si las suposiciones hechas acerca del mismo se cumplen, es decir, queremos comprobar si para se verifica para la hazard function de cada individuo que $h_i(t_k) = h_o(t_k)e^{\beta_i z_i}$.

- **Hipótesis de linealidad:** Para comprobar si la relación entre los coeficientes β_i y las covariables z_i sigue una dependencia lineal haremos tres modelos nuevos (uno para cada covariable), donde además de introducir la covariable con la dependencia lineal, introduciremos una dependencia no lineal utlizando una función suave (spline). Analizando los p-valores podremos ver si nuestra suposición de linealidad es correcta. Los valores obtenidos son los siguientes:

	p-valor(lineal)	p-valor(no lineal)
Edad	3.4e-05	0.051
WDist	3.6e-05	0.27
FEV	1.3e-04	0.73

  Vemos que en todos los casos el término no lineal no es significativo. Sin embargo en el caso de la edad está solo un 0.01 por encima del valor límite, por lo que debemos utilizar otro método para asegurarnos de que esta variable no entra de forma no lineal en nuestro modelo.
  Para ello ,teniendo el cuenta la expresión de los riesgos propocionales, podemos hacer una gráfica del log-hazard frente a a los valores de las covariables para ver si hay una tendencia lineal, obteniendo las gráficas que se ven en la Figura 3:

  Como vemos, en el caso de la edad, se ve una clara tendencia lineal, excepto para los valores más pequeños de la edad. Sin embargo, para estos valores el intervalo de confianza es muy ancho, por lo que no tenemos la suficiente exactitud como para rechazar la linealidad. Por lo tanto podemos dar por buena nuestra hipótesis de relación lineal.

- **Hipótesis de proporcionalidad** Para comprobar que la relación entre $h_i(t_k)$ y $h_o(t_k)$ es constante a lo largo del tiempo, debemos comprobar si el $e^{\beta_i z_i}$ es constante a lo largo del tiempo. Para ello podemos utilizar la función cox.zph de R, que hace un test de hipótesis donde la hipótesis nula es

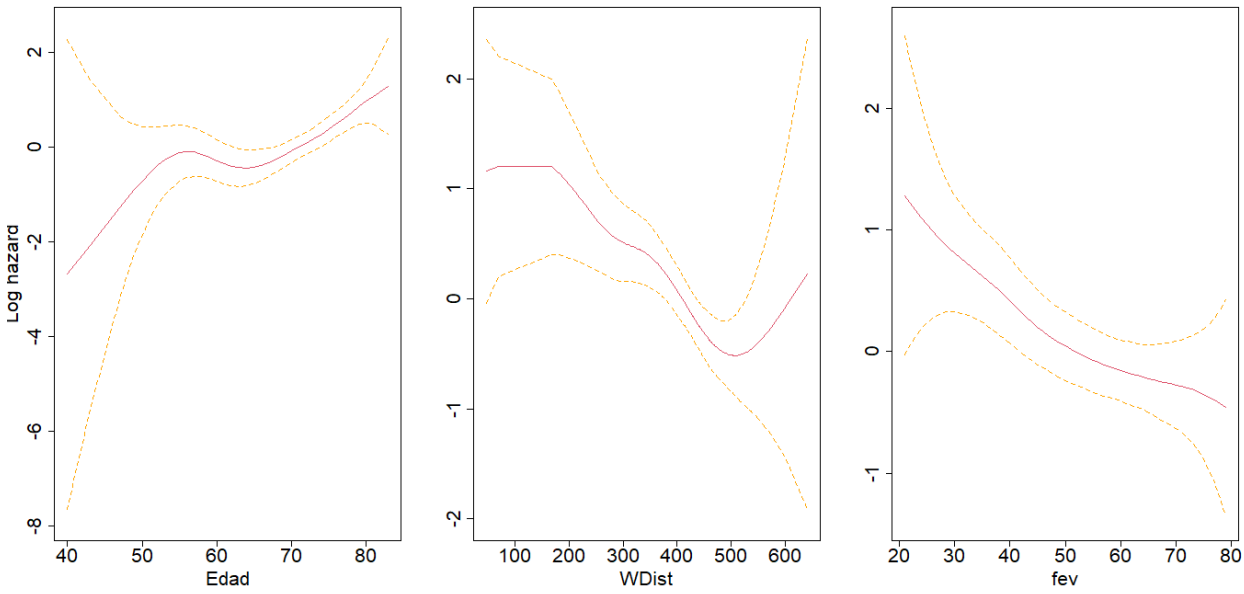

Figura 3: Variación del hazard con las covariables

que los diferentes β_i son constantes a lo largo del tiempo. Los valores obtenidos para cada una de las covariables son los siguientes:

	p-valor(proporcionalidad)
Edad	0.84
WDist	0.48
FEV	0.43
Global	0.51

Como vemos todos los valores son mayores que 0.05, por lo que no podemos rechazar la hipótesis nula de que las β son independientes del tiempo. Por lo tanto podemos concluír que los riesgos son proporcionales.

3.3. Validación del modelo

Finalmente, debemos comprobar la capacidad de predicción de nuestro modelo haciendo una validación del mismo. Al disponer de sólo una muestra, haremos una validación interna con los mismos datos.
Para ello utilizaremos un método Bootstrap, que hace B iteraciones (en nuestro caso 100) en las cuales de cada vez se toma una muestra aleatoria de nuestros datos (con repetición y el tamaño muestral original), y se ajusta el modelo con las variables edad WDist y FEV. Se calcula la concordancia del modelo (C_{boot}) y se le resta la concordancia obtenida cuando aplicamos los predictores de este modelo a los datos originales (C_o).Al promediar esta resta en las B iteraciones, obtendremos finalmente un valor llamado optimismo que debemos restar a la concordancia obtenida en el modelo original.
El valor de optimismo obtenido en nuestro caso es $Opt = 0.011$. Recordando que la concordancia obtenida originalmente fue de $C_{app} = 0.711$ obtenemos finalmente que la concordancia corregida vendrá dada por $C = 0.7$, lo que nos dice que la capacidad de predicción del modelo es buena.

4. Conclusiones

En este trabajo hemos modelizado la supervivencia de pacientes con bronquitis mediante modelos de riesgos proporcionales de Cox.

En primer lugar hemos planteado modelos univariantes con todas las variables disponibles, y hemos concluido que las variables significativas de forma aislada son Edad, WDist, FEV y Disnea.

A continuación se ha seleccionado el mejor modelo multivariante mediante el criterio del AIC, que ha dado como resultado Edad+WDist+FEV. Este modelo implica que tener mayor edad aumenta el riesgo de fallecer, y que mayor valor en las pruebas de WDist y FEV disminuyen el riesgo de fallecer. Por lo tanto, debe dispensarse mayor atención a los pacientes de edad avanzada y/o malos resultados en dichas pruebas.

El modelo se ha validado mediante el método Bootstrap de 100 muestras, obteniendo una concordancia corregida de 0.7. Esto lo convierte en el mejor modelo entre todos aquellos que se han probado en este trabajo.