1 Modelando supervivencia con covariables

1.1. Introducción

Hay tres diferentes formas de modelar los datos de superviencia con covariables:

1. Familias paramétricas
2. Tiempo de vida acelerada
3. Riesgos Proporcionales (*Proportional Hazards*)
4. Nomios Proporcionales (*Proportional Odds*)

Familias paramétricas

Una forma de incorporar a algunas covaribles es considerar una de las distribuciones paramétricas de vida y considerar que su parámetros o parámetros dependa de estas covariables. Por ejemplo,

1. Si el tiempo de vida se modela con una distribución exponencial, su parámetro (λ) puede modelarse en relación a un vector de covariables x usando el modelo log lineal. Esto es

$$log(\lambda) = x'\beta$$

2. Si el tiempo de vida tiene distribución con más de una variable, una puede considerarse fija y la otra modelarse con una relación similar al caso anterior. Por ejemplo si el tiempo de vida tiene distribución Weibull(λ, β) se puede considerar β fija.

Si se tienen k subpoblaciones entonces a cada una se le puede modelar con distribuciones de vida diferentes, pero esto puede hacer innecesariamente más complejo el análisis y no se tendría una interpretación sencilla del impacto de las covariables.

Modelo de tiempo de vida acelerada

El modelo de tiempo de vida acelerado (TVA o *Accelerate failure time, AFT*). Sea $S_1(t)$ y $S_2(t)$ funciones de supervivencia de dos poblaciones. El TVA establece que hay una constante $c > 0$ tq

$$S_1(t) = S_2(ct) \qquad \forall t \geq 0.$$

Este modelo asume que la tasa de desgaste de la primera población es c veces como la de la segunda población. Por ejemplo, si $S_1(t)$ y $S_2(t)$ son las funciones de supervivencia de los seres humanos y perros, respectivamente,

entonces la "sabiduría popular" nos dice que un año-humano equivale a 7 años-perro, por lo que $c = 7$ y $S_1(t) = S_2(7t)$. Entonces la probabilidad de que un perro sobreviva 10 años calendarios (humano) es el mismo que el que un humano sobreviva 70 años.

Por otro lado, si T es continua y denotamos a μ_i como el tiempo de supervivencia media de la i-ésima población y ϕ_i el cuantil tal que $S_i(\phi_i) = \theta$ para algún valor de θ en (0,1), entonces

$$\begin{aligned} \mu_2 &= \int_0^\infty S_2(t)dt = c\int_0^\infty S_2(cu)du \qquad (t = cu) \\ &= c\int_0^\infty S_1(u)du = c\mu_1 \end{aligned}$$

Considere el modelo de regresión para el tiempo de supervivencia T

$$\ln(T) = \beta_0 + \boldsymbol{x}'\boldsymbol{\beta} + \sigma W \tag{1.1}$$

donde el término de error W tiene una distribución "adecuada". Por ejemplo las distribuciones de W "Valor extremo (VE)", "Valor extremo generalizado", Normal o Logistica están ligadas a las distribuciones Weibull, Gama generalizada, Log normal o log-logística de T, respectivamente.

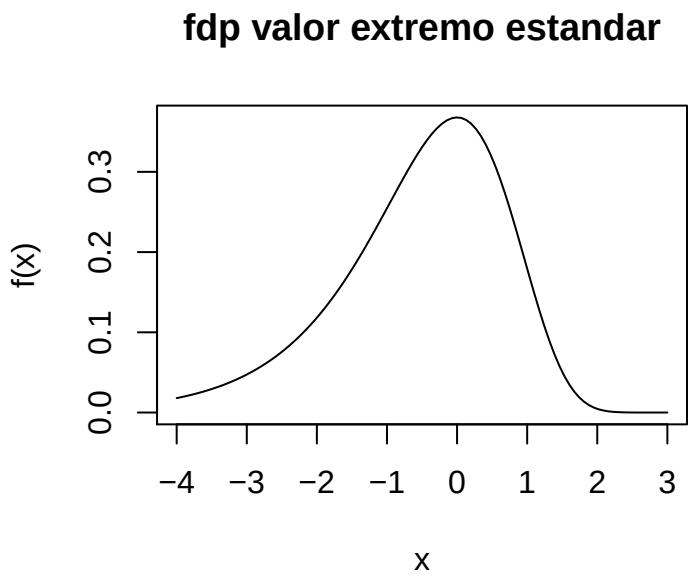
Ejemplo 1

Si $X \sim Exp(1)$ y $Y = \ln(X)$, la distribución de Y se conoce como "valor extremo estándar" y tiene funciones de riesgo y de densidad

$$\begin{aligned} f(x) &= \exp[y - \exp(y)] \\ h(t) &= \exp(t) \end{aligned}$$

Nótese que su soporte no son los reales positivos.

```
dsev<-function(x){exp(x-exp(x))}
curve(dsev,-4,3,ylab="f(x)",main="fdp valor extremo estandar")
```



Se puede demostrar que σW tiene distribución Weibull y entonces en escala logarítmica la distribución Weibull con tasa 1 corresponde a una familia con distribución VE.

Reconsideremos (1.1) expresado como

$$ln(T_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \sigma W_i,$$

donde W_i son iid. Para interpretar a los coeficientes en este modelo, consideremos β_k $(k = 1, \ldots, p)$ y fijemos a todos los otros coeficientes fijos. Si incrementamos el valor de la covariable x_k por una unidad de x_k a $x_k + 1$ y denotamos por T_1 y T_2 los correspondientes tiempos de supervivencia de las poblaciones con x_k y $x_k + 1$ (todas las demás covariables iguales), entonces T_1 y T_2 se pueden expresar como

$$\begin{aligned} T_1 &= e^{\beta_0 + b_1 x_{i1} + \cdots + \beta_k x_{ik} + \cdots \beta_p x_{ip}} e^{\sigma W_1} = c_1 e^{\sigma W_1} \\ T_2 &= e^{\beta_0 + b_1 x_{i1} + \cdots + \beta_k (x_{ik}+1) + \cdots \beta_p x_{ip}} e^{\sigma W_2} = c_1 e^{\beta_k} e^{\sigma W_2} \end{aligned}$$

Entonces

$$\begin{aligned} S_1(t) = P(e^{\sigma W_1} > c_1^{-1} t) \\ S_2(t) = P(e^{\sigma W_2} > (c_1 e^{\beta_k})^{-1} t) \end{aligned}$$

y como W_1 y W_2 son id entonces

$$S_2(e^{\beta_k} t) = S_1(t).$$

Entonces tenemos un tiempo de falla acelerado entre poblaciones 1 y 2, ya que si incrementamos el valor de x_k por una unidad de tiempo (mientras que las otras se mantienen constantes) el tiempo medio de supervivencia μ_2 y μ_1 cumplen con la relación $\mu_2 = \exp(\beta_k)\mu_1$. Cuando β_k es pequeño

$$\frac{\mu_2 - \mu_1}{\mu_1} = e^{\beta_k} - 1 \approx \beta_k \qquad \frac{\phi_2 - \phi_1}{\phi_1} = e^{\beta_k} - 1 \approx \beta_k$$

Ejemplo 2

Retomemos el Ejemplo 1. Uno de los modelos más sencillo es cuando T tienen distribución exponencial cuando $\mathbf{x} = \mathbf{0}$ (base de referencia -*baseline*-). En este caso se considera que la media (y función de riesgo) de la distribución tiene función de rieso $\exp(-\beta_0)$ en la base de referencia. Esto equivle a asumir que $\sigma = 1$ y W tienen distribución VE. Así e^W tiene distribución exponencial estándar con función de riesgo 1.

Entonces cuando se tienen cualquier vector de covariables $\mathbf{x}$, se tiene que T se distribuye exponencial con función de riesgo

$$h(t|\mathbf{x}) = e^{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}.$$

Entonces el efecto de incrementar el valor de una sola de las covariables x_k por una unidad se puede ver en la razón de las funciones de riesgo:

$$\frac{h(t|x_1, \ldots, x_k + 1, \ldots, x_p)}{h(t|x_1, \ldots, x_k, \ldots, x_p)} = e^{-\beta_k}$$

Ejemplo 3: El modelo Weibull

La única diferencia entre el modelo exponencial y Weibull es que en este segundo caso el parámetro σ puede ser diferente a 1 y se estima. Entonces la distribución de σW es es valor extremo con parámetro de escala σ. La función de supervivencia de T dado el valor de la covariable $\mathbf{x} = (x_0, \ldots, x_p)'$ se puede mostrar ser

$$S(t|\mathbf{x}) = \exp\left\{-\left(t\mathrm{e}^{\boldsymbol{x}'\boldsymbol{\beta}}\right)^{1/\sigma}\right\},$$

donde $\boldsymbol{\beta} = \beta_0, \ldots, \beta_p$.

Equivalentemente, en términos del logaritmo de la función de riesgo

$$\ln(h(t|\boldsymbol{x})) = \left(\frac{1}{\sigma} - 1\right)\ln(t) - \ln(\sigma) - \frac{\boldsymbol{x}'\boldsymbol{\beta}}{\sigma}. \tag{1.2}$$

Si denotamos como $\alpha = 1/\sigma$, $\beta_0^* = -\ln(\sigma) - \beta_0/\sigma$ y $\beta_j^* = -\beta_j/\sigma$ $(j = 1, \ldots, p)$ entonces

$$\ln(h(t|\boldsymbol{x})) = (\alpha - 1)\log(t) + \beta_0^* + x_i\beta_1^* + \cdots, x_p\beta_p^*$$

que es un modelo de riesgos proporcionales.

Modelo de riesgos proporcionales

El modelo de riesgos proporcionales se puede utilizar con tiempos de vida discretos o continuos y puede incorporar covariables que varían en el tiempo.

Si $h(t|\boldsymbol{x})$ es la función de riesgo para una persona en el tiempo t con covariables $\boldsymbol{x}$. La base de referencia en el tiempo t, $h_0(t)$ es cuando $\boldsymbol{x} = \mathbf{0}$. Esta base es análoga al intercepto y aunque no se especifica, debe ser positiva.

La razón entre las funciones de riesgo $h_1(t)/h_2(t)$ se puede interpretar como el riesgo relativo de un evento en el tiempo t.

El logaritmo de la razón de funciones de riesgo es una combinación lineal de los parametros y covariables

$$\ln\left(\frac{h(t|\boldsymbol{x})}{h_0(t)}\right) = \beta_1 x_1 + \cdots, \beta_p x_p$$

La razón de las funciones de riesgos se puede entonces considerar como una razón de funciones de exposición.

El modelo, en términos de la función de riesgo al tiempo t es

$$h(t|X_{1i}, X_{2i}, ..., X_{pi}) = h_0(t)\exp(\beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi})$$

A pesar de que un modelo distribución no se selecciona para representar la supervivencia, el modelo de riesgos proporcionales contiene un supuesto. Éste es que la función de riesgos de cualquier individuo es una proporción fija del riesgo de cualquier otro individuo. Si $h_0(t)$ es la función de riesgos de un sujeto (con $\boldsymbol{x} = \mathbf{0}$), entonces la razón de riesgos depende sólo de las covariables y no del tiempo t. Esto significa que las covariables se duplican de un momento a otro, también se duplica el riesgo del evento.

Ejemplo 4

El modelo de tiempo de vida acelerada Weibull puede realizarse en `R` usando la función `survreg` del paquete `survival`. El primer argumento corresponde a la `formula`. El argumento `dist` tiene varias opciones para

describir el modelo paramétrico a usar (`"weibull"`, `"exponential"`, `"gaussian"`, `"logistic"`, `"lognormal"`, `"loglogistic"`).

Los datos `larynx` en `KMsurv` es una base de datos con 90 casos de pacientes con cáncer de laringe reportados entre 1970 y 1978 en un hospital de holanda. Esta base tiene 5 columnas:

stage Etapa de la enfermedad (1=stage 1, 2=stage2, 3=stage 3, 4=stage 4)

time Tiempo de muerte o censura en el estudio (meses) desde la fecha del primer tratamiento

age Edad de fallecimeinto por cáncer de laringe

diagyr Año de diagnostico del cáncer

delta Indicador de muerte o censura (0=vivo, 1=muerte)

```
library(survival)
library(KMsurv)
data(larynx)
attach(larynx)
srFit <- survreg(Surv(time, delta) ~ as.factor(stage) + age, dist="weibull")
detach(larynx)
summary(srFit)

##
## Call:
## survreg(formula = Surv(time, delta) ~ as.factor(stage) + age,
##     dist = "weibull")
##                      Value Std. Error      z        p
## (Intercept)         3.5288     0.9041  3.903 9.50e-05
## as.factor(stage)2 -0.1477     0.4076 -0.362 7.17e-01
## as.factor(stage)3 -0.5866     0.3199 -1.833 6.68e-02
## as.factor(stage)4 -1.5441     0.3633 -4.251 2.13e-05
## age               -0.0175     0.0128 -1.367 1.72e-01
## Log(scale)        -0.1223     0.1225 -0.999 3.18e-01
##
## Scale= 0.885
##
## Weibull distribution
## Loglik(model)= -141.4   Loglik(intercept only)= -151.1
##  Chisq= 19.37 on 4 degrees of freedom, p= 0.00066
## Number of Newton-Raphson Iterations: 5
## n= 90
```

Note que la covariable stage se introduce como as.factor(stage) en la formula para especificar que la variable es categórica.

En la salida el `Intercept` y `Log(scale)` corresponden a los estimadores de β_0 y σ, respectivamente.

```
attach(larynx)
srFitExp <- survreg(Surv(time, delta) ~ as.factor(stage) + age, dist="exponential")
detach(larynx)
summary(srFitExp)

##
## Call:
## survreg(formula = Surv(time, delta) ~ as.factor(stage) + age,
##     dist = "exponential")
##                      Value Std. Error      z        p
## (Intercept)         3.7550     0.9902  3.792 1.49e-04
## as.factor(stage)2 -0.1456     0.4602 -0.316 7.52e-01
## as.factor(stage)3 -0.6483     0.3552 -1.825 6.80e-02
## as.factor(stage)4 -1.6350     0.3985 -4.103 4.08e-05
## age               -0.0197     0.0142 -1.388 1.65e-01
##
## Scale fixed at 1
##
## Exponential distribution
## Loglik(model)= -141.9   Loglik(intercept only)= -151.1
##  Chisq= 18.44 on 4 degrees of freedom, p= 0.001
## Number of Newton-Raphson Iterations: 4
## n= 90
```

Como ya vimos, cuando $\sigma = 1$, el modelo Weibull es equivalente al modelo exponencias. La selección entre uno y otro modelo puede hacerse a trav
es de una prueba de razón de verosimilutes con $h_0 : \sigma = 1$ contra la alternativa de dos lados, o examinando el nivel de significancia de `Log(scale)`. En el ejemplo, usando ambos análisis se obtiene que no hay evidencia suficiente que contradiga H_0.

Libros:

- David Collett (2003) *Modelling Survival Data in Medical Research* Chapman and Hall CRC. Secciones 3.1.1, 3.1.2, 3.2.1, 3.2.2, 3.2.3
- Jerald F. Lawless (2002)Statistical Models and Methods for Lifetime Data Wiley-Interscience. Capítulo 6.

1.2. Modelo de riesgos proporcionales de Cox

Los modelos semiparamétricos de función de riesgo multiplicativo establece

$$h(t|\boldsymbol{x}) = h_0(t)\phi(\beta, \boldsymbol{x})$$

donde $h(t|\boldsymbol{x})$ es la función de riesgo del individuo con covariables $\boldsymbol{x} = (x_1, \ldots, x_p)'$ y $h_0(t)$ es la función de riesgo de referencia.

En este tipo de modelación solo se asume sobre T que es una v.a. continua.

El modelo de riesgos proorcionales de Cox es

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{\beta}'\boldsymbol{x})$$

Si n es el número de individuos que se observan, de los cuales r tienen tiempos de muerte que se registrasn son $t_1, \ldots, t_r$ y $n-r$ son censurados a la derecha. Sea n_j el número de individuos en riesgo al tiempo $t_{(j)}$ y $R(t_{(j)})$ el grupo de (indices) individuos en riesgo al tiempo $t_{(j)}$.

En 1972 Cox propuso la función de verosimilitud parcial para estimar $\boldsymbol{\beta}$ como

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{r} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_{(j)})}{\sum_{l\in R(t_{(j)})}\exp(\boldsymbol{\beta}'\boldsymbol{x}_l)} \tag{1.3}$$

Considerando $\delta_i = 1$ si t_i es tiempo de falla (y 0 eoc) y

$$Y_i(t) = \begin{cases} 1 & \text{si } i \in R(t) \\ 0 & \text{eoc} \end{cases}$$

entonces (1.3) se puede escribir como

$$L(\boldsymbol{\beta}) = \prod_{j=1}^{n} \left(\frac{\exp(\boldsymbol{\beta}'\boldsymbol{x}_j)}{\sum_{l=1}^{n} Y_l(t_j)\exp(\boldsymbol{\beta}'\boldsymbol{x}_l)} \right)^{\delta_j}$$

La log verosimilitud es

$$l(\beta) = \sum_{j=1}^{n} \delta_j \left[\boldsymbol{\beta}'\boldsymbol{x}_j - \log\left(\sum_{l=1}^{n} Y_l(t_j)\exp(\boldsymbol{\beta}'\boldsymbol{x}_l) \right) \right]$$

El vector score y la matriz de información se pueden expresar de forma simple si se define

$$\bar{\boldsymbol{x}}(t,\beta) = \frac{\sum_{l=1}^{n} Y_l(t)\boldsymbol{x}_l \exp(\boldsymbol{\beta}'\boldsymbol{x}_l)}{\sum_{l=1}^{n} Y_l(t)\exp(\boldsymbol{\beta}'\boldsymbol{x}_l)}.$$

Se puede demostrar que el vector de score y la matriz de información son respectivamente

$$S(\boldsymbol{\beta}) = \sum_{j=1}^{n} \delta_j [\boldsymbol{x}_j - \bar{\boldsymbol{x}}(t_j,\beta)]$$

y

$$I(\boldsymbol{\beta}) = \sum_{j=1}^{n} \delta_j \left\{ \frac{\sum_{l=1}^{n} Y_l(t_j)\exp(\boldsymbol{\beta}'\boldsymbol{x}_l)[\boldsymbol{x}_l - \bar{\boldsymbol{x}}(t_j,\boldsymbol{\beta})][\boldsymbol{x}_l - \bar{\boldsymbol{x}}(t_j,\boldsymbol{\beta})]'}{\sum_{l=1}^{n} Y_l(t_j)\exp(\boldsymbol{\beta}'\boldsymbol{x}_l)} \right\}.$$

Observaciones

- No es necesario especificar la función de riesgo base $h_0(t)$.
- En la función de verosimilitud parcial (FVP), $L(\boldsymbol{\beta})$, no se introduce información sobre el tiempo entre eventos.
- En la FVP la información de los datos censurados solo se utiliza al establecer los grupos en riesgo.

Ejemplo 5 (Dirk F. Moore)

Consider a hypothetical comparative clinical trial with six subjects assigned to either a control or treatment group. The survival data for the control group are 6, 7+, and 15, and for the treatment group they are 10, 19+, and 25 (Table). In tabular form, with the survival times in increasing order, we have where "C" denotes a control patient and "T" denotes a treatment patient.

Paciente	t_i	δ_i	Grupo
1	6	1	C
2	7	0	C
3	10	1	T
4	15	1	C
5	19	0	T
6	25	1	T

```
library("survival")
tt <- c(6, 7, 10, 15, 19, 25)
delta <- c(1, 0, 1, 1, 0, 1)
trt <- c(0, 0, 1, 0, 1, 1)
result.cox <- coxph(Surv(tt, delta) ~ trt)
summary(result.cox)

## Call:
## coxph(formula = Surv(tt, delta) ~ trt)
##
##   n= 6, number of events= 4
##
##        coef exp(coef) se(coef)     z Pr(>|z|)
## trt -1.3261    0.2655   1.2509 -1.06    0.289
##
##     exp(coef) exp(-coef) lower .95 upper .95
## trt    0.2655      3.766   0.02287     3.082
##
## Concordance= 0.7  (se = 0.187 )
## Rsquare= 0.183   (max possible= 0.76 )
## Likelihood ratio test= 1.21  on 1 df,   p=0.2715
## Wald test            = 1.12  on 1 df,   p=0.2891
## Score (logrank) test = 1.27  on 1 df,   p=0.2591
```

Ya que el EMVP $\hat{\beta} = -1.326129$ y $\exp(\hat{\beta}) = 0.2655 < 1$ entonces el riesgo de las personas en tratamiento disminuye en relación a las personas en el grupo control. Para verificar este resultado realizamos una prueba de hipótesis, que se basa en la distribución asintótica del EMVP a la Normal.

A continuación se verifican algunos de los resultados obtenidos con `coxph` obteniendoles directamente.

```
library(numDeriv)
#The log partial likelihood function
plsimple <- function(beta) {
  psi <- exp(beta)
  result <- log(psi) - log(3*psi + 3) -
  log(3*psi + 1) - log(2*psi + 1)
  result }
```

We may find the m.p.l.e. (maximum partial likelihood estimate) using the `optim` function. The control parameter `fnscale` is set to -1 so that optim will find the maximum of the function `plsimple`. (The default would be to find the minimum.)

```
result <- optim(par=0, fn = plsimple, method = "L-BFGS-B", control=list(fnscale = -1),
lower = -3, upper = 1)
result$par

## [1] -1.326129
```

We may compute the derivative of the log-likelihood (i.e. the score) evaluated at $\beta = 0$ numerically using the `gradient` function in the package `numDeriv` (which must be separately downloaded and installed),

```
library(numDeriv)
grad(func=plsimple, x=0)

## [1] -0.9166667
```

The result -0.917 is thus the score evaluated at the null hypothesis. To carry out the score test, we also need the information, which we obtain using the `hessian` function as follows:

```
h<-hessian(func=plsimple, x=0)
h

##             [,1]
## [1,] -0.6597171
```

The **score test statistic**, expressed as Z_s^2, is the square of the score at $\beta = 0$ divided by minus the hessian (information), also at $\beta = 0$. This is the result given on last line of the summary output. We compare this to a chi-square distribution with one degree of freedom. The score test p-value is given by the upper tail.

```
Z2<-grad(func=plsimple, x=0)^2/-h
Z2

##          [,1]
## [1,] 1.273694

pchisq(Z2, df=1, lower.tail=F)

##           [,1]
## [1,] 0.2590748
```

To compute the **Wald test**, we need the maximum partial likelihood estimate $\hat{\beta}$ and the information at this point. We compute this as for the score test, but evaluated at $\hat{\beta}$. This is `result.cox$par`, and here is the hessian for the Wald test. The square root of minus the reciprocal of the hessian is the approximated standard error,

```
hp<-hessian(func=plsimple, x=result$par)
hp

##           [,1]
## [1,] -0.639117

sqrt(-1/hp)
```

```
##          [,1]
## [1,] 1.250863
```

The parameter estimate and standard error are given also in the summary. Finally, the Wald test statistic Z_w and two-sided p-value for the test are given by

```
result$par/1.2508

## [1] -1.060225

2*pnorm(1.060, lower.tail=F)

## [1] 0.2891446
```

These results may also be found in the summary of `result.cox`. The square of the test statistic is 1.124, and the same Wald p-value of 0.289.

The **likelihood ratio statistic** is twice the difference in the log partial likelihoods evaluated at $\hat{\beta}$ and at 0:

```
betahat <- result.cox$coefficients
2*(plsimple(betahat) - plsimple(0))

##      trt
## 1.209369
```

This result may be found on the "likelihoo ratio test" line, along with the p-value derived from the chi-square distribution,

```
pchisq(1.209, 1, lower.tail=F)

## [1] 0.2715303
```

Two additional portions of the output are often useful. The statistic "r-square" is an adaptation to survival analysis of the R^2 statistic from linear regression. Here it is defined as follows:

$$R^2 = 1 - \left(\frac{l(0)}{l(\hat{\beta})}\right)^{2/n}$$

and reflects the improvement in the fit of the model with the covariate compared to the null model. The "Concordance" is the C-statistic, a measure of the predictive discrimination of a covariate. See Harrell (2015)[1] for more details.

Cuando las covariables en el modelo son constantes entonces $\exp(\boldsymbol{\beta})$ es independiente del tiempo y la función de riesgo según diferentes covariables son proporcionales.El termino *riesgos proporcionales* se utiliza en forma general para describir el modelo de riesgo relativo de Cox, pero la propiedad de *proporcionalidad* se cumple para casos particulares como cuando las covariables son fijas.

[1] Harrell, F.E.: Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis, 2nd edn. Springer Science Business Media, New York (2015)

Tiempo de vida T continua pero registrada con empates

Cuando T se concibe como continua pero todavía se pueden tener algunos datos con tiempos de falla idénticos debido a la resolución de nuestros "instrumentos de medición".

A: Solución exacta (Kalbfleisch y Prentice)
B: Aproximación de Breslow
C: Aproximación de Efron

Ejemplo 6

```
library(KMsurv)
data(larynx)
head(larynx)
```

```
##   stage time age diagyr delta
## 1     1  0.6  77     76     1
## 2     1  1.3  53     71     1
## 3     1  2.4  45     71     1
## 4     1  2.5  57     78     0
## 5     1  3.2  58     74     1
## 6     1  3.2  51     77     0
```

```
## Vemos si hay empates en los tiempos de muerte
any(duplicated(larynx$time[larynx$delta==1]))
```

```
## [1] TRUE
```

```
## Vemos cuales son los valores de d_j que son duplicados
tab<-table(larynx$time[larynx$delta==1])
tab[tab>1]
```

```
##
## 0.3 0.8   1 1.3 1.8 1.9   2 3.2 3.5 3.6   4 6.4
##   3   3   2   2   2   2   2   2   3   2   3   2
```

A continuación usamos la aproximación de Breslow para tratar los empates.

```
fit.breslow <- coxph( Surv(time, delta) ~ age + factor(stage), data=larynx, method="breslow")
summary(fit.breslow)
```

```
## Call:
## coxph(formula = Surv(time, delta) ~ age + factor(stage), data = larynx,
##     method = "breslow")
##
##   n= 90, number of events= 50
##
##                    coef exp(coef) se(coef)     z Pr(>|z|)
## age            0.01890   1.01908  0.01425 1.326    0.185
```

```
## factor(stage)2 0.13856   1.14862  0.46231 0.300    0.764
## factor(stage)3 0.63835   1.89335  0.35608 1.793    0.073 .
## factor(stage)4 1.69306   5.43607  0.42221 4.010 6.07e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##                exp(coef) exp(-coef) lower .95 upper .95
## age                1.019     0.9813    0.9910     1.048
## factor(stage)2     1.149     0.8706    0.4642     2.842
## factor(stage)3     1.893     0.5282    0.9422     3.805
## factor(stage)4     5.436     0.1840    2.3763    12.436
##
## Concordance= 0.682  (se = 0.045 )
## Rsquare= 0.182   (max possible= 0.988 )
## Likelihood ratio test= 18.07  on 4 df,   p=0.001197
## Wald test            = 20.82  on 4 df,   p=0.0003443
## Score (logrank) test = 24.33  on 4 df,   p=6.867e-05
```

Como en los modelos lineales, el primer valor de la variable categorica `factor(stage)` se considera como de referencia.

La aproximación de Efron se puede implementar usando

```
fit.efron <- coxph( Surv(time, delta) ~ age + factor(stage), data=larynx )
```

ya que este método es el de default para `R` dada que esta aproximación es mejor cuando se tiene muchos empates.

Aunque el parámetro `method` en la función `coxph` admite la opción `"exact"`, el método que implementa no es el valor exacto presentado por Kalbfleish y Prentice, sino que utiliza un modelo de regresión logístico condicional (*matched logistic regression*, *conditional logistic regression*).[2]

Evaluación

Cuando las covariables son categóricas, una forma simple de examinar si el modelo evidentemente no viola los supuestos, es graficando las diferencias de las log-log transmormación de las $\hat{S}(t)$ (obtenidas con K-M) en cada categorá. Ya que $\log(H(t|\boldsymbol{x}^*) - \log(H(t|\boldsymbol{x}) = \boldsymbol{\beta}'(\boldsymbol{x}^* - \boldsymbol{x})$ por el supuesto de riesgos proporcionales, debemos ver que esta diferencia de las log-log curvas de supervivencia es constante a lo largo de los valores de t.

Se puede realizar un aálisis de los residuos para verificar el cumplimiento de los supuestos del modelo de Cox. `R` implementa en la función `coxph.zph` una prueba de tal ajuste, de acuerdo a Grambsch y Therneau (1994)[3].Dicha prueba se basa en un modelo de regresión para probar la no proporcionalidad y brinda una graficas auxiliares que presentan los residuos escalados de Schoenfeld[4], junto con una curva suavizada.

La base de datos `colon` del paquete `suvival` contiene las primeras pruebas exitosas de quimioterapia para cáncer de colon. *Levamisole* es un compuesto de baja toxicidad que se uso previamente para tratar los gusanos en animales, 5-FU es un componente de mediana toxicidad. Hay dos registros por personas, uno para la recurencia y otra que indica la muerte.

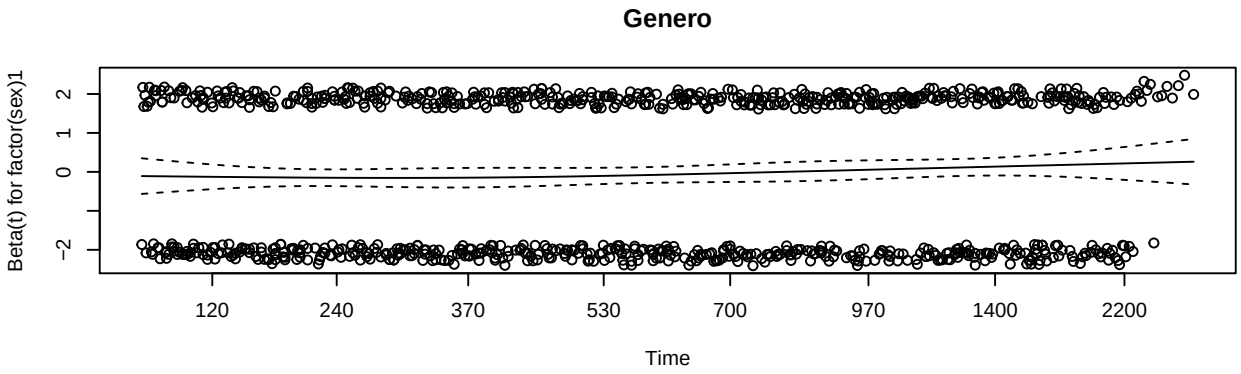
[2] Hosmer and Lemeshow (2000) *Applied Logistic Regression*, Kleinbaum and Klein (2010) *Logistic Regression: A Self-Learning Text.*

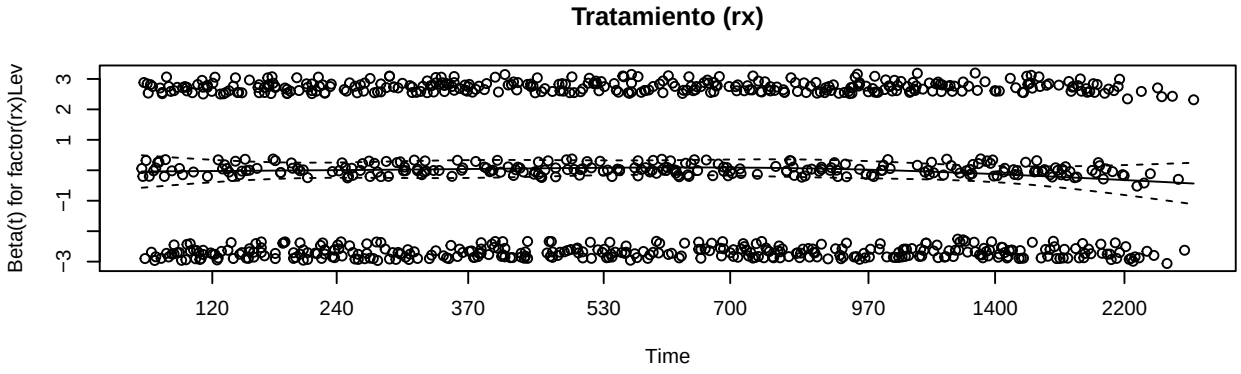
[3], P. Grambsch y T. Therneau (1994) *Proportional hazards tests and diagnostics based on weighted residuals.* Biometrika, 81, 515-26.

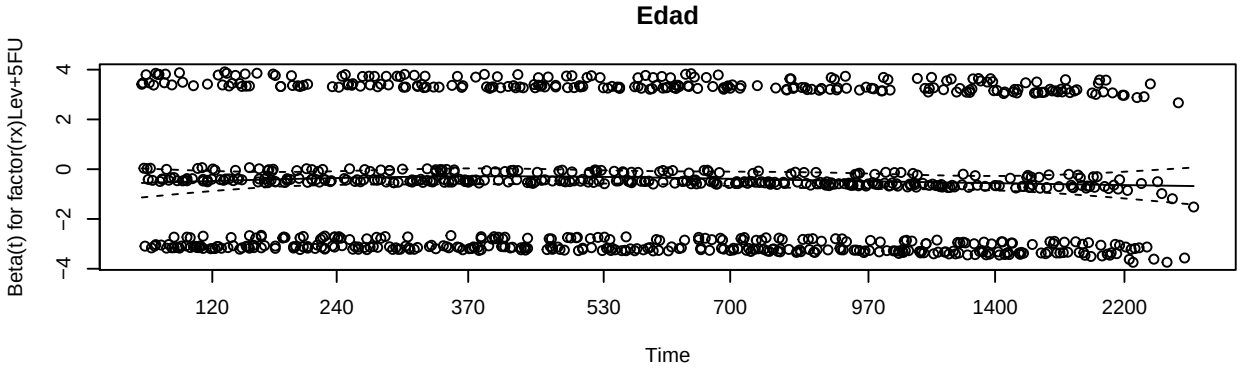
[4]Schoenfeld D. (1982)*Residuals for the proportional hazards regresssion model.* Biometrika. 69(1):239-241.

La función `cox.zph` "Test the Proportional Hazards Assumption of a Cox Regression".

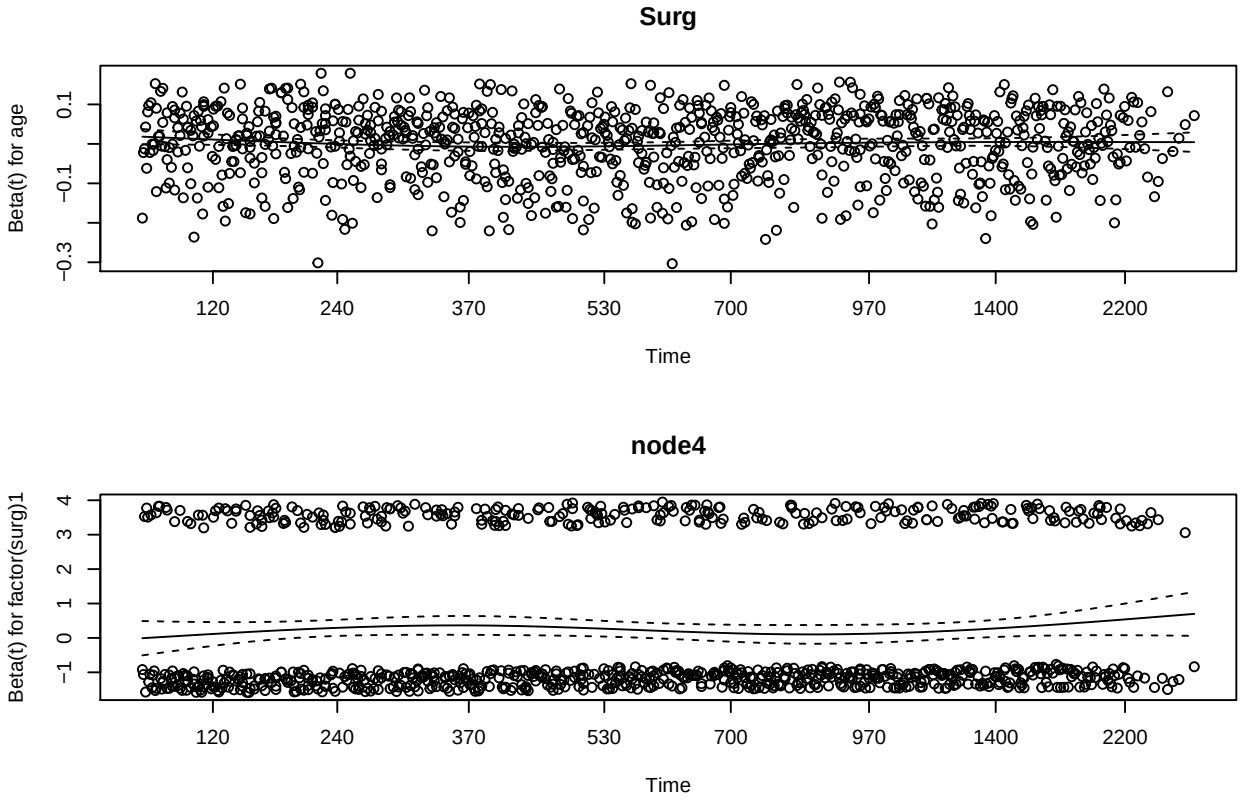
```
model <- coxph(Surv(time,status) ~ factor(sex) + factor(rx) + age +
              factor(surg) + factor(node4), data = colon)
a <- cox.zph(model)
a

##                        rho   chisq       p
## factor(sex)1       0.05602  2.8629 0.09064
## factor(rx)Lev     -0.02282  0.4809 0.48803
## factor(rx)Lev+5FU -0.02436  0.5515 0.45771
## age               -0.00219  0.0048 0.94477
## factor(surg)1      0.01689  0.2647 0.60694
## factor(node4)1    -0.10290  9.4187 0.00215
## GLOBAL                 NA 13.7842 0.03214

par(mfrow = c(3, 1))
plot(a[1], main = "Genero")
plot(a[2], main = "Tratamiento (rx)")
plot(a[3], main = "Edad")
```








```
plot(a[4], main = "Surg")
plot(a[5], main = "node4")
```



Covariables que varían en el tiempo

Libros:

- David Collett (2003) *Modelling Survival Data in Medical Research.* Chapman and Hall CRC. Sección 3.3
- Dirk F. Moore (2016) *Applied Survival Analysis Using R.* Springer International Publishing. Capítulo 5.
- Hosmer, Lemeshow y May (2008) *Applied Survival Analysis: Regression Modeling of Time to Event Data* Wiley. Capítulo 3.