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 de 1 a 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.

Selección del modelo

En primer lugar se compara, para cada variable, su modelo univariado contra un modelo sin covariables. Previo a esto, se definen 2 *dummies* referentes a la variable *Tamaño* y otras 2 para la variable *Grado*:

size	S_1	S_2	grade	G_1	G_2
< 20	0	0	1	0	0
20-50	1	0	2	1	0
> 50	0	1	3	0	1

Las hipótesis en contraste son:

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

Los resultados obtenidos son los siguientes:

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

[1]Fuente: Organización Panamericana de la Salud

Se determina entonces que las variables significativas en esta etapa de la selección de variables son: Edad, Menopausia, Receptores de estrógeno, Grado de diferenciación y Tamaño del tumor.

En segundo lugar, se evaluará si cada una de esas variables siguen siendo significativas en presencia de las demás.

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

- Modelo aditivo:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{meno}_i + \beta_3 \cdot \text{er}_i + \beta_4 \cdot G_{1i} + \beta_5 \cdot G_{2i} + \beta_6 \cdot S_{1i} + \beta_7 \cdot S_{2i})$$

- Modelos de comparación:

1) Sin la variable *Edad*:

$$h_i(t) = h_0(t) \cdot \exp(\beta_2 \cdot \text{meno}_i + \beta_3 \cdot \text{er}_i + \beta_4 \cdot G_{1i} + \beta_5 \cdot G_{2i} + \beta_6 \cdot S_{1i} + \beta_7 \cdot S_{2i})$$

2) Sin la variable *Menopausia*:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_3 \cdot \text{er}_i + \beta_4 \cdot G_{1i} + \beta_5 \cdot G_{2i} + \beta_6 \cdot S_{1i} + \beta_7 \cdot S_{2i})$$

3) Sin la variable *Receptores de estrógeno*:

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{meno}_i + \beta_4 \cdot G_{1i} + \beta_5 \cdot G_{2i} + \beta_6 \cdot S_{1i} + \beta_7 \cdot S_{2i})$$

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

$$h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{meno}_i + \beta_3 \cdot \text{er}_i + \beta_6 \cdot S_{1i} + \beta_7 \cdot S_{2i})$$

5) Sin la variable *Tamaño del tumor*:

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

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

- Caso 1): la $j-$ésima variable tiene un solo β_j asociado a ella

$$\begin{cases} H_0) & \beta_j = 0 \\ \\ H_1) & \beta_j \neq 0 \end{cases} \qquad \text{con } j = \overline{1,3}$$

- Caso 2): la $j-$ésima variable tiene más de un β_j asociado a ella

$$\begin{cases} H_0) \quad \beta_j = \beta_{j'} = 0 \\ \\ H_1) \quad \text{Al menos un} \quad \beta_j \neq 0 \end{cases} \quad \text{con } (j, j') = (4,\ 5) \text{ o } (6,\ 7)$$

Teniendo en cuenta que los resultados de las comparaciones de los modeloes fueron los siguientes:

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

El modelo que se obtendría sería $h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot G_{1i} + \beta_3 \cdot G_{2i} + \beta_4 \cdot S_{1i} + \beta_5 \cdot S_{2i})$. Sin embargo, este no es el modelo final dado que falta evaluar si ahora, con otras variables en el modelo, aquellas que no resultaron significativas de manera univariada lo son.

Nuevamente se analizan los resultados brindados por la comparación entre los modelos:

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

De esta manera, el modelo sería $h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{age}_i + \beta_2 \cdot \text{chemo}_i + \beta_3 \cdot G_{1i} + \beta_4 \cdot G_{2i} + \beta_5 \cdot S_{1i} + \beta_6 \cdot S_{2i})$

Una vez definido, se prueban las distintas interacciones dobles entre las variables y resulta que ninguna es significativa. Por lo que no se altera el modelo.

Linealidad

Cuando se tienen variables de tipo continua debe analizarse en qué forma se las incluye en el modelo y hay dos opciones para ello: lineal y no lineal; esta última contempla tanto la idea de categorizar la variable como la de trabajar con una función de ella (logaritmo, al cuadrado, etc). Las hipótesis que se plantean son:

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

La probabilidad asociada es mayor al 5%, entonces la variable es lineal. Con este resultado, se concluye que el modelo definitivo es el presentado con anterioridad:

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

Comprobación de supuestos

Como es sabido, todo el trabajo realizado se desarrolló suponiendo que los hazards son proporcionales pero ¿Se cumple esto? Se comprabará mediante el uso de los residuos de Schoenfeld obtenidos para cada variable.

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

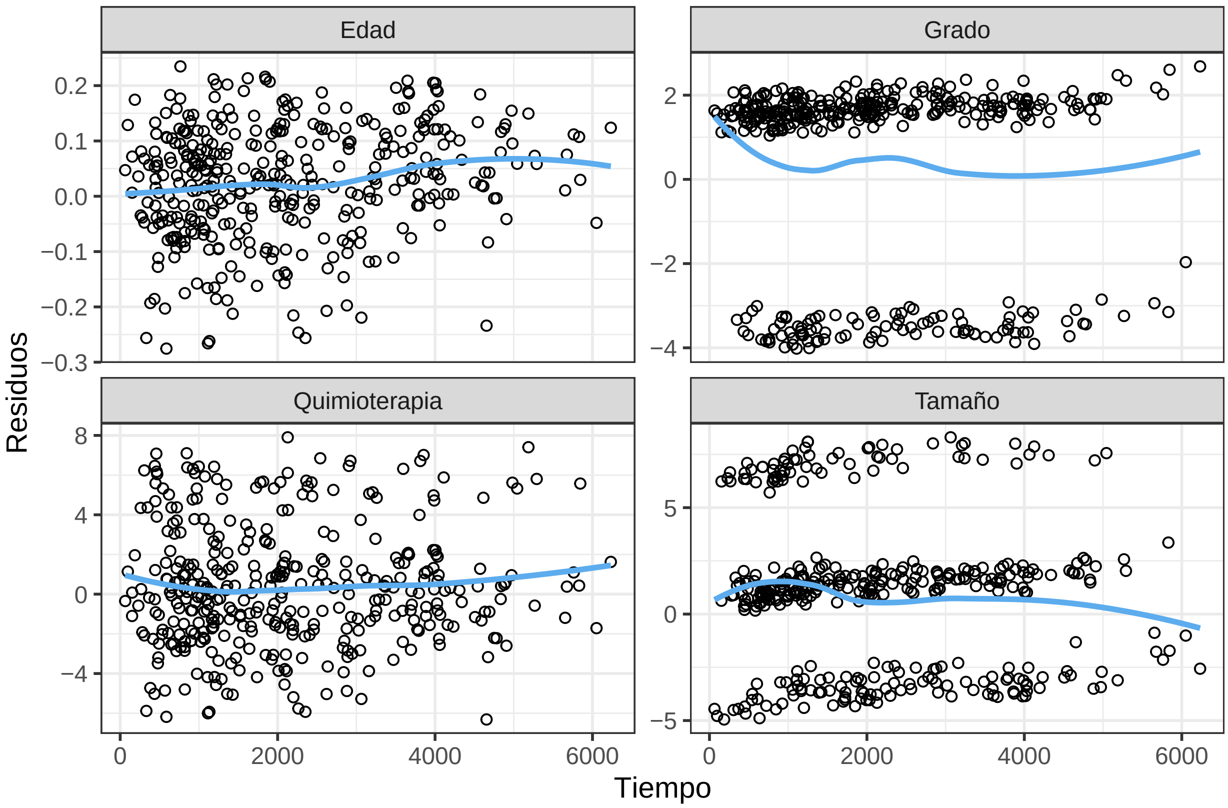

Gráfico 1: Evaluación de la proporcionalidad

La función `cox.zph()`, de manera global, dice que el supuesto no es cierto ya que se rechaza la hipótesis nula. Además, brinda información acerca de qué variables no estarían cumpliendo la proporcionalidad de los hazards: *Edad* y *Quimioterapia*. Esto puede constatarse también de forma gráfica 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 incluir las variables que son significativas pero no cumplen el supuesto de proporcionalidad, se decide lo siguiente:

- Categorizar la variable *Edad* con valores del 1 al 4.
- Plantear un modelo nuevo con la variable *Quimioterapia* estratificada.

Variable	p-value
Edad	0.07
Grado	0.22
Tamaño	0.15
Global	0.04

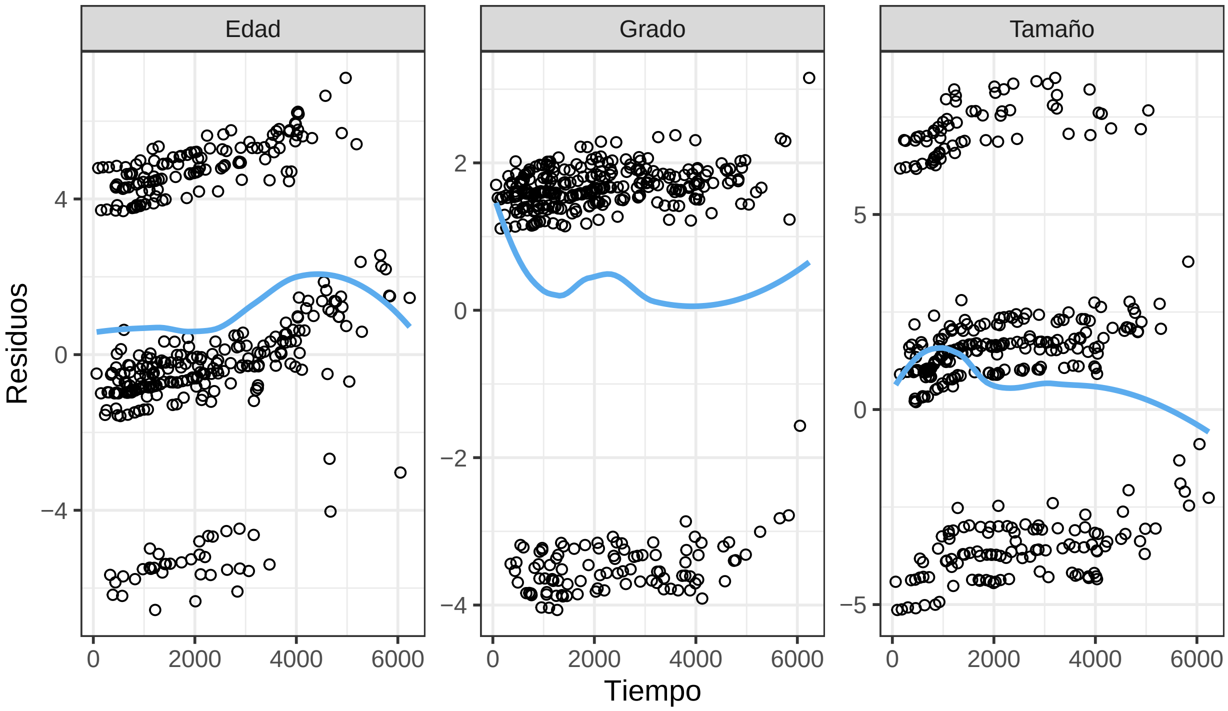

Gráfico 2: Evaluación de la proporcionalidad

Se puede observar que si bien el test global es significativo, las probabilidades asociadas a cada variable no lo son y gráficamente parece que el supuesto se cumple.

Se concluye entonces que el supuesto de hazards proporcionales se satisface para las siguientes variables significativas:

- Edad (categorizada)
- Quimioterapia (variable de estratificación del modelo)
- Grado de diferenciación.
- Tamaño del tumor.

Por lo que el modelo final estimado resulta ser

$$\hat{h_i}(t) = \hat{h_0}(t) \cdot exp(-0.05 \cdot A_{1i} + 0.47 \cdot A_{2i} + 1 \cdot A_{3i} + 0.31 \cdot G_{2i} + 0.38 \cdot S_{1i} + 0.81 \cdot S_{2i})$$

para $Quimioterapia = 0$ y

$$\hat{h_i}(t) = \hat{h_0}(t) \cdot exp(0.95 \cdot A_{1i} + 1.6 \cdot A_{2i} + 2.73 \cdot A_{3i} + 1.36 \cdot G_{2i} + 1.46 \cdot S_{1i} + 2.25 \cdot S_{2i})$$

para $Quimioterapia = 1$.