UNIVERSIDAD NACIONAL DE ROSARIO

Facultad de Ciencias Económicas y Estadística

Análisis de datos de duración en pacientes con cáncer de mama

Rotterdam tumor bank - 1978-1985

Alumnas: Agustina Mac Kay y Rocio Canteros

Año 2024

Introducción

El cáncer de mama es un tipo de cáncer primario, que se origina en la mama y puede propagarse a otros tejidos u órganos del cuerpo. Es el tipo de cáncer más frecuente y la causa más común de muerte por cáncer en mujeres a nivel mundial.[1]

En este estudio se trabajará con información acerca de 583 mujeres que fueron sometidas, entre 1978 y 1985, a una cirugía primaria para extirpar el tumor.

Los datos fueron obtenidos de la base *rotterdam* del paquete *survival* de R. La misma cuenta con el tiempo desde la cirugía hasta la muerte o pérdida de seguimiento de las pacientes, junto a otras covariables basales que se detallan a continuación:

- **Age**: edad al momento de la cirugía (en años).
- **Meno**: estado menopáusico, donde 0 = premenopáusico y 1 = postmenopáusico.
- **Hormon**: variable indicadora de haber recibido un tratamiento hormonal.
- **Chemo**: variable indicadora de haber recibido quimioterapia.
- **Pgr**: receptores de progesterona (en fmol/l).
- **Er**: receptores de estrógeno (en fmol/l).
- **Grade**: grado de diferenciación del tumor, con valores 2 o 3.
- **Size**: tamaño del tumor, con niveles: menos de 20mm, entre 20 y 50mm, 50mm.

De la totalidad de mujeres en estudio, se cuenta con el tiempo exacto hasta la muerte de 377 de ellas y 206 censuras.

Elección del modelo

Para hallar el modelo de Cox más adecuado, se utilizará el método de selección de modelos propuesto por Collet y un nivel de significación de 0.05 en todos los pasos.

Selección de variables

Para tratar las variables *Tamaño* y *Grado* se definen las siguientes 3 variables dummies:

size	S_1	S_2	grade	G
< 20	0	0	2	0
20-50	1	0	3	1
> 50	0	1		

En primer lugar, se comparan modelos que incluyen una sola variable a la vez contra un modelo sin covariables.

Las hipótesis en contraste son:

$$\begin{cases} H_0) & \beta_j = 0 \\ H_1) & \beta_j \neq 0 \end{cases} \quad \text{con } j = \overline{1,8}$$

Con los resultados observados en la *Tabla 1* se determina que las variables significativas en esta etapa de la selección son: *Edad*, *Menopausia*, *Receptores de estrógeno*, *Grado de diferenciación* y *Tamaño del tumor*.

[1] Fuente: Organización Panamericana de la Salud

Variable	p-value	Decisión
Edad	~ 0	Rechazo H_0
Menopausia	~ 0	Rechazo H_0
Tratamiento hormonal	0.8381	No rechazo H_0
Quimioterapia	0.4217	No rechazo H_0
Receptores de progesterona	0.1316	No rechazo H_0
Receptores de estrógeno	0.0017	Rechazo H_0
Grado de diferenciación	0.0004	Rechazo H_0
Tamaño	~ 0	Rechazo H_0

Tabla 1: Test de hipótesis para la comparación de modelos univariados contra el modelo nulo

En segundo lugar, se evaluará si cada una de esas variables sigue siendo significativa en presencia de las demás seleccionadas.

Para cada variable, se compara entonces un modelo que contenga todas las variables significativas hasta el momento, excepto la variable en cuestión, contra un modelo que sí la incluya.

- Modelo con todas las variables significativas:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{meno}_i + \beta_3 \cdot \text{er}_i + \beta_4 \cdot G_i + \beta_5 \cdot S_{1i} + \beta_6 \cdot S_{2i})$$

- Modelos de comparación:

1) Sin la variable *Edad*:

$$h_i(t) = h_0(t) \cdot \exp(\beta_2 \cdot \text{meno}_i + \beta_3 \cdot \text{er}_i + \beta_4 \cdot G_i + \beta_5 \cdot S_{1i} + \beta_6 \cdot S_{2i})$$

2) Sin la variable *Menopausia*:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_3 \cdot \text{er}_i + \beta_4 \cdot G_i + \beta_5 \cdot S_{1i} + \beta_6 \cdot S_{2i})$$

3) Sin la variable *Receptores de estrógeno*:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{meno}_i + \beta_4 \cdot G_i + \beta_5 \cdot S_{1i} + \beta_6 \cdot S_{2i})$$

4) Sin la variable *Grado de diferenciación del tumor*:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{meno}_i + \beta_3 \cdot \text{er}_i + \beta_5 \cdot S_{1i} + \beta_6 \cdot S_{2i})$$

5) Sin la variable *Tamaño del tumor*:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{meno}_i + \beta_3 \cdot \text{er}_i + \beta_4 \cdot G_i)$$

Hipótesis: las hipótesis del test se dividen en 2 casos:

- Variables con 1 solo coeficiente asociado en el modelo:

$$\begin{cases} H_0) & \beta_j = 0 \\ & \\ H_1) & \beta_j \neq 0 \end{cases} \quad \text{con } j = \overline{1,4}$$

- Variable *Tamaño del tumor*, que cuenta con 2 coeficientes asociados en el modelo:

$$\begin{cases} H_0) & \beta_5 = \beta_6 = 0 \\ H_1) & \text{Al menos un} \quad \beta_j \neq 0, \quad \text{con } j = \overline{5,6} \end{cases}$$

Variable	p-value	Decisión
Edad	0.0229	Rechazo H_0
Menopausia	0.5586	No rechazo H_0
Receptores de estrógeno	0.1276	No rechazo H_0
Grado de diferenciación	0.0056	Rechazo H_0
Tamaño	~ 0	Rechazo H_0

Tabla 2: Test de hipótesis para probar la significancia de las variables en presencia de las demás

Teniendo en cuenta los resultados de las comparaciones en la *Tabla 2*, el modelo que se obtiene es:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot G_i + \beta_3 \cdot S_{1i} + \beta_4 \cdot S_{2i})$$

Sin embargo, este no es el modelo final. El siguiente paso es evaluar si ahora las variables que se descartaron en el primer paso son o no significativas.

Variable	p-value	Decisión
Tratamiento hormonal	0.5420	No rechazo H_0
Quimioterapia	0.0209	Rechazo H_0
Receptores de progesterona	0.3811	No rechazo H_0

Tabla 3: Significancia de las variables descartadas en el paso 1 en presencia de las seleccionadas en el paso 2

Por lo observado en la *Tabla 3*, se debe agragar *Quimioterapia* al modelo, quedando de la siguiente manera:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{chemo}_i + \beta_3 \cdot G_i + \beta_4 \cdot S_{1i} + \beta_5 \cdot S_{2i})$$

Por último, se prueban las interacciones dobles entre las variables pero ninguna resulta significativa, por lo que el modelo actual no se altera.

Linealidad

Cuando se tienen variables de tipo continua se debe analizar si incluirlas en el modelo de forma lineal o no lineal.

En este caso, la única variable continua que se tiene es *Edad*, por lo que se plantean las siguientes hipótesis:

$$\begin{cases} H_0) & \text{El efecto de Edad es lineal} \\ H_1) & \text{El efecto de Edad NO es lineal} \end{cases}$$

Como la probabilidad asociada al test es mayor a 0.05, es correcto incluir la edad al modelo de forma lineal.

Con este resultado, se concluye que el modelo seleccionado es el presentado con anterioridad:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{chemo}_i + \beta_3 \cdot G_i + \beta_4 \cdot S_{1i} + \beta_5 \cdot S_{2i})$$

Comprobación de supuestos

El modelo propuesto solo es válido si se cumple el supuesto de hazards proporcionales. Para comprobarlo, se utilizan los residuos de Schoenfeld obtenidos para cada variable y el test de Grambsch y Therneau.

Variable	p-value
Edad	~ 0
Quimioterapia	0.0438
Grado	0.2969
Tamaño	0.1809
Global	0.0004

Tabla 4: Test de hipótesis para la comprobación del supuesto de hazards proporcioneales

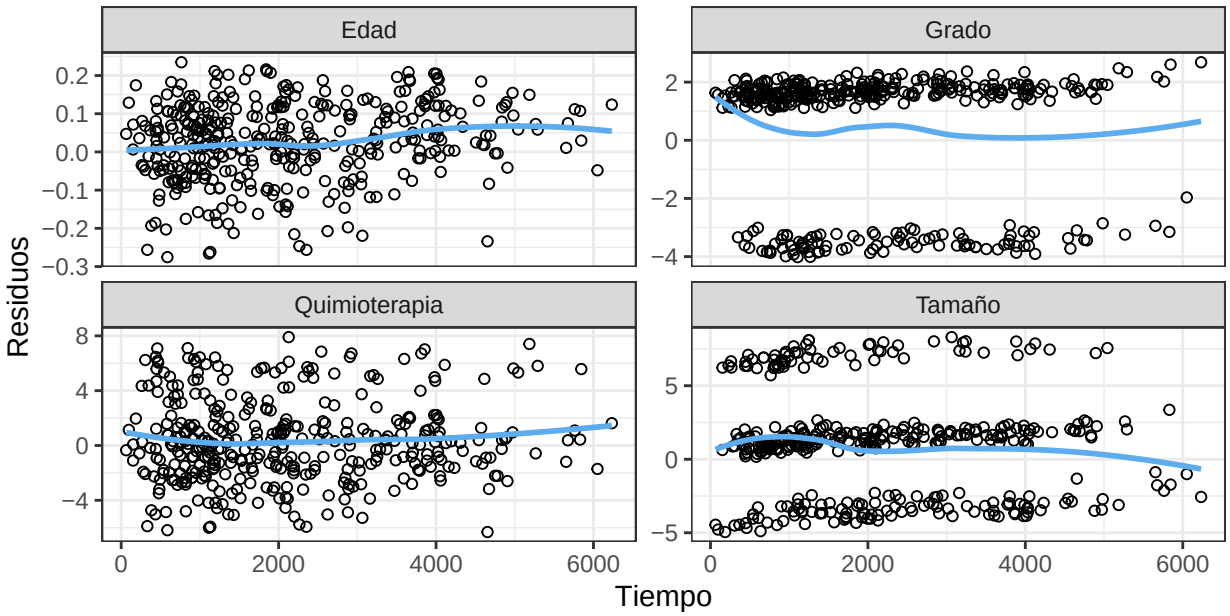

Figura 1: Comprobación gráfica del supuesto de hazards proporcinales

Los resultados de los test (*Tabla 4*), de manera global, indican que el supuesto no se cumple ya que se rechaza la hipótesis nula. En particular, las variables que no cumplen con la proporcionalidad de los hazards son *Edad* y *Quimioterapia*. Esto puede constatarse también de forma gráfica (*Figura 1*) al ver que los residuos de *Quimioterapia* no están divididos en dos grupos (cantidad de categorías que tiene), lo que sí sucede con *Grado* y *Tamaño*.

De la variable *Edad*, por ser continua, se debería observar un patrón aleatorio, pero se observa que los puntos están menos dispersos al avanzar el tiempo.

Para poder incluir las 2 variables que son significativas pero no cumplen el supuesto de proporcionalidad, se decide lo siguiente:

- Categorizar la variable *Edad* en 4 categorías: Menores de 40 años, Entre 40 y 50 años, Entre 50 y 65 años y Mayores de 65 años.
- Estratificar el modelo por la variable *Quimioterapia*.

Variable	p-value
A_1	0.540
A_2	0.500
A_3	0.063
Grado	0.201
Tamaño	0.123
Global	0.087

Variable	p-value
A_1	0.208
A_2	0.592
A_3	0.014
Grado	0.259
Tamaño	0.117
Quimioterapia	0.065
Global	0.046

Tabla 5: Test de comprobación del supuesto de hazards proporcionales estratificando por Quimioterapia (tabla derecha) y sin estratificar (tabla izquierda)

En la *Tabla 5* se puede observar que el supuesto se cumple solo para el modelo estratificado. Sin embargo, la probabilidad asociada al test del modelo sin estratificar es muy cercano a 0.05, y dicho modelo brinda la posibilidad de estimar un parámetro para *Quimioterapia*, que es una variable importante para el problema.

Como otra forma de analizar el supuesto de hazards proporcionales, se utiliza un modelo sin estratificar y con la edad categorizada dependiente del tiempo. Al comparar dicho modelo con un modelo cuyas covariables no dependen del tiempo, se obtiene un p-*value* de 0.03, lo cual indica nuevamente que el supuesto de hazards proporcionales no se cumple en un modelo con la edad categorizada.

Sin embargo, se considera que la *Edad* de las pacientes es una variable sumamente importante a la hora de analizar la experiencia de supervivencia al cáncer de mama. Por ello, no se la descartará a pesar de no cumplir el supuesto.

Advertencia

Como la *Edad* categorizada no cumple con el supuesto HP, sus razones de hazard serán una suerte de promedio a través del tiempo

Por lo tanto, se decide igualmente utilizar el modelo sin estratificar. De esta forma, el modelo final contiene las siguientes variables:

- **Edad** (categorizada).
- **Quimioterapia**.
- **Grado de diferenciación tumoral**.

- **Tamaño del tumor**.

Por lo que el modelo estimado resulta ser:

$$\hat{h_i}(t) = \hat{h_0}(t) \cdot exp(-0.05 \cdot A_{1i} + 0.5 \cdot A_{2i} + 0.77 \cdot A_{3i} + 0.35 \cdot \text{chemo}_i + 0.31 \cdot G_i + 0.38 \cdot S_{1i} + 0.84 \cdot S_{2i})$$

Interpretación del modelo

Coeficientes

Con solo mirar los coeficientes del modelo se puede concluir:

- El perfil de paciente femenino con mejor prónostico al realizar la cirugía primaria de extracción del tumor mamario es: mujer entre 40 y 50 años de edad, que no recibió quimioterapia y que posee un tumor de grado 2 con menos de 20cm de diámetro.
- El perfil de paciente femenino con el peor pronóstico es: mujer mayor a 65 años de edad, que recibió quimioterapia y posee un tumor de grado 3 con más de 50cm de diámetro.

Razones de hazards

Otra forma de interpretar el modelo es obteniendo razones de hazards, y sus respectivos intervalos de confianza, para distintos valores de las variables.

1) Grado y Tamaño: Con la interpretación de los coeficientes del modelo se pudo concluir que el tumor más grave es el de grado 3 y tamaño mayor a 50mm, y el menos grave es el de grado 2 y tamaño menor o igual a 20mm. Para cuantificar la diferencia entre ambos tumores se calcula la razón de hazards para un grupo etario y tratamiento de quimioterapia fijos.

2) Tamaño y Quimioterapia: La quimioterapia podría no ser tan beneficiosa en mujeres con tumores chicos, o no ser eficiente en mujeres con un cáncer muy avanzado. Por eso resulta de interés analizar la relación entre *Tamaño* y *Quimioterapia*.

3) Edad: Puede interesar, por ejemplo, comparar las funciones hazard de mujeres para dos grupos etarios: las menores de 40 años y aquellas entre 50 y 65 años.

RH	Estimación puntual	Intervalo
$G = 3;\ T \leq 20$ vs $G = 2,\ T \geq 50$	3.155	$(2.159; 4.610)$
$Q =$ no, $T > 50$ vs $Q =$ si, $T \leq 20$	1.63	$(1.034; 2.578)$
$Q =$ si, $T > 50$ vs $Q =$ no, $T \leq 20$	3.284	$(2.135; 5.051)$
Mujeres entre 50 y 65 años vs menores de 40	1.640	$(1.086; 2.476)$

1) La tasa de mortalidad de pacientes con grado de diferenciación tumoral 3 y un tumor de más de 50mm de diámetro es, como mínimo un 116% y como máximo un 361% mayor que esa misma tasa para pacientes con grado 2 de diferenciación y un tumor de 20mm de diámetro o menos, para un grupo etario y *Quimioterapia* fijos.

2) La tasa de mortalidad para mujeres que no recibieron quimioterapia teniendo un tumor de más de 50 mm es, como mínimo, un 3% mayor y, a lo sumo, un 158% mayor que esa misma tasa para mujeres que sí recibieron quimioterapia pero tienen un tumor de 20 mm o menos; para mujeres del mismo rango etario y grado de tumor.

Por otro lado, la tasa de mortalidad para mujeres que tienen un tumor de más de 50 mm y recibieron quimioterapia es, al menos, un 113% mayor y, como máximo, 405% mayor que esa misma tasa para mujeres cuyo tumor mide 20 mm o menos y no recibieron quimioterapia; para mujeres del mismo rango etario y grado de tumor.

3) La tasa de mortalidad de mujeres entre 50 y 65 años es, al menos, un 8.6% mayor y, como mucho, un 148% mayor que esa misma tasa para mujeres de menos de 40 años, manteniendo fijas las demás variables.