Modelo de riesgos proporcionales

Giancarlo Sal y Rosas

Departmento de Ciencias
Pontificia Universidad Católica del Perú

May 5, 2017

Outline

Modelo

- El modelo de riesgos proporcionales (PH) tiene la forma

$$\lambda(t \mid Z) = \lambda_0(t) \exp\left[Z^t \beta\right]$$

  donde λ_0 es una función de riesgo basal (es decir no depende de covariables)

- El modelo tambien se puede expresar como

$$\log\left[\lambda(t \mid Z)\right] = \log\left[\lambda_0(t)\right] + Z^t \beta$$

  es decir el logaritmo de la función de riesgo en el tiempo $T = t$ es una <u>función lineal</u> de las covariables.

Modelo

- El modelo Exponencial es un caso particular

$$\lambda_0(t) = 1$$

  es decir el riesgo basal es constante.

- El modelo Weibull es otro caso particular

$$\lambda_0(t) = at^{a-1}$$

- El modelo log logístico no es un caso particular del modelo PH

Modelo

- La verosimilitud del model PH tiene la forma

$$
\begin{aligned}
L_n(\beta, \lambda_0) &= \prod_{i=1}^{n} f(x_i \mid Z_i)^{\delta_i} S(x_i \mid Z_i)^{1-\delta_i} \\
&= \prod_{i=1}^{n} \lambda(x_i \mid Z_i)^{\delta_i} S(x_i \mid Z_i) \\
&= \prod_{i=1}^{n} [\lambda_0(x_i) e^{Z_i^t \beta}]^{\delta_i} \exp\left[- \int_0^{x_i} \lambda_0(t) e^{Z_i(t)\beta} dt \right] \qquad (1)
\end{aligned}
$$

donde $\beta \in R^k$ (k = # de covariables) y λ_0 pertenece a un espacio de funciones

- Optimizar (1) es un problema muy complicado dada la naturaleza de λ_0

Modelo: Verosimilitud parcial

- Cox (1972) descompuso la verosimilitud total en la forma

$$L_n(\beta, \lambda_0) = L_p(\beta) \times L_{resto}(\lambda_0, \beta)$$

y llama a L_p la "**verosimilitud parcial**"

- Cox (1972) afirmo que bastaba usar $L_p(\beta)$ para estimar los coeficientes de regresión:

$$L_p(\beta) = \prod_{i=1}^{m} \frac{\exp\,[Z_i^t \beta]}{\sum_{j \in R(t_i)} \exp\,[Z_j^t \beta]}$$

donde $R(t_i)$ es el conjunto de individuos que estan en riesgo en el el tiempo t_i y m es el número de eventos observados.

Historia

- En 1972, Sir David Cox (no era Sir aun) publicó su artículo: **Regression models and life tables** en el Journal of the Royal Statistical Society
- Este fue leido ante la Royal Statistical Society el 8 de marzo de 1972
- En este artículo, Cox discute modelos de regressión para datos censurados y como tablas de vida son un caso particular de estos. En este proceso introduce el concepto de **verosimilitud parcial**
- En el artículo original se incluyen opiniones de Peto, Oakes, Barton, Prentice, Kalbfleisch, Breslow, entre otros.

Historia

- La idea de **verosimilitud parcial** consiste en usar solo parte de la data para estimar los parámetros de regresión del modelo PH.
- En 1975, en su articulo **Partial Likelihood**, Cox fundamenta y formaliza su idea de **verosimilitud parcial** (pero sin mucho detalle ...)
- En 1981, Tsiatis en su articulo **A large sample study of the Cox regression model** publica una prueba rigurosa que el estimador de los coeficientes de regresión obtenido via la verosimilitud parcial **converge** al valor real de estos.

Historia

- David Cox ha recibido imnumerables premios por su trayectoria como estadístico entre los cuales podemos mencionar:
    - Guy Medals in Silver y Gold (1961 y 1963) de la Royal Statistical Society
    - Nombrado caballero por la reina Elizabeth en 1985
    - Miembro asociado extranjero de la academia nacional de ciencias de los EU
    - Miembro de la Royal Danish Academy of Sciences and Letters
    - Honorary Fellow de la Academia Británica en el 2000

Interpretación: Variable binaria

- Nuestro modelo univariado es de la forma

$$\lambda(t \mid Z) = \lambda_0(t)e^{\beta Z}$$

- En el caso de una variable binaria (Ejemplo: sexo) tenemos

$$\frac{\lambda(t \mid Z = hombre)}{\lambda(t \mid Z = mujer)} = \frac{\lambda_0(t)e^{\beta\times 1}}{\lambda_0(t))e^{\beta\times 0}} = e^{\beta}$$

- **Interpretación**: El riesgo de desarrollar el evento en hombres es e^{β} veces el riesgo de desarrollar el evento en mujeres

Regresión: Variable binaria

- Codigo de R

```
> model1 <- coxph(Surv(time,status)~sex,data=cancer)
> summary(model1)
Call:
coxph(formula = Surv(time, status) ~ sex, data = cancer)

  n= 228, number of events= 165

       coef exp(coef) se(coef)      z Pr(>|z|)
sex -0.5310    0.5880   0.1672 -3.176  0.00149 **

    exp(coef) exp(-coef) lower .95 upper .95
sex     0.588      1.701    0.4237     0.816

Concordance= 0.579  (se = 0.022 )
Rsquare= 0.046   (max possible= 0.999 )
Likelihood ratio test= 10.63  on 1 df,   p=0.001111
Wald test            = 10.09  on 1 df,   p=0.001491
Score (logrank) test = 10.33  on 1 df,   p=0.001312
```

- Modelo PH tiene la forma

$$\lambda(t) = \lambda_0(t)e^{-0.53Z}$$

donde $Z = 0$ (1) si la persona es hombre (mujer)

Regresión: Variable binaria

- Intervalo de confianza para $\hat{\beta}$

```
> confint(model1)
         2.5 %     97.5 %
sex -0.8586875 -0.2033595
```

- Intervalo de confianza para el cociente de riesgos: $e^{\hat{\beta}}$

```
> exp(confint(model1))
        2.5 %    97.5 %
sex 0.4237178 0.8159848
```

- **Interpretación**: Tengo un 95% de confianza que la diferencia en el riesgo en mujeres (en comparación con hombres) se encuentra entre una reducción del 18% y 58%

Interpretación: Variable politómica

- Codigo de R

```
> cancer$ph.ecog <- ifelse(cancer$ph.ecog < 2,cancer$ph.ecog,2)
> model2          <- coxph(Surv(time,status)~factor(ph.ecog),data=cancer)
> summary(model2)

  n= 227, number of events= 164
   (1 observation deleted due to missingness)

                    coef exp(coef) se(coef)     z Pr(>|z|)
factor(ph.ecog)1 0.3682    1.4451   0.1987 1.853   0.0638 .
factor(ph.ecog)2 0.9309    2.5369   0.2235 4.166 3.11e-05 ***

                 exp(coef) exp(-coef) lower .95 upper .95
factor(ph.ecog)1     1.445     0.6920    0.9791     2.133
factor(ph.ecog)2     2.537     0.3942    1.6371     3.931
```

- Modelo PH tiene la forma

$$\lambda(t) = \lambda_0(t)e^{0.37Z_1+0.93Z_2}$$

donde $Z_i = 1$ (0) si la persona tuvo un score ECOG igual a i (o no)

Interpretación: Variable politómica

- Persona con score 1 vs. score 0

$$\frac{\lambda(t \mid Z=1)}{\lambda(t \mid Z=0)} = \frac{\lambda_0(t)e^{0.37\times 1+0.93\times 0}}{\lambda_0(t)e^{0.37\times 0+0.93\times 0}} = e^{0.37} = 1.4$$

Interpretación: Pacientes cuyo score ECOG fue 1, tiene 40% mayor riesgo de morir que aquellos que tuvieron score ECGO igual a 0.

- Persona con score 2 o 3 vs. score 0

$$\frac{\lambda(t \mid Z=2,3)}{\lambda(t \mid Z=0)} = \frac{\lambda_0(t)e^{0.37\times 0+0.93\times 1}}{\lambda_0(t)e^{0.37\times 0+0.93\times 0}} = e^{0.93} = 2.5$$

Interpretación: Pacientes cuyo score ECOG fue 2 o 3, tiene 2.5 veces el riesgo de morir que pacientes cuyo score ECOG igual a 0.

Regresión: Variable politómica

- Intervalos de confianza

```
> confint(model2)
                        2.5 %     97.5 %
factor(ph.ecog)1 -0.02117005 0.7575425
factor(ph.ecog)2  0.49292109 1.3689575
> exp(confint(model2))
                     2.5 %   97.5 %
factor(ph.ecog)1 0.9790525 2.133028
factor(ph.ecog)2 1.6370913 3.931250
```

- **Interpretación**:
  - Pacientes cuyo score ECOG fue 2 o 3, tiene 2.5 veces el riesgo riesgo de morir que aquellos que tuvieron score ECGO igual a 0. Con una confianza del 95%, esta diferencia podria estar entre 1.6 y 3.9 veces.
  - No tenemos evidencia, con una significancia del 5% , para afirmar que un score ECOG 1 incrementa el riesgo de muerte en comparación con un score ECOG de 0 (95%IC: 0.97-2.13)

Interpretación: Variable politómica

- Deseamos comparar el riesgo entre pacientes con score ECOG 1 vs. pacientes con score ECOG 2 o 3:

$$\frac{\lambda(t \mid Z = 2,3)}{\lambda(t \mid Z = 1)} = \frac{\lambda_0(t)e^{0.37\times 0+0.93\times 1}}{\lambda_0(t)e^{0.37\times 1+0.93\times 0}}$$
$$= e^{0.93-0.37} = e^{0.56} = 1.75$$

- **Interpretación**: Pacientes cuyo score ECOG fue 2 o 3, tiene 1.8 veces el riesgo de morir que pacientes cuyo score ECOG igual a 1.

Interpretación: Variable politómica

- **R** nos entrega la matriz de varianza covarianza de los coeficientes de regresión.

```
> model2$var
           [,1]       [,2]
[1,] 0.03946374 0.02734994
[2,] 0.02734994 0.04994456
```

$$\begin{aligned} Var(\hat{\beta}_3 - \hat{\beta}_2) &= Var(\hat{\beta}_3) + Var(\hat{\beta}_2) - Cov(\hat{\beta}_3, \hat{\beta}_2) \\ &= 0.05 + 0.04 - 0.03 = 0.06 \end{aligned}$$

- El intervalo de confianza al 95%, para el CRI, es

```
> c(0.56 - 1.96*sqrt(0.06),0.56 + 1.96*sqrt(0.06))
[1] 0.07990001 1.04009999
> exp(c(0.56 - 1.96*sqrt(0.06),0.56 + 1.96*sqrt(0.06)))
[1] 1.083179 2.829500
```

Interpretación: Tengo un 95% de confianza que el riesgo para los pacientes con score ECOG 2 o 3 aumenta entre 1.08 y 2.83 veces en comparación con pacietnes con score ECOG 1

Interpretación: Variable continua

- Relación entre la edad en el momento de diagnóstico y el tiempo de vida

```
> model3 <- coxph(Surv(time,status)~age,data=cancer)
> summary(model3)
Call:
coxph(formula = Surv(time, status) ~ age, data = cancer)

  n= 228, number of events= 165

        coef exp(coef) se(coef)     z Pr(>|z|)
age 0.018720  1.018897 0.009199 2.035   0.0419 *

    exp(coef) exp(-coef) lower .95 upper .95
age     1.019     0.9815     1.001     1.037
```

- **Interpretación**: Un año adicional en la edad de diagnostico esta asociado con un incremento enel riesgo de 2 % (95%IC: 0.1-3.7,)

Interpretación: Variable continua

- El aumento o disminución en el riesgo al comparar una personas de edad $Z = a$ con una de edad $Z = a + 10$ esta dado por

$$\frac{\lambda(t \mid Z = a + 10)}{\lambda(t \mid Z = a)} = \frac{\lambda_0(t)e^{0.02(a+10)}}{\lambda_0(t)e^{0.02a}} = e^{0.02\times 10} = 1.22$$

- **Interpretación**: Una diferencia de 10 años en la edad en el momento de diagnostico esta asociada con un incremento en el riesgo del 22%.

- Su intervalo de confianza al 95% esta dado por

```
> c(0.02*10 - 1.96*10*sqrt(model3$var),0.02 + 1.96*10*sqrt(model3$var))
[1] 0.01969555 0.20030445

> exp(c(0.02*10 - 1.96*10*sqrt(model3$var),0.02 + 1.96*10*sqrt(model3$var)))
[1] 1.019891 1.221775
```

Interpretación: Modelo multivariado

- Modelo ajustado por edad, score ECOG y sexo del paciente

```
> model4 <- coxph(Surv(time,status)~age + sex + factor(ph.ecog),data=cancer)
> summary(model4)

  n= 227, number of events= 164
   (1 observation deleted due to missingness)

                      coef  exp(coef)  se(coef)      z  Pr(>|z|)
age               0.011031  1.011092  0.009297  1.186  0.23544
sex              -0.551322  0.576188  0.167987 -3.282  0.00103 **
factor(ph.ecog)1  0.409461  1.506006  0.199596  2.051  0.04022 *
factor(ph.ecog)2  0.915752  2.498654  0.227042  4.033  5.5e-05 ***

                 exp(coef) exp(-coef) lower .95 upper .95
age                 1.0111     0.9890    0.9928    1.0297
sex                 0.5762     1.7355    0.4145    0.8009
factor(ph.ecog)1    1.5060     0.6640    1.0184    2.2270
factor(ph.ecog)2    2.4987     0.4002    1.6012    3.8991
```

- Despues de controlar por el sexo del paciente y el score ECOG, la edad en el momento del diagnostico no esta asociado con el tiempo de vida del paciente ($p = 0.24$)

Regresión: Multivariado

- Parte de interpretación

	exp(**coef**)	**exp**(-**coef**)	**lower** .95	**upper** .95
age	1.0111	0.9890	0.9928	1.0297
sex	0.5762	1.7355	0.4145	0.8009
factor(ph.ecog)1	1.5060	0.6640	1.0184	2.2270
factor(ph.ecog)2	2.4987	0.4002	1.6012	3.8991

- **Sexo**: Las pacientes mujeres tienen 42% menos riesgo de morir (HR= 0.58, 95% IC: 0.41-0.80), comparación con los hombres, despues de controlar por otras variables
- **Score ECOG**: Pacientes con score ECOG 2 o 3 tienen 2.5 veces el riesgo de morir en comparación con pacientes con score ECOG 0 (HR=2.5, 95%: 1.6 - 3.9)

Regresión: Multivariado

- Comparación de modelos

```
> model4 <- coxph(Surv(time,status)~age + sex + factor(ph.ecog),data=cancer)
> model5 <- coxph(Surv(time,status)~ sex + factor(ph.ecog),data=cancer)
> anova(model5,model4)
Analysis of Deviance Table
 Cox model: response is  Surv(time, status)
 Model 1: ~ sex + factor(ph.ecog)
 Model 2: ~ age + sex + factor(ph.ecog)
   loglik  Chisq Df P(>|Chi|)
1 -730.15
2 -729.44 1.4273  1    0.2322
```

- Esta prueba evalua la hipótesis

$$H_0 : \beta_{edad} = 0 \ vs. \ H_0 : \beta_{edad} \neq 0$$

- El estadistico es el cociente de *log verosimilitud*:

$$2[l_n(modelo\ 4) - l_n(modelo\ 5)] \rightarrow_{H_0} \chi^2_{(1)}$$

Estimación del riesgo basal

- Coeficientes de regresión se estiman usando la verosimilitud parcial
- Esta estimación se realiza sin necesidad de estimar la función de riesgo basal, $\lambda_0(\cdot)$.
- En algunos circunstancias es de interes estimar $\lambda_0(\cdot)$
  - Estimos interesados en la predicción del tiempo de sobrevida
  - Usar $\lambda_0(\cdot)$ para explicar mejor las implicancias de nuestro modelo de riesgos proporcionales.

Estimación del riesgo basal

- Recordemos que el modelo de Cox se expresa como

$$\lambda(t \mid Z) = \lambda_0(t) \exp\left[Z^t \beta\right]$$

que en base a la función de riesgo acumulada se expresa como

$$\Lambda(t \mid Z) = \Lambda_0(t) \exp\left[Z^t \beta\right]$$

y que es equivalente a

$$S(t \mid Z) = \left[S_0(t)\right]^{\exp\left[Z^t \beta\right]}$$

- La verosimilitud esta dada por

$$L_n(\beta, \lambda_0) = \prod_{i=1}^{n} \left[\lambda_0(x_i) e^{Z_i^t \beta)}\right]^{\delta_i} \exp\left[-\int_0^{x_i} \lambda_0(t) e^{Z_i^t \beta} dt\right]$$

Estimación del riesgo basal

- Supongamos que observamos k eventos en tiempos distintos: $t_1, \ldots, t_k$
- Supongamos que conocemos el valor de β.
- La función de verosimilitud, en función de $\lambda_0(.)$, nos sugiere:
  - *Mirando el segundo factor*: Quisieramos que λ_0 sea lo mas pequeño posible.
  - *Mirando el primer factor*: Para observaciones $\delta_i = 1$, deseamos que el valor de $\lambda_0(t_i)$ sea el mayor posible.
  - Podemos reducir el problema asumiendo que

$$\hat{\lambda}_0(t) = 0 \; , \; t \notin \{t_1, \ldots, t_k\}$$

es decir para los puntos donde observamos censuras.

Estimación del riesgo basal

- El logaritmo de la función de verosimilitud se reduce a

$$l_n(\lambda_0) = \sum_{i=1}^{k} \left[\log\left(\lambda_0(t_i)\right) + Z_i^t\beta\right] - \sum_{i=1}^{k} \left(\lambda_0(t_i) \sum_{j \in R(t_i)} \exp Z_j^t \beta\right)$$

  donde $R(t_i)$ es el conjunto de riesgo hasta antes del tiempo t_i.

- Si asumimos que β es conocido, tenemos k parámetros a estimar:

$$\lambda_0(t_1), \ldots, \lambda_0(t_k)$$

Estimación del riesgo basal

- El estimador de máxima verosimilitud para $\lambda_0(t_i)$ es

$$\hat{\lambda}_0(t_i) = \frac{1}{\sum_{j \in R(t_i)} \exp\left(Z_j^t \hat{\beta}\right)}$$

- La función de riesgo acumulada es

$$\hat{\Lambda}_0(t) = \sum_{t_i \leq t} \hat{\lambda}_0(t_i) = \sum_{t_i \leq t} \frac{1}{\sum_{j \in R(t_i)} \exp\left(Z_j^t \hat{\beta}\right)}$$

que conocemos como el *estimador de Breslow*

- La función de supervivencia en base a este modelo es

$$\hat{S}_0(t) = e^{-\hat{\Lambda}_0(t)}$$

y es una función de saltos en los puntos $\{t_1, \ldots, t_k\}$

Ejemplo: Cáncer

- Codigo de R

```
> model1 <- coxph(Surv(time,status)~sex,data=cancer)
> summary(model1)

  n= 228, number of events= 165

        coef exp(coef) se(coef)      z Pr(>|z|)
sex -0.5310    0.5880   0.1672 -3.176  0.00149 **
```

- Al tener ambos estimadores $\hat{\beta}$ y $\hat{S}_0$, podemos estimar la función de supervivencia

$$\hat{S}(t, \hat{\beta}) = [\hat{S}_0(t)]^{\exp(Z^t \hat{\beta})}$$

- En el caso de la variable historia reciente de consumo tenemos:

$$\begin{aligned} \hat{S}(t_i \mid Hombre) &= \hat{S}_0(t_i) \\ \hat{S}(t_i \mid Mujer) &= \hat{S}_0(t_i)^{\exp(0.32)} \end{aligned}$$

Ejemplo: Cáncer

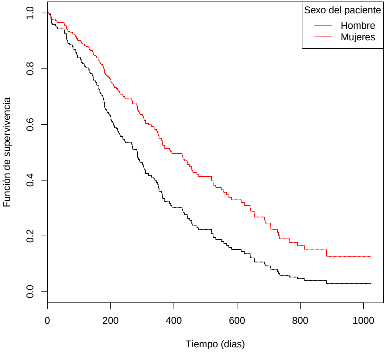

Figure : Función de supervivencia estimada bajo el modelo de proportionalidad de Cox

Ejemplo: Cáncer

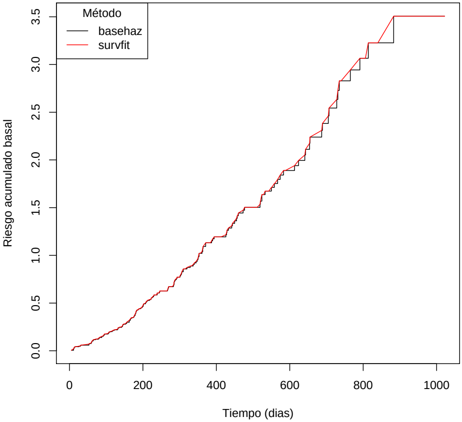

Figure : Función basal: Usando *basehaz* y *survfit*

Función de supervivencia ajustada

- Supongamos que queremos estudiar las diferencias en el riesgo por sexo ajustando por edad

```
> model3 <- coxph(Surv(time,status)~age+sex,data=cancer)
> summary(model3)
Call:
coxph(formula = Surv(time, status) ~ age + sex, data = cancer)

  n= 228, number of events= 165

         coef exp(coef)  se(coef)      z Pr(>|z|)
age  0.017045  1.017191  0.009223  1.848  0.06459 .
sex -0.513219  0.598566  0.167458 -3.065  0.00218 **

    exp(coef) exp(-coef) lower .95 upper .95
age    1.0172     0.9831    0.9990    1.0357
sex    0.5986     1.6707    0.4311    0.8311
```

- Modelo

$$\Lambda(t \mid Z) = \Lambda_0(t) e^{0.02 Z_1 - 0.51 Z_2}$$

donde Z_1 y Z_2 son edad y sexo, respectivamente

Funciones de supervivencia ajustadas

- Una idea popular es usar una medida de tendencia central de la variable de ajuste (Ejemplo: Media).

$$
\begin{aligned}
\hat{S}(t_i \mid Hombre, edad = 62.5) &= \hat{S}_0(t_i)^{\exp(0.02\times 62.5)} \\
\hat{S}(t_i \mid Mujer, edad = 62.5) &= \hat{S}_0(t_i)^{\exp(-0.60+0.02\times 62.5)}
\end{aligned}
$$

- Estas dos curvas describen la experiencia de supervivencia, estimada por el modelo de Cox, para dos cohortes de sujetos de 63 años de edad (hombres y mujeres)

- En general para cualquier edad, la proporcionalidad entre las curvas se mantiene pero cada curva se reajusta en base a la edad.

Ejemplo: Cancer

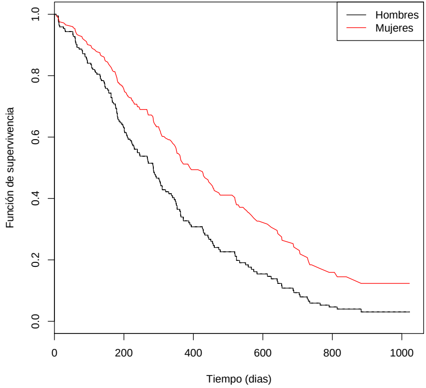

Figure : Función de supervivencia en base al modelo de Cox para una cohorte de hombres (negro) y mujeres (rojo) de 63 años de edad

Ejemplo: Recaida en drogas

- Modelo multivariado

```
> model4 <- coxph(Surv(time,censor)~age+beck+race+treat+site,data=uis)
> summary(model4)
Call:
coxph(formula = Surv(time, censor) ~ age + beck + race + treat +
    site, data = uis)

  n= 585, number of events= 471
   (43 observations deleted due to missingness)

           coef  exp(coef)  se(coef)       z  Pr(>|z|)
age   -0.013638  0.986454  0.007506 -1.817  0.06921 .
beck   0.010255  1.010307  0.004831  2.123  0.03377 *
race  -0.309739  0.733638  0.111168 -2.786  0.00533 **
treat -0.241273  0.785627  0.092667 -2.604  0.00922 **
site  -0.162555  0.849969  0.103407 -1.572  0.11595

      exp(coef) exp(-coef) lower .95 upper .95
age      0.9865     1.0137    0.9720    1.0011
beck     1.0103     0.9898    1.0008    1.0199
race     0.7336     1.3631    0.5900    0.9122
treat    0.7856     1.2729    0.6551    0.9421
site     0.8500     1.1765    0.6940    1.0409
```

- ¿Cómo presentamos graficamente estos resultados?

Score de riesgo

- En un modelo con k covariables, el *score de riesgo* es definido por

$$
\begin{aligned}
\hat{r}_{\hat{\beta}}(\mathbf{Z}) &= \hat{\beta}_1 Z_1 + \hat{\beta}_2 Z_2 + \cdots + \hat{\beta}_k Z_k \\
&= \sum_{l=1}^{k} \hat{\beta}_l Z_l
\end{aligned}
$$

- El *score de riesgo* para el sujeto i-ésimo es

$$
\hat{r}_j = \sum_{l=1}^{k} \hat{\beta}_l Z_{li}
$$

que mide como la combinación de las covariables de una persona con estas caracteristicas, afectan su riesgo.

Score de riesgo

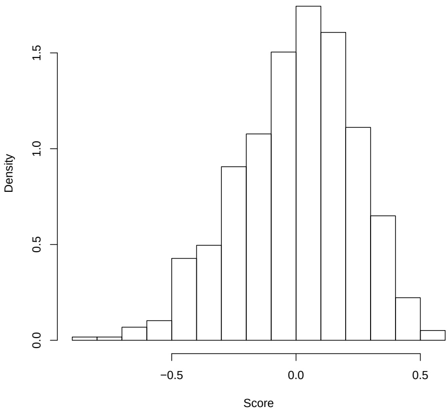

Figure : Distribución del score de riesgo

Score de riesgo

- Función de supervivencia en base al score de riesgo

$$\hat{S}(t, \hat{r}_q) = \hat{S}_0(t)^{\exp(\hat{r}_q)}$$

  donde $\hat{r}_q$ corresponde a el q-ésimo cuantil empírico de los scores de riesgo.

- En el caso que quisieramos estudiar el efecto de una variable ajustando por el score de riesgo podemos definir

$$\hat{r}_{jl} = \hat{r}_j - \hat{\beta}_l Z_{jl}$$

Ejemplo: Recaida en drogas

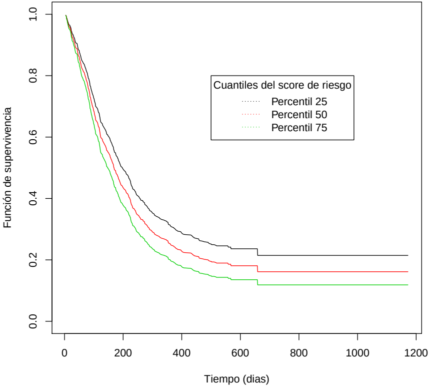

Figure : Función estimada de supervivencia en base a el score de riesgo

Ejemplo: Recaida en drogas

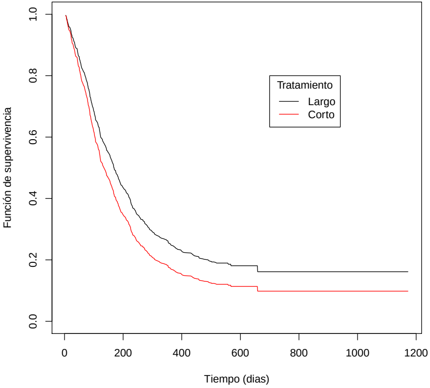

Figure : Función estimada de supervivencia para los dos tipos de tratamiento ajustando por el score de riesgo