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

```
#Modelo sin covariables
modelo_nulo <- coxph(Surv(dtime, death) ~ 1, ties = "breslow", data = datos)
loglik_nulo <- modelo_nulo$loglik

#Modelos con una sola variable:
#1) Edad:
modelo_edad <- coxph(Surv(dtime, death) ~ age, ties = "breslow", data = datos)

loglik_edad <- modelo_edad$loglik[2]

#Comparación del modelo con el modelo nulo
lrtest(modelo_nulo, modelo_edad) #Edad es significativo
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "coxph.null", updated model is of class "coxph"

## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ 1
## Model 2: Surv(dtime, death) ~ age
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   0 -2194.4
## 2   1 -2178.1  1 32.726  1.061e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

[1] Fuente: Organización Panamericana de la Salud

```r
#2) Menopausia
modelo_meno <- coxph(Surv(dtime, death) ~ meno, ties = "breslow", data = datos)

loglik_meno <- modelo_meno$loglik[2]

#Comparo con el modelo nulo:
lrtest(modelo_nulo, modelo_meno) #Menopausia es significativo
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "coxph.null", updated model is of class "coxph"

## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ 1
## Model 2: Surv(dtime, death) ~ meno
##   #Df  LogLik Df Chisq Pr(>Chisq)
## 1   0 -2194.4
## 2   1 -2181.7  1 25.51    4.4e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#3) Terapia hormonal:
modelo_hormon <- coxph(Surv(dtime, death) ~ hormon, ties = "breslow", data = datos)

#Comparo con el modelo nulo:
lrtest(modelo_nulo, modelo_hormon) #No es significativo
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "coxph.null", updated model is of class "coxph"

## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ 1
## Model 2: Surv(dtime, death) ~ hormon
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   0 -2194.4
## 2   1 -2194.4  1 0.0417     0.8381
```

```r
#4) Quimioterapía:
modelo_chemo <- coxph(Surv(dtime, death) ~ chemo, ties = "breslow", data = datos)


#Comparo con el modelo nulo:
lrtest(modelo_nulo, modelo_chemo) #No es significativo
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "coxph.null", updated model is of class "coxph"

## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ 1
```

```
## Model 2: Surv(dtime, death) ~ chemo
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   0 -2194.4
## 2   1 -2194.1  1 0.6455     0.4217
```

```
#5) Receptores de progesterona:
modelo_pgr <- coxph(Surv(dtime, death) ~ pgr, ties = "breslow", data = datos)



#Comparo con el modelo nulo:
lrtest(modelo_nulo, modelo_pgr)  #No significativo
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "coxph.null", updated model is of class "coxph"
```

```
## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ 1
## Model 2: Surv(dtime, death) ~ pgr
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   0 -2194.4
## 2   1 -2193.3  1 2.2737     0.1316
```

```
#6) Receptores de estrógeno:
modelo_er <- coxph(Surv(dtime, death) ~ er, ties = "breslow", data = datos)


#Comparo con el modelo nulo:
lrtest(modelo_nulo, modelo_er)  #Receptores es significativo
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "coxph.null", updated model is of class "coxph"
```

```
## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ 1
## Model 2: Surv(dtime, death) ~ er
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   0 -2194.4
## 2   1 -2189.5  1 9.8979   0.001655 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#7) Grado de diferenciación:
modelo_grade <- coxph(Surv(dtime, death) ~ grade, ties = "breslow", data = datos)


#Compraro con el modelo nulo:
lrtest(modelo_nulo, modelo_grade)  #Es significativo
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "coxph.null", updated model is of class "coxph"

## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ 1
## Model 2: Surv(dtime, death) ~ grade
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   0 -2194.4
## 2   1 -2188.2  1 12.436  0.0004211 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#8) Tamaño del tumor:
modelo_size <- coxph(Surv(dtime, death) ~ size, ties = "breslow", data = datos)


#Comparación con el modelo nulo:
lrtest(modelo_nulo, modelo_size) #Es significativo
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "coxph.null", updated model is of class "coxph"

## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ 1
## Model 2: Surv(dtime, death) ~ size
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   0 -2194.4
## 2   2 -2175.8  2 37.335  7.812e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
datos_graf <- datos %>%
  mutate(pgr_cat = ifelse(pgr, 0, 1),
         er_cat = ifelse(er, 0, 1))
mod_alt <- coxph(Surv(dtime, death) ~ er_cat, ties = "breslow", data = datos_graf)

lrtest(modelo_nulo, mod_alt)
```

```
## Warning in modelUpdate(objects[[i - 1]], objects[[i]]): original model was of
## class "coxph.null", updated model is of class "coxph"

## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ 1
## Model 2: Surv(dtime, death) ~ er_cat
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   0 -2194.4
## 2   1 -2193.8  1 1.3392     0.2472
```

Tras comparar los modelos simples o univariados con aquel sin covariables, se determinó que las variables más "importantes" son: Edad, Menopausía, Receptores de progesterona, Grado de diferenciación y Tamaño del tumor. Sin embargo, no sé conoce si todas en conjunto son significativas o solo algunas, por esto se planteará un modelo aditivon con ellas y se probará, cada una por separado, si afectdn el ajuste del modelo aditivo.

Previo a esta comparación de modelos, 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

Modelo aditivo:

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

- Modelos de comparación:

1) Sin la variable *Tamaño*:

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

2) Sin la variable *Grado*:

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

3) Sin la variable *Receptores de estrógeno*:

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

4) Sin la variable *Menopausia*:

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

5) Sin la variable *Edad*:

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

Hipótesis: $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$ con $j = \overline{1:7}$

El subíndice de β tomará el valor correspondiente según cuál variable es de interés probar su significancia.

```
#Se define el modelo con las variables importantes:
modelo_cc <- coxph(Surv(dtime, death) ~ age + meno + er + grade + size, ties = "breslow", data = datos)


##Modelos combinando las variables:

#1)Edad, Menopausía, Receptores de estrógeno y Grado:
modelo_nsize <- coxph(Surv(dtime, death) ~ age + meno + er + grade, ties = "breslow", data = datos)

lrtest(modelo_nsize, modelo_cc) #La variable tamaño es significativa con las demás en modelo
```

```
## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ age + meno + er + grade
## Model 2: Surv(dtime, death) ~ age + meno + er + grade + size
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2171.9
## 2   6 -2157.2  2 29.401  4.128e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#2)Edad, Menopausía, Receptores de estrógeno y Tamaño:
modelo_ngrade <- coxph(Surv(dtime, death) ~ age + meno + er + size, ties = "breslow", data = datos)

lrtest(modelo_ngrade, modelo_cc) #EL grado es significativo con las demás variables en el modelo
```

```
## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ age + meno + er + size
## Model 2: Surv(dtime, death) ~ age + meno + er + grade + size
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -2161.0
## 2   6 -2157.2  1 7.6713   0.005611 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#3) Edad, Menopausía, Grado y Tamaño:
modelo_ner <- coxph(Surv(dtime, death) ~ age + meno + grade + size, ties = "breslow", data = datos)


lrtest(modelo_ner, modelo_cc) #La variable receptores no es significativa con las deás variables en el
```

```
## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ age + meno + grade + size
## Model 2: Surv(dtime, death) ~ age + meno + er + grade + size
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -2158.3
## 2   6 -2157.2  1 2.3216     0.1276
```

```
#4)Edad, Receptores de estrógeno, Grado y Tamaño:
modelo_nmeno <- coxph(Surv(dtime, death) ~ age + er + grade + size, ties = "breslow", data = datos)

lrtest(modelo_nmeno, modelo_cc) #Menopausía no es significativa con las demás variables en el modelo
```

```
## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ age + er + grade + size
## Model 2: Surv(dtime, death) ~ age + meno + er + grade + size
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -2157.4
## 2   6 -2157.2  1 0.3421     0.5586
```

```r
#5)Menopausía, Receptores de estrógeno, Grado y Tamaño:
modelo_nage <- coxph(Surv(dtime, death) ~ meno + er + grade + size, ties = "breslow", data = datos)

lrtest(modelo_nage, modelo_cc) #La edad es significativa con las demás variables en el modelo
```

```
## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ meno + er + grade + size
## Model 2: Surv(dtime, death) ~ age + meno + er + grade + size
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   5 -2159.8
## 2   6 -2157.2  1 5.1704    0.02297 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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

Modelo	$p - asociada$	*Decisión*
1	0	Rechazo H_0
2	0.006	Rechazo H_0
3	0.128	No rechazo H_0
4	0.558	No rechazo H_0
5	0.023	Rechazo H_0

El modelo que se obtendría sería $h_i(t) = h_0(t) \cdot \exp(\beta_1 \cdot \text{edad}_i + \beta_2 \cdot G_{1i} + \beta_3 \cdot G_{2i} + \beta_4 \cdot S_{1i} + \beta_5 \cdot S_{2i})$. 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 singular (NOTA PARA AGUS: sé que guille usó una palabra para ese tipo de ajustes pero no puedo recordar cuál es; lo hizo en el ejercicio de seleccion de variables cuando hay una de interés principal -tratamiento-).

```r
#Modelo con las variables significativas en conjunto
modelo_cc_2 <- coxph(Surv(dtime, death) ~ age + grade + size, ties = "breslow", data = datos)

##Modelos con las variables no importantes:

 #1)Se incluye "Terapía hormonal"
 modelo_chormon <- coxph(Surv(dtime, death) ~ age + grade + size + hormon, ties = "breslow", data = dat

 lrtest(modelo_cc_2, modelo_chormon) #La variable "Terapái hormonal" sigue siendo no significativa
```

```
## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ age + grade + size
## Model 2: Surv(dtime, death) ~ age + grade + size + hormon
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2158.6
## 2   5 -2158.4  1 0.3718      0.542
```

```r
 #2) Se inlcuye "Quimioterapía":
 modelo_cchemo <- coxph(Surv(dtime, death) ~ age + grade + size + chemo,
                        ties = "breslow", data = datos)

 lrtest(modelo_cc_2, modelo_cchemo) #Con las demás, "Quimioterapía" es significativa
```

```
## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ age + grade + size
## Model 2: Surv(dtime, death) ~ age + grade + size + chemo
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2158.6
## 2   5 -2155.9  1 5.3353     0.0209 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#3)Se incluye "Receptores de progesterona":
modelo_cpgr <- coxph(Surv(dtime, death) ~ age + grade + size + pgr,
                     ties = "breslow", data = datos)

lrtest(modelo_cc_2, modelo_cpgr) #No es significativa
```

```
## Likelihood ratio test
##
## Model 1: Surv(dtime, death) ~ age + grade + size
## Model 2: Surv(dtime, death) ~ age + grade + size + pgr
##   #Df  LogLik Df  Chisq Pr(>Chisq)
## 1   4 -2158.6
## 2   5 -2158.2  1 0.7672     0.3811
```

##Linealidad

Cuando se tienen variables de tipo continua debe analizarse en que forma se las incluye en modelo y hay dos opciones de inclusión (?): lineal y no lineal; esta última contempla tanto la idea de categorizar la variable como la de una función de ella (logaritmo, al cuadrado, etc).

```
datos1 <- datos %>%
  mutate(
    age_d1 = ifelse(age >= 40 & age < 50, 1, 0),
    age_d2 = ifelse(age >= 50 & age < 70, 1, 0),
    age_d3 = ifelse(age >= 70, 1, 0)
  ) %>%
  mutate(
    age_factor = case_when(
      age < 40 ~ 1,
      age >= 40 & age < 50 ~ 2,
      age >= 50 & age < 70 ~ 3,
      age >= 70 ~ 4
    )
  )

mod_edad_dummie <- coxph(Surv(dtime, death) ~ age_d1 + age_d2 + age_d3, ties = "breslow", data = datos1)

mod_edad_factor <- coxph(Surv(dtime, death) ~ age_factor, ties = "breslow", data = datos1)


lrtest(mod_edad_factor, mod_edad_dummie)
```

La probabilidad asociada en mayor al 5%, entonces la variable es lineal.

Ahora probamos si receptores de progesterona y receptores de estrógeno son lineales con la variable Edad en el modelo.

```
# Receptores de progesterona

quantile(datos$pgr, probs = c(0.25, 0.5, 0.75))

datos2 <- datos %>%
  mutate(
    pgr_d1 = ifelse(pgr >= 4 & pgr < 39, 1, 0),
    pgr_d2 = ifelse(pgr >= 39 & pgr < 187, 1, 0),
    pgr_d3 = ifelse(pgr >= 187, 1, 0)
  ) %>%
  mutate(
    pgr_factor =case_when(
      pgr < 4 ~ 1,
      pgr >= 4 & pgr <  39 ~ 2,
      pgr >= 39 & pgr <187 ~ 3,
      pgr >= 187 ~ 4
    )
  )


mod_eda_pgr <- coxph(Surv(dtime, death) ~ age + pgr_d1 + pgr_d2 + pgr_d3, ties = "breslow", data = datos


mod_edad_pgrf <- coxph(Surv(dtime, death) ~ age + pgr_factor, ties = "breslow", data = datos2)


lrtest(mod_edad_pgrf, mod_edad_pgrd) # Test significativo. No hay linealidad

# Probamos el efecto de los receptores de progesterona incluidos al modelo
# con varaibles dummies.

lrtest(modelo_edad, mod_edad_pgrd) # Test no significativo


# Receptores de estrógeno
quantile(datos$er, probs = c(0.25, 0.5, 0.75))

datos2 <- datos %>%
  mutate(
    er_d1 = ifelse(er >= 9 & er < 62, 1, 0),
    er_d2 = ifelse(er >= 62 & er < 188, 1, 0),
    er_d3 = ifelse(er >= 188, 1, 0)
  ) %>%
  mutate(
    er_factor =case_when(
      er < 9 ~ 1,
      er >= 9 & er < 62 ~ 2,
      er >= 62 & er <188 ~ 3,
      er >= 188 ~ 4
    )
  )
```

```r
mod_edad_erd <- coxph(Surv(dtime, death == 0) ~ age + er_d1 + er_d2 + er_d3, ties = "breslow", data = da

mod_edad_erf <- coxph(Surv(dtime, death == 0) ~ age + er_factor, ties = "breslow", data = datos2)


lrtest(mod_edad_erf, mod_edad_erd) # Test significativo. No hay linealidad

# Probamos el efecto de los receptores de estrógeno incluidos al modelo con # varaibles dummies.

lrtest(modelo_edad, mod_edad_erd) # Test no significativo
```