Modelo de riesgos proporcionales

Giancarlo Sal y Rosas

Departmento de Ciencias
Pontificia Universidad Católica del Perú

May 10, 2017

Outline

Estudio: Prevención de abuso de drogas

- Estudio randomizado desarrollado por la Universidad de Massachusetts entre 1989 y 1994.
- Pacientes eran aleatorizados a recibir un tratamiento corto o prolongado para la prevención de recaida en abuso de drogas.
- Los datos presentados corresponden a 612 de estos pacientes.
- Se midieron varias variables en el momento del enrolamiento: Edad, tipo de adicción, depresión entre otras

Estudio: Prevención de abuso de drogas

Entre las variables medidas tenemos:

a) *id*: Número de identificación

b) *age*: Edad en el momento de enrolamiento en años

c) *beck*: Score de BECK para medir depresión en el enrolamiento: 0 - 54

d) *ivhx*: Historia de uso de drogas en admisión
   - Nunca (1)
   - Antiguo (2)
   - Reciente (3)

e) *ndrugtx*: Número de tratamiento de drogas previo: 0 - 40

f) *race*: Raza del paciente: Blanco (1) y otros (0)

Estudio: Prevención de abuso de drogas

c) *hercoc*: Uso de heroina/cocaina en los 3 meses previous a adminisión
   - Heroina y cocaina (1)
   - Solo heroina (2)
   - Solo cocaina (3)
   - Ni cocaina ni heroina (4)

g) *treat*: Tratamiento (randomizado): Corto (0) y largo (1)

h) *site*: Lugar de tratamiento: A (0) y B (1)

i) *los*: Tiempo de tratamiento (dias)

j) *time*: Tiempo a recaida en drogas (dias)

k) *censor*: Retorno a drogas: Si (1) y No (0)

Estudio: Prevención de abuso de drogas

```
> uis[1:20,]
   id age  beck hercoc pdrug ndrugtx race treat site los time censor
1   1  39  9.00      4     3       1    0     1    0 123  188      1
2   2  33 34.00      4     2       8    0     1    0  25   26      1
3   3  33 10.00      2     3       3    0     1    0   7  207      1
4   4  32 20.00      4     3       1    0     0    0  66  144      1
5   5  24  5.00      2     1       5    1     1    0 173  551      0
6   6  30 32.55      3     3       1    0     1    0  16   32      1
7   7  39 19.00      4     3      34    0     1    0 179  459      1
8   8  27 10.00      4     3       2    0     1    0  21   22      1
9   9  40 29.00      2     3       3    0     1    0 176  210      1
10 10  36 25.00      2     3       7    0     1    0 124  184      1
11 11  35    NA     NA    NA      12    1     1    0   2    5      1
12 12  38 18.90      2     3       8    0     1    0 176  212      1
13 13  29 16.00      3     1       1    0     1    0  79   87      1
14 14  32 36.00      3     3       2    1     1    0 182  598      0
15 15  41 19.00      1     3       8    0     1    0 174  260      1
16 16  31 18.00      1     3       1    0     1    0 181  210      1
17 17  27 12.00      2     3       3    0     1    0  61   84      1
18 18  28 34.00      1     3       6    0     1    0 177  196      1
19 19  28 23.00      4     2       1    0     1    0  19   19      1
20 20  36 26.00      3     1      15    1     1    0  27  441      1
```

Estudio: Prevención de abuso de drogas

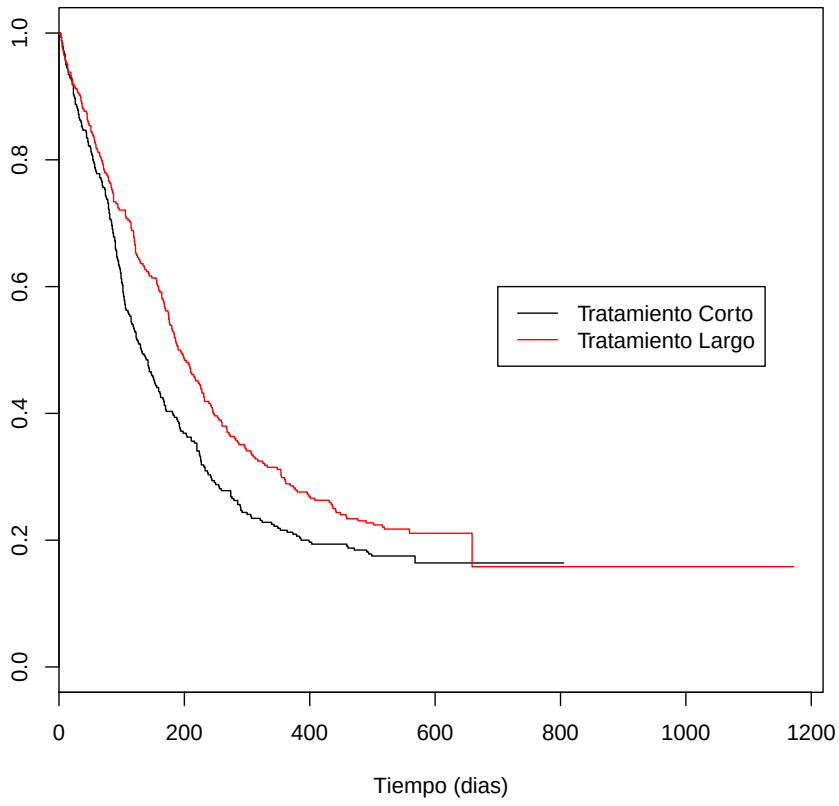


Figure : Función de supervivencia estimada para ambos grupos

Estudio: Prevención de abuso de drogas

- Codigo de R

```
> summary(model1)
Call:
coxph(formula = Surv(time, censor) ~ treat, data = uis)

  n= 628, number of events= 508

          coef exp(coef) se(coef)      z Pr(>|z|)
treat -0.23163   0.79324  0.08899 -2.603  0.00925 **

      exp(coef) exp(-coef) lower .95 upper .95
treat    0.7932      1.261    0.6663    0.9444
```

- Interpretación
  - El tratamiento prolongado reduce el riesgo de recaida en drogas en un 21% (HR=0.79, 95%IC: 0.66-0.94)
  - Existen factores que mejoran (disminuyen) el efecto del tratamiento ?

Estudio: Prevención de abuso de drogas

- Codigo de R

```
> summary(model1)
Call:
coxph(formula = Surv(time, censor) ~ treat, data = uis)

  n= 628, number of events= 508

          coef exp(coef) se(coef)      z Pr(>|z|)
treat -0.23163   0.79324  0.08899 -2.603  0.00925 **

      exp(coef) exp(-coef) lower .95 upper .95
treat    0.7932      1.261    0.6663    0.9444
```

- Interpretación
  - El tratamiento prolongado reduce el riesgo de recaida en drogas en un 21% (HR=0.79, 95%IC: 0.66-0.94)
  - Existen factores que mejoran (disminuyen) el efecto del tratamiento ?

Introducción

- En la actualidad es facil constuir un modelo
- Escoger un modelo adecuado es un problema muy dificil

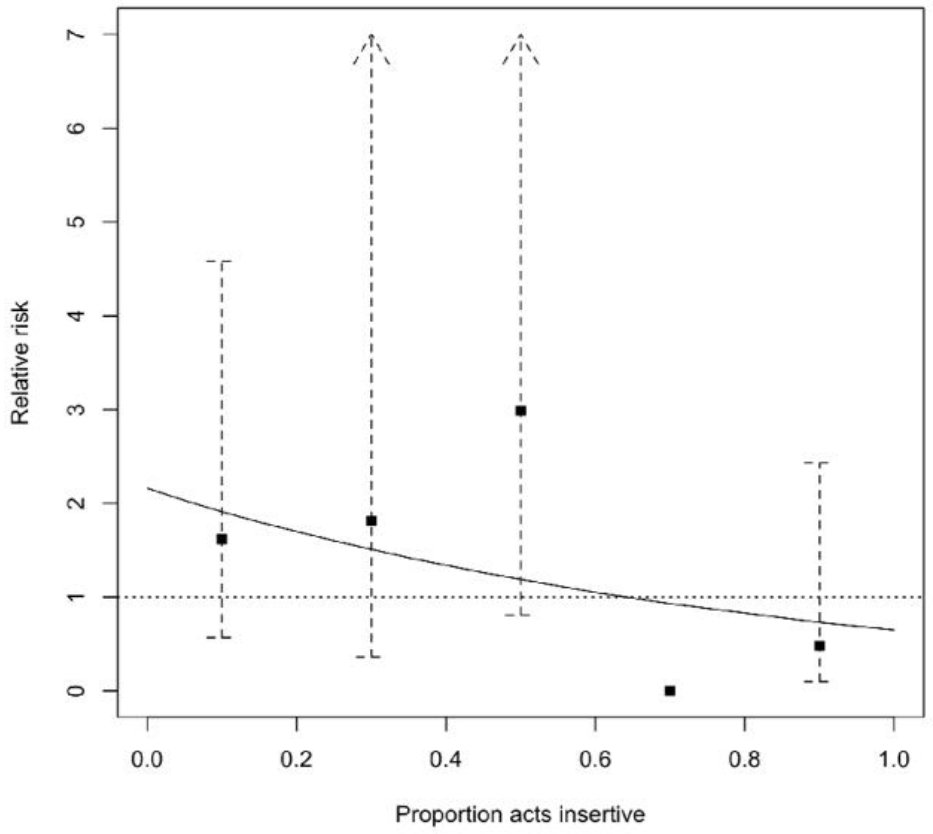
Supongamos que medimos las variables $\mathbf{X} = (X_1, \ldots, X_p)$ y el tiempo T a la ocurrencia del evento de interes.

- Nos vamos a enfrentar con dos escenarios:
  - Estudio de confirmación: Deseamos determinar la asociación entre una variable X_k y T
  - Estudio exploratorio o predictivo: Deseamos estudiar la relación entre $\underline{X} = (X_1, \ldots, X_p)$ y T

Analisis confirmatorio

Sánchez et al. (2011) analizo si la circumcisión podría ser un factor de protección contra la infección de VIH

Figure : Riesgo de acquisión de VIH asociado a circuncisión bajo dos modelos



Introducción

- Muchas personas que usan estos modelos no saben distinguir la diferencia entre estos dos escenarios.

> Es su trabajo explicarles la diferencia.

- Existen varios metodos para seleccionar variables
  - Selección dirigida
  - Metodo Lasso
  - Procedimiento Stepwise

Selección dirigida

- *Paso 0*: En base a su entendimiento del problema, seleccionar el conjunto de variables candidatos a ser incluidas en el modelo.
- *Paso 1*: Seleciona todas las variables
  - Significativas en el analisis univariado (al 20-25% de confianza)
  - Aquellas cientificamente importantes.

Desarrollo del modelo: Selección dirigida

- *Paso 2*: Implementar un modelo multivariado con las variables consideradas y en base a la prueba de Wald considera las variables a ser removidas. La remoción es confirmada por
  - El estadistico del cociente de funciones de verosimilitud
  - Su remoción genera cambios en los coeficientes del modelo menores al 20%
- *Paso 3*: Considerar todas las variablas no consideradas en el *Paso 1* para confirmar que no son importantes.

Desarrollo del modelo: Selección dirigida

- *Paso 4*: Analizar la escala de las variables continuas
  - Categorización de la variable continua
  - Asumir un modelo polinomico para estas.
- *Paso 5*: Analizar la necesidad de posibles iteracciones.

Al final de este paso tenemos un modelo preliminar

Estudio: Prevención de abuso de drogas

Paso 1

```
> coxph(Surv(time, censor)~age,data=uis)
        coef exp(coef) se(coef)     z     p
age -0.01288   0.98720  0.00719 -1.79 0.073

> coxph(Surv(time, censor)~beck,data=uis)
        coef exp(coef) se(coef)    z    p
beck 0.01098   1.01104  0.00471 2.33 0.02

> coxph(Surv(time, censor)~factor(hercoc),data=uis)
                   coef exp(coef) se(coef)     z     p
factor(hercoc)2  0.0776    1.0807   0.1445  0.54 0.591
factor(hercoc)3 -0.2549    0.7750   0.1351 -1.89 0.059
factor(hercoc)4 -0.1622    0.8503   0.1302 -1.25 0.213

> coxph(Surv(time, censor)~factor(ivhx),data=uis)
                coef exp(coef) se(coef)    z       p
factor(ivhx)2 0.196     1.216    0.129 1.52 0.12829
factor(ivhx)3 0.386     1.471    0.101 3.81 0.00014

> coxph(Surv(time, censor)~ndrugtx,data=uis)
           coef exp(coef) se(coef)    z       p
ndrugtx 0.0294    1.0299   0.0075 3.92 8.7e-05
```

Estudio: Prevención de abuso de drogas

Paso 1

```
> coxph(Surv(time, censor)~race,data=uis)
       coef exp(coef) se(coef)     z      p
race -0.285     0.752    0.106 -2.69 0.0072

> coxph(Surv(time, censor)~treat,data=uis)
        coef exp(coef) se(coef)    z      p
treat -0.232     0.793    0.089 -2.6 0.0092

> coxph(Surv(time, censor)~site,data=uis)
        coef exp(coef) se(coef)     z    p
site -0.1516    0.8593   0.0986 -1.54 0.12

> coxph(Surv(time, censor)~los,data=uis)
         coef exp(coef)  se(coef)     z       p
los -0.008365  0.991670  0.000764 -10.9 <2e-16
```

- La unica que que luce no importante es la variable *hercoc*

Estudio: Prevención de abuso de drogas

- Modelo multivariado

```
> model2 <- coxph(Surv(time,censor)~age + beck + factor(hercoc) + factor(ivhx) + ndrugt
+           + race + treat + site,data=uis)
> summary(model2)

  n= 575, number of events= 464
   (53 observations deleted due to missingness)

                     coef exp(coef)  se(coef)      z Pr(>|z|)
age             -0.028919  0.971496  0.008171 -3.539 0.000401 ***
beck             0.008361  1.008396  0.004976  1.680 0.092866 .
factor(hercoc)2  0.065278  1.067456  0.150009  0.435 0.663446
factor(hercoc)3 -0.094206  0.910096  0.165472 -0.569 0.569142
factor(hercoc)4  0.027523  1.027906  0.160275  0.172 0.863653
factor(ivhx)2    0.174657  1.190837  0.138635  1.260 0.207729
factor(ivhx)3    0.280944  1.324379  0.146933  1.912 0.055870 .
ndrugtx          0.028443  1.028851  0.008307  3.424 0.000617 ***
race            -0.203049  0.816238  0.116696 -1.740 0.081862 .
treat           -0.240484  0.786248  0.094371 -2.548 0.010825 *
site            -0.103029  0.902100  0.109279 -0.943 0.345779
```

- Estudiaremos si *hercoc* puede ser retirada
- *site* es de importancia cientifica asi obtamos por mantenerla en el modelo.

Estudio: Tratamiento de prevención de abuso de drogas

- Usando el cociente de funciónes de verosimilitud podemos estudiar si mantenemos a **hercoc** en el modelo o no
- Codigo de R

```
> model2      <- coxph(Surv(time,censor)~age + beck + factor(ivhx) +
  ndrugtx + race + treat + site,data=uis)
> pvalue <- 1 - pchisq(2*(model1$loglik[2] - model2$loglik[2]),3)
> pvalue
[1] 0.7055953
```

- La prueba estadística concluye que no hay evidencia para pensar que los dos modelos (con *hercoc* vs. no *hercoc*) son diferentes

Estudio: Prevención de abuso de drogas

- Modelo en R

```
> summary(model2)

  n= 575, number of events= 464
   (53 observations deleted due to missingness)

                    coef exp(coef)  se(coef)      z Pr(>|z|)
age            -0.028262  0.972134  0.008169 -3.460 0.000541 ***
beck            0.007957  1.007989  0.004967  1.602 0.109177
factor(ivhx)2   0.196279  1.216867  0.137212  1.430 0.152580
factor(ivhx)3   0.333371  1.395665  0.119909  2.780 0.005432 **
ndrugtx         0.027830  1.028221  0.008287  3.358 0.000784 ***
race           -0.209428  0.811048  0.115901 -1.807 0.070770 .
treat          -0.232313  0.792698  0.093712 -2.479 0.013175 *
site           -0.099996  0.904841  0.108551 -0.921 0.356952
```

- Este modelo contiene todas las variables que son importantes.
- Nuestro siguiente paso es estudiar la estructura en que estas variables esta presentadas.

Estudio: Prevención de abuso de drogas

- Notemos que la variable **ivhx** tiene dos coeficientes que miden su efecto (no vs. pasado y no vs. reciente) y solo el segunda es significativo.
- Una solución es pensar en dos categorias en lugar de tres: **no reciente vs. reciente**

```
> uis$nivhx <- ifelse(is.na(uis$ivhx),NA,ifelse(uis$ivhx < 3,0,1))
> model3    <- coxph(Surv(time,censor)~age + beck + nivhx + ndrugtx +
                       race + treat + site,data=uis)
> pvalue    <- 1 - pchisq(2*(model2$loglik[2] - model3$loglik[2]),1)
> pvalue
[1] 0.1563473
```

- La prueba estadística nos dice que ell nuevo modelo con dos categorias es al menos tan bueno como el que tiene tres categorias.

Estudio: Prevención de abuso de drogas

- Nuestro modelo multivariado hasta ahora luce asi

```
> summary(model3)

  n= 575, number of events= 464
    (53 observations deleted due to missingness)

             coef exp(coef)  se(coef)      z Pr(>|z|)
age     -0.026196  0.974145  0.008048 -3.255 0.001135 **
beck     0.008417  1.008453  0.004952  1.700 0.089157 .
nivhx    0.256599  1.292527  0.106301  2.414 0.015783 *
ndrugtx  0.029143  1.029572  0.008213  3.548 0.000388 ***
race    -0.224670  0.798779  0.115272 -1.949 0.051290 .
treat   -0.232971  0.792177  0.093732 -2.485 0.012937 *
site    -0.087180  0.916512  0.107871 -0.808 0.418982
```

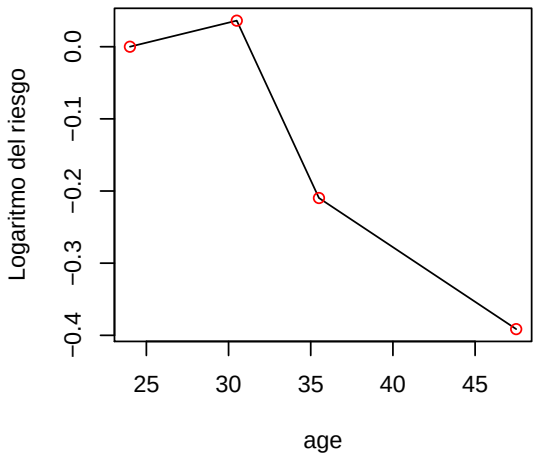
- El siguiente paso es ver si *age*, *beck* y *ndrugtx* deben permanecer como funcionales lineales.
- ¿ Alguna forma no lineal para alguna de estas variables podria ser mejor para el modelo ?

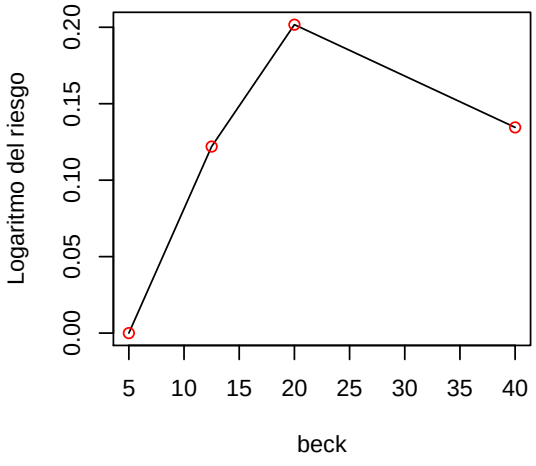
Estudio: Prevención de abuso de drogas

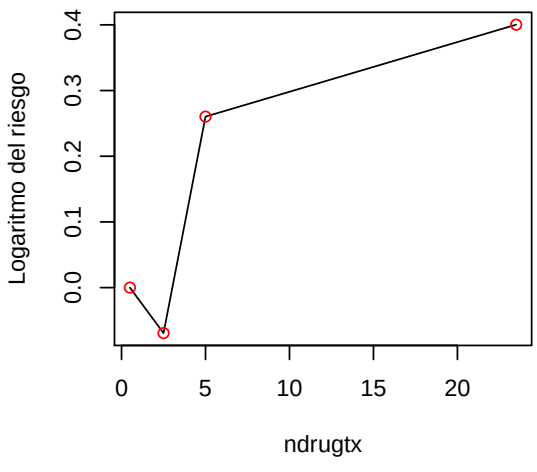
```
> uis$cage     <- cut(uis$age,breaks=quantile(uis$age,probs=c(0,0.25,0.5,0.75,1),
na.rm=TRUE),include.lowest=TRUE)
> uis$cbeck    <- cut(uis$beck,breaks=quantile(uis$beck,probs=c(0,0.25,0.5,0.75,1),
na.rm=TRUE),include.lowest=TRUE)
> uis$cndrugtx <- cut(uis$ndrugtx,breaks=quantile(uis$ndrugtx,probs=c(0,0.25,0.5,0.75,1),
na.rm=TRUE),include.lowest=TRUE)
>
> model4 <- coxph( Surv(time, censor)~factor(cage)+beck+ndrugtx+nivhx+race+treat+
site, data=uis)
> model5 <- coxph( Surv(time, censor)~age+factor(cbeck)+ndrugtx+nivhx+race+treat+
site, data=uis)
> model6 <- coxph( Surv(time, censor)~age+beck+factor(cndrugtx)+nivhx+race+treat+
site, data=uis)
>
> cbind(c(24,30.5,35.5,47.5),c(0,as.numeric(summary(model4)$coef[1:3,1])))
     [,1]       [,2]
[1,] 24.0  0.0000000
[2,] 30.5  0.0361321
[3,] 35.5 -0.2096805
[4,] 47.5 -0.3913728
```

- Categorizamos las variabes y vemos si se cumple un comportamiento lineal

Estudio: Prevención de abuso de drogas







Estudio: Prevención de abuso de drogas

La libraria *mfp* permite modelar los factores usando polinomios fraccionados

```
> library(mfp)
> # Tenemos que transformar las variables para que no tomen valores 0
> uis$fdrugtx  <- 10/(uis$ndrugtx+1) # Esta es una de las transformaciones mas usadas
> m8 <- mfp( Surv(time, censor)~ fp(fdrugtx,df=2) + beck+ age +nivhx+race+treat
+site, data=uis,family=cox)
> summary(m8)
                          coef exp(coef)  se(coef)      z Pr(>|z|)
age                 -0.027171  0.973195  0.008091 -3.358 0.000785 ***
nivhx                0.245779  1.278617  0.106569  2.306 0.021094 *
treat               -0.229330  0.795066  0.093663 -2.448 0.014347 *
race                -0.225719  0.797943  0.115295 -1.958 0.050260 .
beck                 0.008358  1.008393  0.004950  1.688 0.091320 .
I((fdrugtx/10)^-0.5) 0.179495  1.196613  0.051253  3.502 0.000462 ***
site                -0.079109  0.923939  0.107746 -0.734 0.462814
```

En el caso $m = 1$, el modelo selecionado sigue siendo el modelo lineal.

Estudio: Prevención de abuso de drogas

Si probamos el caso de $m = 2$

```
> library(mfp)
> # Tenemos que transformar las variables para que no tomen valores 0
> uis$fdrugtx  <- 10/(uis$ndrugtx+1) # Esta es una de las transformaciones mas usadas
> m8 <- mfp( Surv(time, censor)~ fp(fdrugtx,df=4) + beck+ age +nivhx+race+treat
+site, data=uis,family=cox)
> summary(m8)
                                          coef exp(coef)  se(coef)      z Pr(>|z|)
age                                  -0.028202  0.972192  0.008132 -3.468 0.000524 ***
nivhx                                 0.259107  1.295772  0.108023  2.399 0.016457 *
treat                                -0.211375  0.809470  0.093688 -2.256 0.024061 *
race                                 -0.242416  0.784730  0.115472 -2.099 0.035786 *
beck                                  0.009181  1.009223  0.004987  1.841 0.065633 .
I((fdrugtx/10)^1)                    -0.744773  0.474842  0.212169 -3.510 0.000448 ***
I((fdrugtx/10)^1 * log((fdrugtx/10))  1.950495  7.032169  0.482460  4.043 5.28e-05 ***
site                                 -0.105826  0.899581  0.109163 -0.969 0.332331
```

Aqui si se nos presenta un modelo que describe mejor el rol de **drugtx** en el modelo

Estudio: Prevención de abuso de drogas

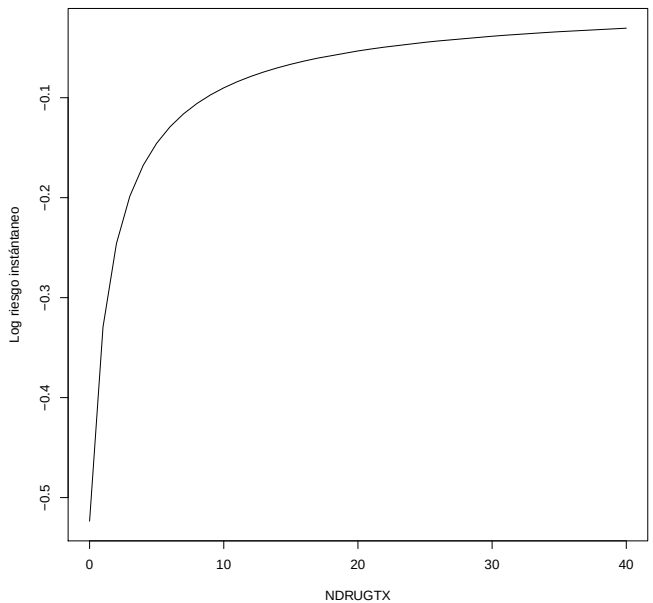
Podemos comparar el nuevo modelo con el modelo lineal usando el cociente de funciones de verosimilitud

Comparación	χ^2	gd	Valor de p
Mejor PF2 vs. modelo nulo		4	< 0.001
Mejor PF2 vs. modelo lineal		3	< 0.001
Mejor PF2 vs. PF1		2	< 0.001

Podemos concluir que el modelo explica mejor los datos

Estudio: Prevención de abuso de drogas

El efecto de la transformación de *ndrugtx* es



Por ende luce un poco mejor que el modelo lineal

Estudio: Prevención de abuso de drogas

El modelo final de efectos principales (Paso 5) es:

	coef	exp(coef)	se(coef)	z	p
age.1	-0.03	0.97	0.01	-3.47	0.00
nivhx.1	0.26	1.30	0.11	2.40	0.02
treat.1	-0.21	0.81	0.09	-2.26	0.02
race.1	-0.24	0.78	0.12	-2.10	0.04
beck.1	0.01	1.01	0.00	1.84	0.07
fdrugtx.1	-0.52	0.59	0.12	-4.21	0.00
fdrugtx.2	0.20	1.22	0.05	4.04	0.00
site.1	-0.11	0.90	0.11	-0.97	0.33

Nuestro siguiente paso es estudiar las interacciones

Estudio: Prevención de abuso de drogas

El modelo final de efectos principales (Paso 5) es:

	coef	exp(coef)	se(coef)	z	p
age.1	-0.03	0.97	0.01	-3.47	0.00
nivhx.1	0.26	1.30	0.11	2.40	0.02
treat.1	-0.21	0.81	0.09	-2.26	0.02
race.1	-0.24	0.78	0.12	-2.10	0.04
beck.1	0.01	1.01	0.00	1.84	0.07
fdrugtx.1	-0.52	0.59	0.12	-4.21	0.00
fdrugtx.2	0.20	1.22	0.05	4.04	0.00
site.1	-0.11	0.90	0.11	-0.97	0.33

Nuestro siguiente paso es estudiar las interacciones

Estudio: Tratamiento de prevención de abuso de drogas

- Interacciones es un tema complicado de explicar a alguien que no es estadistico!
- Mantenganse alejados de estas si es posible!
- En nuestro caso las interacciones posibles son:
  - Edad, raza, beck, nivhx, ndrugtx.i con site
  - ndrugtx.i y beck con edad
  - ndrugtx.i y beck con raza

Estudio: Tratamiento de prevención de abuso de drogas

- Interacciones es un tema complicado de explicar a alguien que no es estadistico!
- Mantenganse alejados de estas si es posible!
- En nuestro caso las interacciones posibles son:
  - Edad, raza, beck, nivhx, ndrugtx.i con site
  - ndrugtx.i y beck con edad
  - ndrugtx.i y beck con raza

Estudio: Tratamiento de prevención de abuso de drogas

- Interacciones es un tema complicado de explicar a alguien que no es estadistico!
- Mantenganse alejados de estas si es posible!
- En nuestro caso las interacciones posibles son:
  - Edad, raza, beck, nivhx, ndrugtx.i con site
  - ndrugtx.i y beck con edad
  - ndrugtx.i y beck con raza

Estudio: Tratamiento de prevención de abuso de drogas

- Interacciones es un tema complicado de explicar a alguien que no es estadistico!
- Mantenganse alejados de estas si es posible!
- En nuestro caso las interacciones posibles son:
  - Edad, raza, beck, nivhx, ndrugtx.i con site
  - ndrugtx.i y beck con edad
  - ndrugtx.i y beck con raza

Estudio: Tratamiento de prevención de abuso de drogas

- Interacciones es un tema complicado de explicar a alguien que no es estadistico!
- Mantenganse alejados de estas si es posible!
- En nuestro caso las interacciones posibles son:
  - Edad, raza, beck, nivhx, ndrugtx.i con site
  - ndrugtx.i y beck con edad
  - ndrugtx.i y beck con raza

Estudio: Tratamiento de prevención de abuso de drogas

- Interacciones es un tema complicado de explicar a alguien que no es estadistico!
- Mantenganse alejados de estas si es posible!
- En nuestro caso las interacciones posibles son:
  - Edad, raza, beck, nivhx, ndrugtx.i con site
  - ndrugtx.i y beck con edad
  - ndrugtx.i y beck con raza

Estudio: Tratamiento de prevención de abuso de drogas

Table : Modelo de regresión multivariado con interacciones

	coef	exp(coef)	se(coef)	z	p
age	-0.06	0.95	0.03	-1.79	0.07
beck	0.01	1.01	0.03	0.32	0.75
ndrugfp1	-0.81	0.45	0.66	-1.23	0.22
ndrugfp2	-0.22	0.80	0.26	-0.84	0.40
nivhx	0.16	1.18	0.13	1.30	0.20
race	-0.49	0.61	0.88	-0.55	0.58
treat	-0.25	0.78	0.10	-2.60	0.01
site	-1.53	0.22	0.76	-2.01	0.04
agesite	0.02	1.02	0.02	1.34	0.18
becksite	-0.00	1.00	0.01	-0.30	0.77
ndrugfp1site	0.29	1.34	0.27	1.08	0.28
ndrugfp2site	0.11	1.11	0.10	1.02	0.31
nivhxsite	0.29	1.34	0.25	1.16	0.25
racesite	0.91	2.48	0.26	3.44	0.00
beckage	0.00	1.00	0.00	0.10	0.92
ndrugfp1age	0.00	1.00	0.02	0.07	0.95
ndrugfp2age	-0.00	1.00	0.01	-0.30	0.76
raceage	-0.01	0.99	0.02	-0.38	0.70
ndrugfp1race	0.19	1.21	0.32	0.60	0.55
ndrugfp2race	0.08	1.09	0.12	0.67	0.50

Estudio: Tratamiento de prevención de abuso de drogas

Table : Modelo multivariado final

	coef	exp(coef)	se(coef)	z	p
age	-0.04	0.96	0.01	-4.18	0.00
beck	0.01	1.01	0.00	1.76	0.08
ndrugfp1	-0.57	0.56	0.13	-4.59	0.00
ndrugfp2	-0.21	0.81	0.05	-4.42	0.00
nivhx	0.23	1.26	0.11	2.10	0.04
race	-0.47	0.63	0.13	-3.46	0.00
treat	-0.25	0.78	0.09	-2.62	0.01
site	-1.32	0.27	0.53	-2.48	0.01
agesite	0.03	1.03	0.02	2.01	0.04
racesite	0.85	2.34	0.25	3.43	0.00

Un tratamiento prolongado reduce el riesgo de recaida en un 22% (HR=0.78, 95%IC: 0.65 - 0.94) en comparación con un tratamiento corto despues de controlar por edad, depresión historia de abuso de drogas, entre otros.

Estudio: Tratamiento de prevención de abuso de drogas

Table : Modelo multivariado final

	coef	exp(coef)	se(coef)	z	p
age	-0.04	0.96	0.01	-4.18	0.00
beck	0.01	1.01	0.00	1.76	0.08
ndrugfp1	-0.57	0.56	0.13	-4.59	0.00
ndrugfp2	-0.21	0.81	0.05	-4.42	0.00
nivhx	0.23	1.26	0.11	2.10	0.04
race	-0.47	0.63	0.13	-3.46	0.00
treat	-0.25	0.78	0.09	-2.62	0.01
site	-1.32	0.27	0.53	-2.48	0.01
agesite	0.03	1.03	0.02	2.01	0.04
racesite	0.85	2.34	0.25	3.43	0.00

Un tratamiento prolongado reduce el riesgo de recaida en un 22% (HR=0.78, 95%IC: 0.65 - 0.94) en comparación con un tratamiento corto despues de controlar por edad, depresión historia de abuso de drogas, entre otros.

Modelo

- Data disponible

$$(y_1, \mathbf{x}_1, \delta_1), \ldots, (y_n, \mathbf{x}_n, \delta_n)$$

donde $\mathbf{x} = (x_1, x_2, \ldots, x_k)$ es el vector de predictores lineales

- Modelo de Cox

$$\lambda(t \mid x) = \lambda_0(t) \exp\left(\sum_j x_j \beta_j\right)$$

- La verosimilitud parcial es

$$L(\beta) = \prod_{r \in D} \frac{\exp(\beta^t x_j)}{\sum_{j \in R_r} \exp(\beta^t x_j)}$$

Modelo

- Tibshirani (1997) propuso estimar β con restricciones

$$\hat{\beta} = \arg\min_{\beta} l(\beta)$$

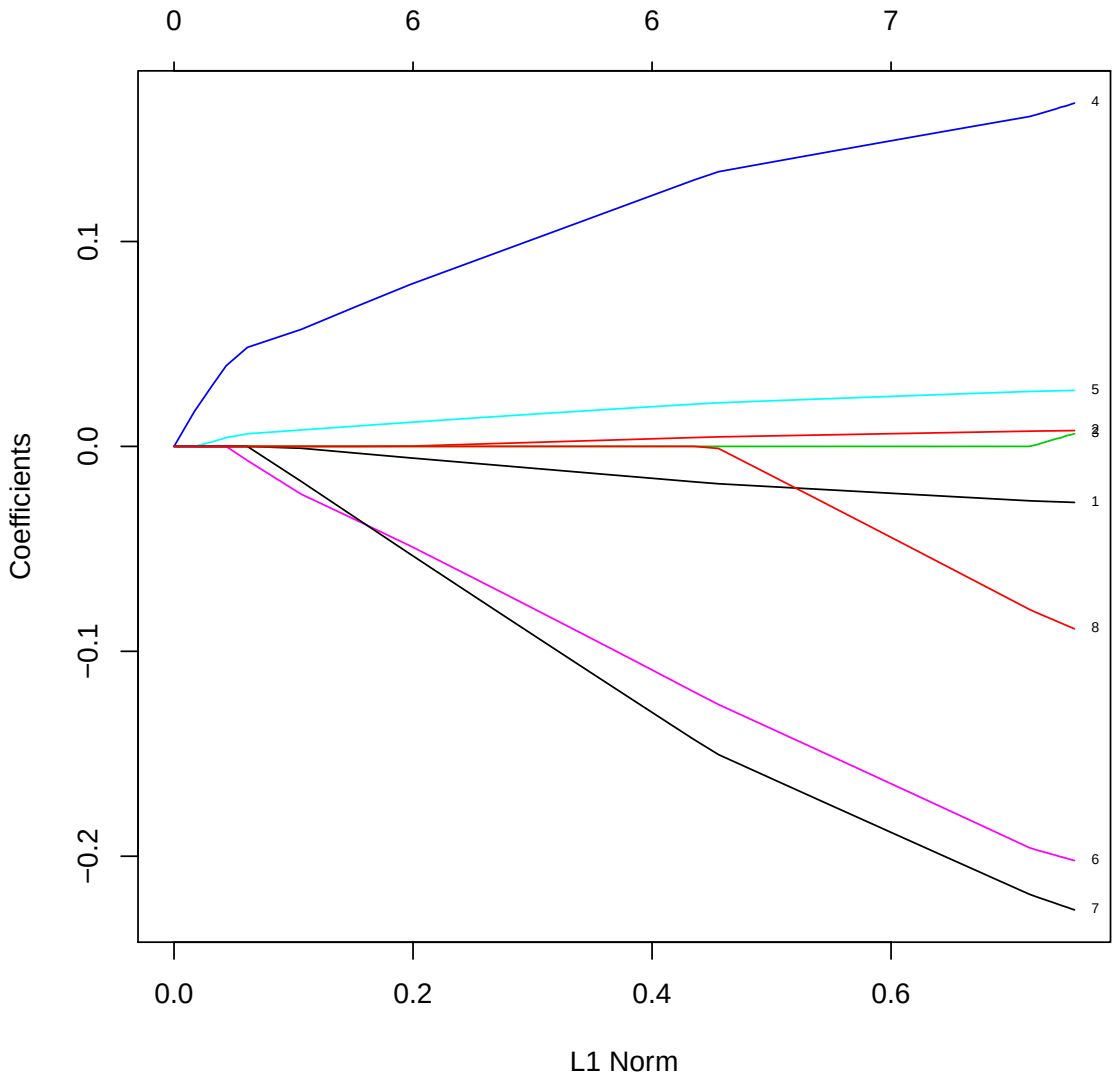
sujeto a $\sum |\beta_j| \leq s$.

- Note que en el caso de regresión lineal lo que se minimiza es la suma del cuadrado de los residuos.
- R usa una net elastica

$$\sum_{j=1}^{p} \left[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j| \right]$$

que es un compromiso entre lasso y ridge

Implementación en R

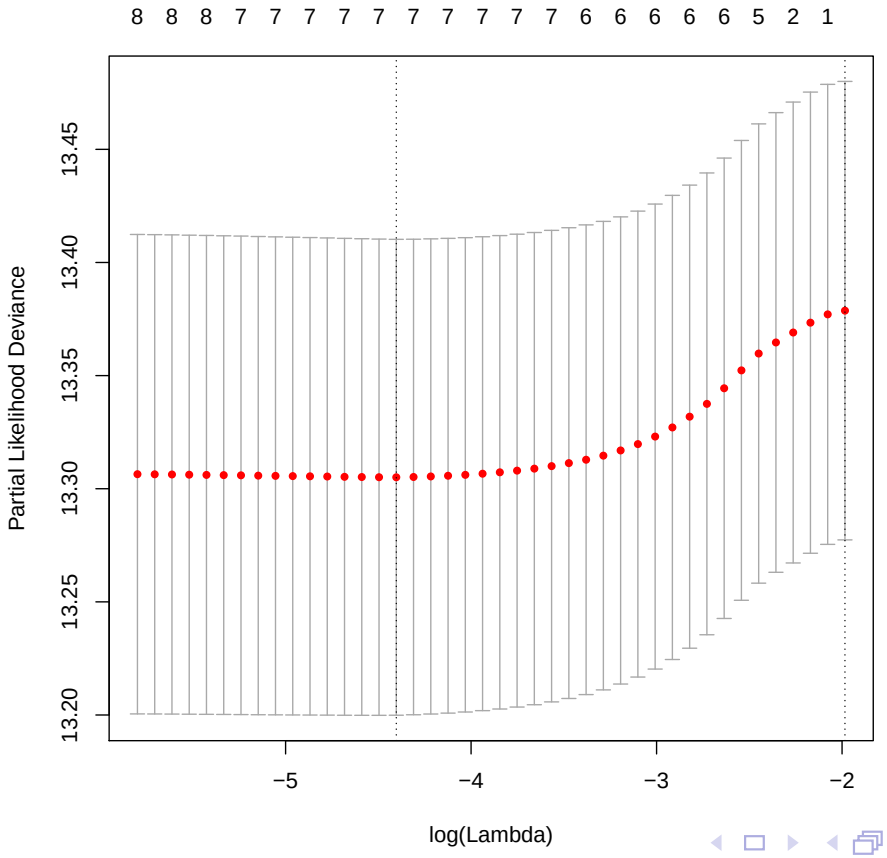


Optimo valor de λ

- Codigo de R

```
cv.fit <- cv.glmnet(x, y, family="cox", alpha=1)
plot(cv.fit)
```

- Verosimilitud parcial para diferentes valores de λ



Modelo optimo

```
> cv.fit$lambda.min
[1] 0.01223744

> cv.fit$lambda.1se
[1] 0.1374662

> coef(cv.fit, s = "lambda.min")
8 x 1 sparse Matrix of class "dgCMatrix"
                      1
age     -0.024171708
beck     0.006648725
hercoc   .
ivhx     0.153394298
ndrugtx  0.025277587
race    -0.175716611
treat   -0.199148046
site    -0.056868661
```

- *lambda.min*: Valor de λ que nos da el mínimo validación cruzada (en base a la verosimilitud parcial)
- *lambda.*1*se*: Mayor valor de λ que se encuentra dentro de $+/-$ error éstandar de el mínimo.

Algoritmo paso a paso

- Al igual que en regresión lineal, podemos usar algoritmos para seleccionar o eliminar variables de un modelo.
- **Paso a paso hacia atraz**
  - Inicia con un modelo complejo y secuencialmente remueve términos.
  - En un determinado paso, elimina el termino en el modelo que tiene el valor mas largo.
  - El proceso para cuando eliminar una variable reduce cuan bien ajusta el modelo.
- **Paso a paso hacia adelante**
  - Añade terminos secuencialmente hasta que una adicional adicional no mejora el ajuste del modelo.

Algoritmo paso a paso: Hacia adelante

- *Paso 0*:
  - Se construye el modelo sin covariables y evalua el valor que maximiza su función de verosimilitud (denotado por L_0).
  - Construya el modelo univariado para cada una de las k variables y evalue el valor que maximiza su función de verosimilitud (denotado por L_j^0 para $j = 1, 2, \ldots, k$).
  - El valor de p para las k pruebas es

$$p_j^{(0)} = P\left[\chi_v^2 > 2(L_j^{(0)} - L_0)\right] \quad , \quad j = 1, 2, \ldots, k$$

  - Seleccionamos la variable x_{e_1} si

$$p_{e_1} = \min_j(p_j^{(0)}) \quad y \quad p_{e_1} < p_E$$

  donde p_E es la probabilidad de entrada al modelo (definido por el usuario).

Algoritmo paso a paso: Hacia adelante

- *Paso 1*:
  - Construir los $k-1$ modelos que contienen: la variable x_{e_1} + las $k-1$ variables restantes
  - Calcular

$$p_j^{(1)} = P\left[\chi_v^2 > 2(L_j^{(1)} - L_{e_1})\right] \quad , \quad j = 1, \ldots, k-1$$

    donde L_{e_1} es el valor de verosimilitud para el modelo que solo incluye a x_{e_1}
  - Seleccionamos a la variable x_{e_2} si

$$p_{e_2}^{(1)} = \min_j(p_j^{(1)}) \quad y \quad p_{e_2} < p_E$$

Algoritmo paso a paso: Hacia adelante

- *Paso 2*:
  - Se construye el modelo con las variables x_{e_1} y x_{e_2}
  - Se evalua si x_{e_1} ya no es relevante una vez que hemos añadido x_{e_1} usando:

$$p^{(2)}_{-e_1} = P\left[\chi^2_v > 2(L^{(2)}_{-e_1} - L^{(2)}_{e_1 e_2})\right]$$

  - Para eliminar esa variable se debe cumplir $p^{(2)}_{-e_1} > p_R$, donde p_R es el limite permitido de nivel de significancia.
  - Para añadir una variable esta debe cumplir

$$p^{(2)}_{e_3} = \min_j(p^{(2)}_j) \quad y \quad p_{e_3} < p_E$$

donde

$$p^{(2)}_{e_j} = P\left[\chi^2_v > 2(L^{(2)}_{e_1 e_2 e_j} - L^{(2)}_{e_1 e_2})\right]$$

Algoritmo paso a paso: Hacia adelante

- *Paso 3*,. . . , k: Son iguales al paso igual al *Paso 2*.
- *Paso final*: El proceso termina en cualquiera de estos dos casos
  - Todas las variables han entrado al modelo
  - Las que están en el modelo tienen un valor de $p < p_R$ y las que no lo están tiene $p > P_E$

Algoritmo paso a paso: Hacia adelante

- No podemos asegurarnos que capturemos ruido en lugar de señal.
- Se recomienda su uso en caso de análisis exploratorio o para construir modelos de predicción.
- La función *selectCox* en la libreria *pec* nos permite implementar este algoritmo.
- Al igual que en la selección metódica, tenemos que especificar la forma de las variables y las posibles iteracciones

Implementación en R

```
> fit2 <- selectCox(Surv(time,censor)~age+beck+hercoc+nivhx+
+                   ndrugtx+race+treat+site+site*race+age*site,data=uis,rule="aic")
> fit2$fit
Cox Proportional Hazards Model

 rms::cph(formula = newform, data = data, surv = TRUE)

                      Model Tests       Discrimination
                                           Indexes
 Obs         575    LR chi2     56.04    R2        0.093
 Events      464    d.f.            8    Dxy       0.215
 Center -1.2837    Pr(> chi2) 0.0000    g         0.403
                   Score chi2  57.76    gr        1.496
                   Pr(> chi2) 0.0000

              Coef    S.E.   Wald Z Pr(>|Z|)
 age          -0.0389 0.0098 -3.95  <0.0001
 nivhx         0.2449 0.1061  2.31  0.0209
 ndrugtx       0.0311 0.0083  3.77  0.0002
 race         -0.4267 0.1344 -3.17  0.0015
 treat        -0.2663 0.0945 -2.82  0.0048
 site         -1.2136 0.5287 -2.30  0.0217
 race * site   0.8118 0.2470  3.29  0.0010
 age * site    0.0298 0.0161  1.86  0.0634
```

- Se elimino la variable *beck* usando como criterio el estadístico AIC

Implementación en R

```
> fit3 <- selectCox(Surv(time,censor)~age+beck+hercoc+nivhx+
+                     ndrugtx+race+treat+site+site*race+age*site,data=uis,rule="p")
> fit3$fit
Cox Proportional Hazards Model

 rms::cph(formula = newform, data = data, surv = TRUE)

                      Model Tests       Discrimination
                                           Indexes
 Obs         575    LR chi2     52.62    R2        0.087
 Events      464    d.f.            7    Dxy       0.207
 Center -0.9479    Pr(> chi2) 0.0000    g         0.392
                   Score chi2  54.20    gr        1.480
                   Pr(> chi2) 0.0000

              Coef    S.E.   Wald Z Pr(>|Z|)
 age          -0.0288 0.0081 -3.55  0.0004
 nivhx         0.2517 0.1065  2.36  0.0181
 ndrugtx       0.0299 0.0082  3.64  0.0003
 race         -0.4204 0.1344 -3.13  0.0018
 treat        -0.2471 0.0939 -2.63  0.0085
 site         -0.2648 0.1212 -2.19  0.0289
 race * site   0.8429 0.2466  3.42  0.0006
```

- Se elimino la variable *beck* y una interacción usando como criterio el valor de *p*

Referencias I

Sánchez, J., Sal Y Rosas, V., Hughes, J., Baeten, J., Fuchs, J., Buchbinder, S., Koblin, B., Casapia, M., Ortiz, A., and Celum, C. (2011). Male circumcision and risk of hiv acquisition among msm. *AIDS*, 25(4):519–23.

Tibshirani, R. (1997). The lasso method for variable seleccion in the Cox model. *Statistics in Medicine*, 16(1):385–395.