Regresión paramétrica

Giancarlo Sal y Rosas

Departmento de Ciencias
Pontificia Universidad Católica del Perú

April 19, 2017

Outline

Modelo acelerado

- Sea T el tiempo a la ocurrencia de un evento y sea Z una variable que se cree afecta T.
- Para una observación, (t_i, z_i), consideremos el modelo

$$\log(t_i) = \beta_0 + \beta_1 z_i + \sigma\epsilon_i$$

donde

  - β_1 es el coeficiente de regresión que mide el efecto de Z sobre T
  - σ es un parámetro de escala
  - ϵ_i son los terminos de las perturbaciones (errores) que son independendientes e identicamente distribuidos

- ¿Qué mide β_1?

Modelo acelerado

- Consideremos dos poblaciones
  - Población A: Aquellos para el cual $Z = z$
  - Población B: Aquellos para el cual $Z = z + 1$
- Sea T_A y T_B los tiempo hasta la ocurrencia de un evento para las poblaciones A y B, respectivamente. Entonces

$$\begin{aligned} T_A &= e^{\beta_0+\beta_1 z} e^{\sigma\epsilon_A} = e^{\beta_0} e^{\sigma\epsilon_A} \\ T_B &= e^{\beta_0+\beta_1 (z+1)} e^{\sigma\epsilon_B} = e^{\beta_k} \left[e^{\beta_0} e^{\sigma\epsilon_B} \right] \end{aligned}$$

  donde los errores ϵ_A y ϵ_B tienen la misma distribución

Modelo acelerado

- Las funciones de supervivencia, para cada población, son

$$
\begin{aligned}
S_A(t) &= P(T_A > t) = P(ce^{\sigma\epsilon_A} > t) = P(e^{\sigma\epsilon_A} > c^{-1}t) \\
S_B(t) &= P(T_B > t) = P(e^{\sigma\epsilon_B} > c^{-1}e^{-\beta_1}t)
\end{aligned}
$$

donde $c = e^{\beta_0}$

- Dado que ϵ_A y ϵ_B tienen la misma distribución

$$
\begin{aligned}
S_B(e^{\beta_1}t) &= P(e^{\sigma\epsilon_B} > c^{-1}e^{-\beta_1}e^{\beta_1}t) = P(e^{\sigma\epsilon_B} > c^{-1}t) \\
&= P(e^{\sigma_A\epsilon} > c^{-1}t) \\
&= S_A(t)
\end{aligned}
$$

- **Nota**: Es importante notar que esto es independiente de la distribución que tienen los errores!

Modelo acelerado

- Supongamos que la probabilidad de supervivencia para la población A a los 30 dias es 0.5 ($S_A(30) = 0.5$)
  - Si $\beta_1 = 0.5$, entonces

$$S_B(49.5) = S_B(\exp(0.5) \times 30) = S_A(30) = 0.5$$

  - Si $\beta_1 = -0.5$, entonces

$$S_B(18.2) = S_B(\exp(-0.5) \times 30) = S_A(30) = 0.5$$

- Quiere decir que un β_1 positivo (negativo) retarda (acelera) la ocurrencia del evento

Modelo acelerado

- Supongamos que S_A y S_B son las funciones de supervivencia de dos poblaciones.
- El modelo acelerado del tiempo de falla establece:

$$S_A(t) = S_B(ct) \ , \ c > 0 \ , \ \forall t \geq 0$$

- Este modelo implica que la tasa de envejecimiento de la población A es <u>c veces</u> la tasa de envejecimiento de la población B.

Modelo acelerado

- Sea μ_A y μ_B la media del tiempo a que ocurra un evento en la población A y B, respectivamente:

$$\begin{aligned}\mu_B &= \int_0^\infty S_B(t)dt = c\int_0^\infty S_B(cu)du \\ &= c\int_0^\infty S_A(u)du \\ &= c\mu_A\end{aligned}$$

es decir, el tiempo esperado a que ocurra el evento en la población B es **c veces** el tiempo esperado a que ocurra el evento en la población A.

Modelo acelerado

- Sean $t_{A,50}$ y $t_{B,50}$ las medianas del tiempo a la ocurrencia del evento en las poblaciones A y B, respectimente:

$$0.5 = S_A(t_{A,50}) = S_B(ct_{A,50})$$

entonces, si S_B es estrictamente decreciente:

$$t_{B,50} = ct_{A,50}$$

es decir, la mediana del tiempo a que ocurra el evento en la población B es **c veces** la mediana del tiempo a que ocurra el evento en la población A.

Modelo acelerado

- Entonces, tenemos que

$$\begin{aligned} \mu_B &= e^{\beta_1}\mu_A \\ t_{B,50} &= e^{\beta_1}t_{A,50} \end{aligned} \qquad (1)$$

donde $t_{A,50}$ y $t_{B,50}$ son las medianas del tiempo a la ocurrencia del evento en las poblaciones A $(Z = z)$ y B $(Z = z + 1)$, respectivamente.

- En general, (1) se cumple para cualquier percentil. Es decir

$$t_{B,s} = e^{\beta_1}t_{A,s}$$

para $s \in \{1, \ldots, 99\}$

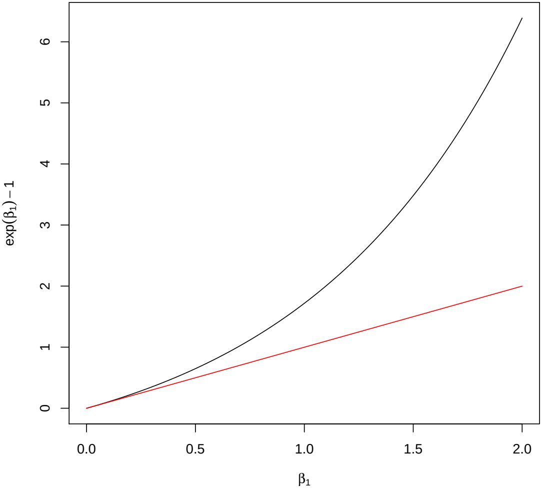
Modelo acelerado

- Si β_1 es pequeño

$$\frac{\mu_B - \mu_A}{\mu_A} = \frac{e^{\beta_1}\mu_A - \mu_A}{\mu_A} = e^{\beta_1} - 1 \approx \beta_1$$

- Entonces, si β_k es pequeño
  - Si $\beta_1 > 0$, entonces β_1 es el incremento porcentual en la media de tiempo cuando Z aumenta en una unidad
  - Si $\beta_1 < 0$, entonces β_1 es la disminución porcentual en la media de tiempo cuando Z aumenta en una unidad
- Si $\beta_1 > 0$, un mayor valor de Z **prolongara** la supervivencia de la población en estudio.

Modelo acelerado



Regresión exponencial

- El modelo satisface

$$\log(T) = \beta_0 + \beta_1 Z + \sigma\epsilon$$

con $\sigma = 1$ y la distribución de ϵ esta dada por

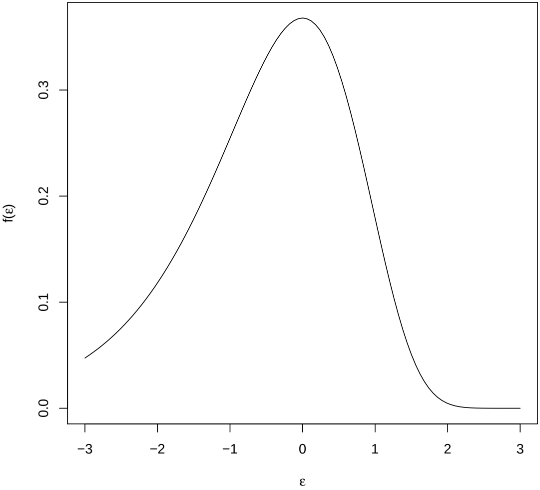
$$f(\epsilon) = e^{\epsilon - e^{\epsilon}}, \ \epsilon > 0$$

que es conocida como la distribución de <u>valores extremos</u>

- Dada la covariable $Z = z$, la distribución de T es exponencial:

$$T \mid Z = z \sim Exponencial(e^{-\beta z})$$

Distribución valores extremos



Regresión exponencial

- Dado que

$$T \mid Z = z \sim Exponencial(e^{-\beta z})$$

entonces la función de riesgo instantáneo

$$\lambda(t \mid Z, \beta) = e^{-Z\beta} \ , \ \ t > 0$$

y la de supervivencia son

$$S(t \mid Z, \beta) = e^{-te^{-Z\beta}} \ , \ \ t > 0$$

y la mediana del tiempo a que ocurra el evento es

$$t_{50}(Z, \beta) = -e^{Z\beta} \times \log(0.5)$$

Regresión exponencial

- Si consideramos una variable dicotómica (ej. sexo, etc), tenemos
$$t_{50}(Z=1,\beta)=e^{\beta_1}t_{50}(Z=0,\beta)$$
- Si por ejemplo $e^{\beta_1}=2$, entonces la mediana del tiempo a la ocurrencia del evento en grupo 1 ($Z=1$) es el doble que en el grupo 0 ($Z=0$).

Riesgos proporcionales

- Si incrementamos el valor de la covariable Z en una unidad de z a $z+1$, entonces el cociente de los riesgos, en un tiempo $T=t$, es

$$\begin{aligned}\frac{\lambda(t \mid Z=z+1)}{\lambda(t \mid Z=z)} &= \frac{e^{-\beta_0-\beta_1(z+1)}}{e^{-\beta_0-\beta_1 z}} \\ &= \frac{e^{-\beta_1}(e^{-\beta_0-\beta_1 z})}{e^{-\beta_0-\beta_1 z}} = e^{-\beta_1}\end{aligned}$$

- **Interpretación**: El riesgo de que ocurra el evento en la población $Z=z+1$ es $e^{-\beta_1}$ veces el riesgo de que ocurra el evento en la población $Z=z$

Regresión exponencial: Cancer

```
> model1 <- survreg(Surv(time,dead)~factor(Group),dist="exp",data=cancer)
> summary(model1)

Call:
survreg(formula = Surv(time, dead) ~ factor(Group), data = cancer,
    dist = "exp")
                         Value  Std. Error       z        p
(Intercept)             1.1083       0.258   4.2926 1.77e-05
factor(Group)Completed -0.0221       0.309  -0.0716 9.43e-01

Scale fixed at 1

Exponential distribution
Loglik(model)= -104.6   Loglik(intercept only)= -104.6
Chisq= 0.01 on 1 degrees of freedom, p= 0.94
Number of Newton-Raphson Iterations: 5
n= 63
```

- Scale fixed at 1 ($\sigma = 1$)
- Modelo

$$\lambda(t \mid Z) = e^{-1.1+0.02Z}, \; Z \in \{0, 1\}$$

Regresión exponencia: Cancer

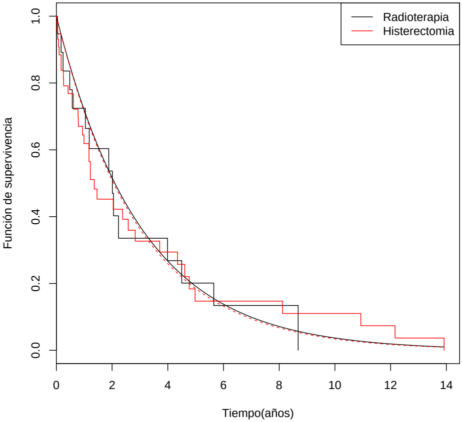


Figure : Estimador de Kaplan-Meier y en base al modelo exponencial para la función de supervivencia para ambos grupos

Regresión exponencial: Cancer

```
> model2 <- survreg(Surv(time,dead)~Age,dist="exp",data=cancer)
> summary(model2)

Call:
survreg(formula = Surv(time, dead) ~ Age, data = cancer, dist = "exp")
               Value Std. Error     z        p
(Intercept)   2.6507     0.5075  5.22 1.76e-07
Age          -0.0346     0.0102 -3.40 6.73e-04

Scale fixed at 1

Exponential distribution
Loglik(model)= -99.7   Loglik(intercept only)= -104.6
Chisq= 9.96 on 1 degrees of freedom, p= 0.0016
```

- Modelo

$$\lambda(t \mid Z) = e^{-2.7+0.03Z}$$

  donde Z es la variable edad

- **Interpretación**: De acuerdo a este modelo, a medida que la edad de diagnostico aumenta, el riesgo es mayor

Regresión exponencia: Cancer

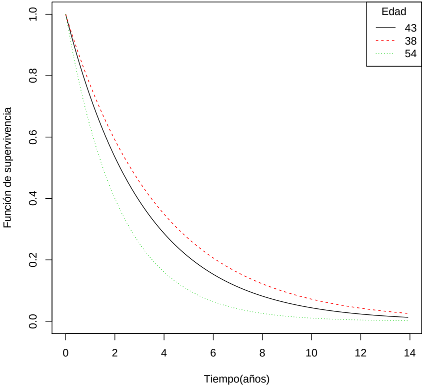


Figure : Estimador en base al modelo exponencial para la función de supervivencia para el primer, segundo y tercer cuartil de edades

Regresión exponencial: Cancer

```
> model3 <- survreg(Surv(time,dead)~factor(Group)+Age,dist="exp",data=cancer)
> summary(model3)

Call:
survreg(formula = Surv(time, dead) ~ factor(Group) + Age, data = cancer,
    dist = "exp")
                          Value Std. Error      z        p
(Intercept)              2.8107     0.5776  4.866 1.14e-06
factor(Group)Completed  -0.1781     0.3114 -0.572 5.67e-01
Age                     -0.0354     0.0102 -3.477 5.07e-04

Scale fixed at 1

Exponential distribution
Loglik(model)= -99.5   Loglik(intercept only)= -104.6
Chisq= 10.29 on 2 degrees of freedom, p= 0.0058
Number of Newton-Raphson Iterations: 5
n= 63
```

- Modelo

$$\lambda(t \mid Z) = e^{-2.81+0.18Z_1+0.04Z_2}$$

donde Z_1 es histerectomia o abandono y Z_2 es edad

Regresión exponencial: Cancer

Interpretación:

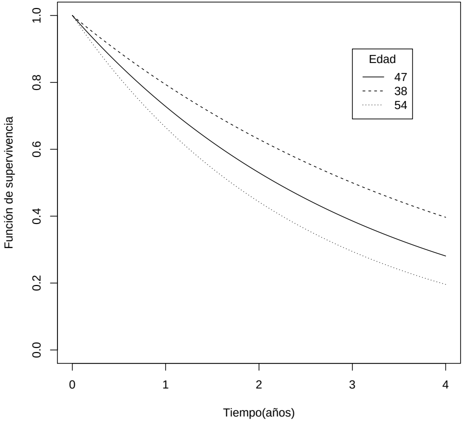
- Controlando por edad, no hay evidencia que la media y/o mediana del tiempo de vida sea diferente para mujeres que se realizaron una histerectomia que para aquellas que no lo hicieron ($p = 0.567$).
- Controlando por edad, no existe diferencia en el riesgo de morir entre mujeres que abandonaron la histerectomia en comparación con aquellas que no lo hicieron ($p = 0.567$).

Regresión exponencial: Cancer

- La mediana del tiempo de vida para mujeres a las cuales se les diagnostico 48, 37 y 54 años (y que abandonaron el proceso de histerectomia) es

$$\begin{aligned}
\hat{t}_{50,Edad=37} &= -e^{2.81-0.04(37)}\log(0.5) = 2.6 \\
\hat{t}_{50,Edad=48} &= -e^{2.81-0.04(48)}\log(0.5) = 1.7 \\
\hat{t}_{50,Edad=54} &= -e^{2.81-0.04(55)}\log(0.5) = 1.3
\end{aligned}$$

Regresión exponencial:Cancer



Regresión exponencial: VIH

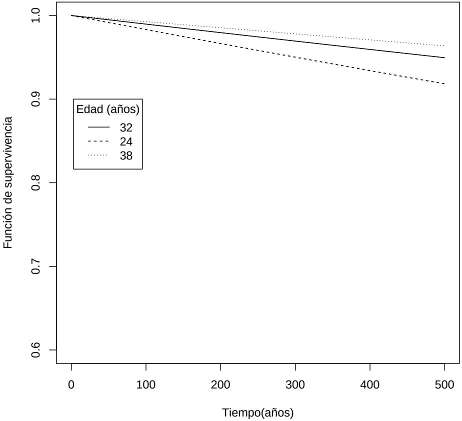
```
> model1 <- survreg(Surv(time,cen)~arm+age+circum+factor(edu)+npartner,dist="exp",data=hiv)
> summary(model1)
```

	Value	Std. Error	z	p
(Intercept)	7.2465	0.72622	9.978	1.89e−23
arm	0.1502	0.21870	0.687	4.92e−01
age	0.0596	0.01486	4.010	6.07e−05
circum	−0.2820	0.30012	−0.939	3.48e−01
factor(edu)1	0.3676	0.60166	0.611	5.41e−01
factor(edu)2	0.0629	0.60865	0.103	9.18e−01
npartner	−0.0055	0.00357	−1.540	1.24e−01

Interpretación:

- La media/mediana del tiempo libre de VIH se incrementa $e^{10\times 0.06} = 1.82$ veces ante un incremento de 10 años en la edad ($p < 0.001$)
- El riesgo de infección por VIH se reduce en un 45% ($HR = e^{-10\times 0.06} = 0.55$) ante un incremento de 10 años en la edad.
- El resto de variables no estan asociadas con infección de VIH.

Regresión exponencial: VIH



Modelo exponencial: Estimación

- Consideremos una muestra $(X_1, \Delta_1, Z_1), \ldots (X_n, \Delta_n, Z_n)$
- La función de verosimilitud esta dada por

$$
\begin{aligned}
L_n(\beta) &= \prod_{i=1}^{n} f(x_i)^{\delta_i} S(x_i)^{1-\delta_i} = \prod_{i=1}^{n} \lambda(x_i)^{\delta_i} S(x_i) \\
&= \prod_{i=1}^{n} e^{-Z_i^t \beta \delta_i} \exp\left(-e^{(-Z_i^t \beta)} x_i\right)
\end{aligned}
$$

por ende el logaritmo de esta esta dado por

$$
l_n(\beta) = -\sum_{i=1}^{n} Z_i^t \beta \delta_i - \sum_{i=1}^{n} e^{(-Z_i^t \beta)} x_i
$$

Regresión exponencial: Estimación

- La función de score

$$\frac{\partial l}{\partial \beta_k} = \sum_{i=1}^{n} z_k(\delta_i - \exp(-z_i^t \beta x_i))$$

- Los elementos de la matriz de información observada $I(\hat{\beta})$

$$\frac{\partial^2 l}{\partial \beta_j \partial \beta_k} = -\sum_{i=1}^{n} z_k z_i \exp(-z_i^t \beta x_i))$$

- La matriz de varianza covarianza esta dada por

$$\hat{Var}(\hat{\beta}) = I(\hat{\beta})^{-1}$$

Distribución Weibull

- Sabemos que si $T \sim Weibull(a, b)$ donde a y b son los parámetros de forma y escala, respectivamente. Entonces

$$
\begin{aligned}
\lambda(t) &= \frac{a}{b}\left(\frac{t}{b}\right)^{a-1} \\
S(t) &= \exp\left[-\left(\frac{t}{b}\right)^{a}\right] \\
\Lambda(t) &= \left(\frac{t}{b}\right)^{a}
\end{aligned}
$$

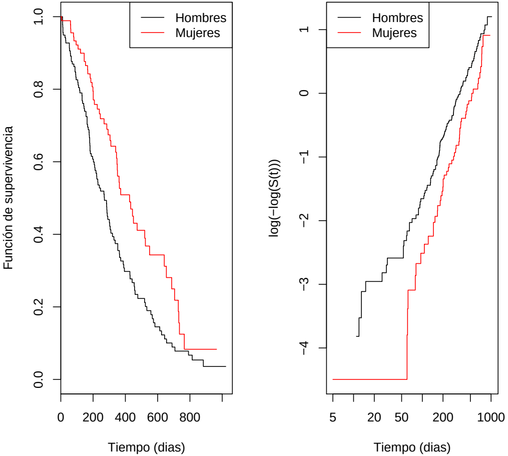
- Otra parametrización es posible si $\lambda = b^{-a}$, entonces

$$
\begin{aligned}
\lambda(t) &= \lambda a t^{a-1} \\
S(t) &= \exp\left(-\lambda t^{a}\right) \\
\Lambda(t) &= \lambda t^{a}
\end{aligned}
$$

Modelo Weibull

- Note que si
$$S(t) = \exp\left(-\lambda t^{a}\right)$$
entonces
$$\log\left[-\log\left(S(t)\right)\right] = \log\left(\lambda\right) + a\log\left(t\right) \qquad (2)$$
- Podriamos reemplazar el estimador de Kaplan-Meier en (2) para evaluar si el modelo Weibull es apropiado para nuestros datos
- Si $a = 1$, el modelo se reduce al modelo exponencial

Distribución Weibull

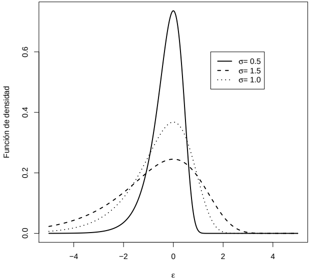


Regresión Weibull

- Asumimos que $\sigma\epsilon$ tiene una distribución de valores extremos con parámetro de escala σ.

$$f(\epsilon) = \frac{1}{\sigma} e^{\epsilon/\sigma - e^{\epsilon/\sigma}}$$

note que $\sigma = 1$ corresponde al caso de regresión exponencial.



Riesgos proporcionales

- Sea Z una covariable de interes dicotómica (0 y 1)
- Supongamos que se cumple que $\lambda = e^{\beta_0+\beta_1 Z}$, entonces

$$\lambda(t) = ae^{\beta_0+\beta_1 Z}t^{a-1}$$

- El cociente de riesgos entre los grupos $Z = 1$ y $Z = 0$ es

$$\begin{aligned}\frac{\lambda(t \mid Z=1)}{\lambda(t \mid Z=0)} &= \frac{ae^{\beta_0+\beta_1 Z}t^{a-1}}{ae^{\beta_0}t^{a-1}} \\ &= e^{\beta_1}\end{aligned}$$

Riesgos proporcionales

Si sabemos que

$$\frac{\lambda(t \mid Z = 1)}{\lambda(t \mid Z = 0)} = e^{\beta_1}$$

entonces

- Si $\beta_1 = 0$, quiere decir que el riesgo es el mismo para ambos grupos.
- Si $\beta_1 < 0$, el riesgo es menor en el grupo $Z = 1$
- Si $\beta_1 > 0$, el riesgo es mayor en el grupo $Z = 1$

Tiempos acelerados

- La función de superviencia tiene la forma

$$S(t) = e^{-\lambda t^a}$$

entonces

$$t = \frac{1}{\lambda^{1/a}} \left[-\log \left[S(t) \right] \right]^{1/a}$$

- Si suponemos que

$$\begin{aligned} \frac{1}{\lambda^{1/a}} &= e^{\alpha_0 + \alpha_1 Z} \\ S(t) &= q \end{aligned}$$

entonces

$$t = \left[-\log \left(q \right) \right]^{1/a} e^{\alpha_0 + \alpha_1 Z}$$

Tiempos acelerados

- Para la mediana, tenemos $q = 0.5$, entonces

$$\begin{aligned} \frac{t_{m,Z=1}}{t_{m,Z=0}} &= \frac{[-\log(0.5)]^{1/a} e^{\alpha_0+\alpha_1}}{[-\log(0.5)]^{1/a} e^{\alpha_0}} \\ &= e^{\alpha_1} \end{aligned}$$

- **Interpretación**: La mediana del tiempo de vida, en el grupo $Z = 1$ es e^{α_1} veces la mediana del tiempo de vida en el grupo $Z = 0$

Conección entre ambos modelos

- Para el modelo de riesgos proporcionales se cumple

$$\log(\lambda) = \beta_0 + \beta_1 Z$$

- Para el modelo de tiempos acelerados

$$\frac{1}{a}\log(\lambda) = -\alpha_0 - \alpha_1 Z$$

entonces

$$\log(\lambda) = -a(\alpha_0 + \alpha_1 Z)$$

- Entonces

$$\beta_i = -a\alpha_i$$

Regresión Weibull

```
> model.weibull.1 <- survreg(Surv(time,status)~ sex,dist="w",data=cancer)
> summary(model.weibull.1)

Call:
survreg(formula = Surv(time, status) ~ sex, data = cancer, dist = "w")
              Value Std. Error     z          p
(Intercept)   5.489     0.1790 30.66 2.10e-206
sex           0.396     0.1276  3.10  1.94e-03
Log(scale)   -0.281     0.0619 -4.54  5.71e-06

Scale= 0.755
```

- En la función *survreg* se implementa el modelo de riesgos acelerados
- a y b son iguales a

$$\begin{aligned} a &= (\textit{survreg scale})^{-1} \\ b &= \exp(\textit{Intercepto}) \end{aligned}$$

Regresión Weibull

- Tiempos acelerados

$$\frac{t_{m,mujer}}{t_{m,hombre}} = e^{0.396} = 1.49$$

Interpetación: La mediana del tiempo de vida en mujeres es 49% mayor que en hombres

- Riesgos proporcionales

$$\begin{aligned} \frac{\lambda(t \mid mujer)}{\lambda(t \mid hombre)} &= e^{\beta_1} = exp\left[-\frac{0.394}{survreg\ scale}\right] \\ &= e^{-0.394/0.755} = 0.59 \end{aligned}$$

Interpetación: El riesgo de muerte en mujeres es 41% mejor que en hombres

Regresión Weibull

- La función de riesgo tiene la forma

$$\lambda_0(t) = at^{a-1}e^{-a(\alpha_0+\alpha_1)Z}$$

- En nuestra aplicación

$$\begin{aligned} a &= 1/0.755 = 1.32 \\ \alpha_0 &= 5.489 \\ \alpha_0 &= 0.396 \end{aligned}$$

- Entonces

$$\begin{aligned} \lambda_M &= 0.0009t^{0.32} \\ \lambda_H &= 0.0005t^{0.32} \end{aligned}$$

Regresión Weibull

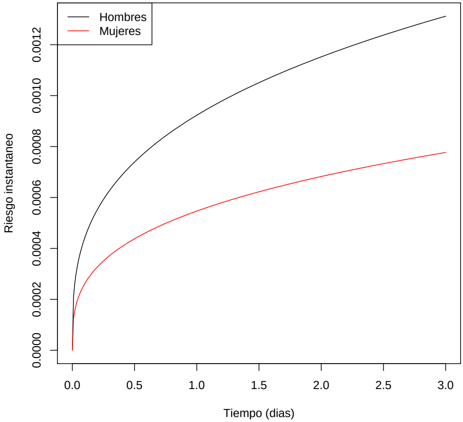


Figure : Funciones de riesgo en base al modelo Weibull para hombres y mujeres

Regresión Weibull

```
> model.weibull.2 <- survreg(Surv(time,status)~ sex +age,dist="w",data=cancer)
> summary(model.weibull.2)

Call:
survreg(formula = Surv(time, status) ~ sex + age, data = cancer,
    dist = "w")
              Value Std. Error     z        p
(Intercept)  6.2749    0.48137 13.04 7.69e-39
sex          0.3821    0.12748  3.00 2.72e-03
age         -0.0123    0.00696 -1.76 7.81e-02
Log(scale)  -0.2823    0.06188 -4.56 5.07e-06

Scale= 0.754
```

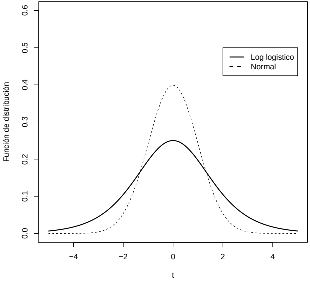
Interpretación:

- Controlando por edad, la media (mediana) del tiempo de vida en mujeres es 1.46 ($e^{0.38}$) veces la mediana del tiempo de vida en hombres ($p = 0.002$).
- Controlando por el sexo del pacientes, pacientes con un año mas de edad en el momento de diagnóstico sufren una reducción de 1% en su mediana del tiempo de vida ($p = 0.07$).

Regresión log logistica

- El modelo log logistico asume que ϵ tiene una distribución logistica

$$f(\epsilon) = \frac{e^{\epsilon}}{(1 + e^{\epsilon})^2}$$



Log logistico

- La función de riesgo tiene la forma

$$\lambda(t) = \frac{\lambda p t^{p-1}}{1 + \lambda t^p}$$

donde $\lambda > 0$ y $p > 0$

- Función de supervivencia

$$S(t) = \frac{1}{1 + \lambda t^p}$$

- El odds de muerte (ocurrencia del evento) es

$$\frac{1 - S(t)}{S(t)} = \frac{P(T \leq t)}{P(T > t)} = \lambda t^p$$