Ejemplo modelo de Regresión Poisson y selección de variables

Los datos corresponden a características de 44 minas de carbón.
La variable respuesta es el **conteo de fracturas** en la capa superior de la mina (frac).
Cabe hacer notar que en este caso la variable tiempo NO mide exposición al riesgo, sino que es una variable explicativa más.

```
library("locfit", lib.loc="~/R/win-library/3.5")
data(mine)
###
###frac fractures in the upper seam of mines (coal fields)
###inb  inner burden thickness in feet (grosor interno de la carga)
###extrp percent extraction of the lower previously mined seam (seam=capa)
###seamh lower seam height
##time that the mine has been opened (years)
```

```
> mine
   frac inb extrp seamh time
1     2  50    70    52  1.0
2     1 230    65    42  6.0
3     0 125    70    45  1.0
4     4  75    65    68  0.5
5     1  70    65    53  0.5
6     2  65    70    46  3.0
7     0  65    60    62  1.0
8     0 350    60    54  0.5
9     4 350    90    54  0.5
10    4 160    80    38  0.0
11    1 145    65    38 10.0
12    4 145    85    38  0.0
13    1 180    70    42  2.0
14    5  43    80    40  0.0
15    2  42    85    51 12.0
16    5  42    85    51  0.0
17    5  45    85    42  0.0
18    5  83    85    48 10.0
19    0 300    65    68 10.0
20    5 190    90    84  6.0
21    1 145    90    54 12.0
22    1 510    80    57 10.0
23    3  65    75    68  5.0
24    3 470    90    90  9.0
25    2 300    80   165  9.0
26    2 275    90    40  4.0
27    0 420    50    44 17.0
28    1  65    80    48 15.0
29    5  40    75    51 15.0
30    2 900    90    48 35.0
31    3  95    88    36 20.0
32    3  40    85    57 10.0
33    3 140    90    38  7.0
34    0 150    50    44  5.0
35    0  80    60    96  5.0
36    2  80    85    96  5.0
37    0 145    65    72  9.0
38    0 100    65    72  9.0
39    3 150    80    48  3.0
40    2 150    80    48  0.0
41    3 210    75    42  2.0
42    5  11    75    42  0.0
43    0 100    65    60 25.0
44    3  50    88    60 20.0
```

Por lo mencionado arriba, aquí NO hay offset, se modela el número de fracturas directamente.

La idea de este ejercicio es ajustar varios modelos, desde un modelo chico con solo una constante hasta un modelo grande, que incluye las 4 variables explicativas.

```
> modchico<-glm(frac~ 1 ,family=poisson,data=mine)
> modgde<-glm(frac~ . ,family= poisson(link = "log"),data=mine)
```

Usando el comando step, se seleccionarán modelos, step tiene tres opciones:
forward, backward y both.

1) Backward

```
> mine.stp1<-step(modgde,
+         scope = list(lower = modchico,upper =modgde),
+         trace = TRUE,direction="backward")
Start:  AIC=144.13
frac ~ inb + extrp + seamh + time

        Df Deviance    AIC
- seamh  1   38.031 142.30
<none>       37.856 144.13
- inb    1   41.023 145.29
- time   1   41.750 146.02
- extrp  1   69.807 174.08

Step:  AIC=142.3
```

Step saca seamh, pues la diferencia

$D2 - D1 = 38.03 - 37.85 = 0.18$

tiene un p-value= 0.77=1-pchisq>(.18,1).

```
frac ~ inb + extrp + time

        Df Deviance    AIC
<none>       38.031 142.30
- inb    1   41.626 143.90
- time   1   42.094 144.37
- extrp  1   70.426 172.70
> mine.stp1

Call:  glm(formula = frac ~ inb + extrp + time, family = poisson(link = "log"),
    data = mine)

Coefficients:
(Intercept)          inb        extrp         time
  -3.720682    -0.001479     0.062701    -0.031651

Degrees of Freedom: 43 Total (i.e. Null);  40 Residual
Null Deviance:       74.98
Residual Deviance: 38.03  AIC: 142.3
```

Step, ya no saca a ninguna otra variable pues los cambios en deviance sí serían significativos.

```
2) Forward.(no se pide registro de como se va haciendo: trace=FALSE)
> mine.stp2<-step(modchico,
+                 scope = list( lower = modchico,upper =modgde),
+                 trace = FALSE,direction="forward")
> mine.stp2

Call:  glm(formula = frac ~ extrp + time + inb, family = poisson, data = mine)

Coefficients:
(Intercept)        extrp         time          inb
  -3.720682     0.062701    -0.031651    -0.001479

Degrees of Freedom: 43 Total (i.e. Null);  40 Residual
Null Deviance:       74.98
Residual Deviance: 38.03  AIC: 142.3

3)Both .(no se pide registro de como se va haciendo: trace=FALSE)
>
> mine.stp3<-step(modchico,
+                  scope = list( lower = modchico,upper =modgde),
+                  trace = FALSE,direction="both")
> mine.stp3

Call:  glm(formula = frac ~ extrp + time + inb, family = poisson, data = mine)

Coefficients:
(Intercept)        extrp         time          inb
  -3.720682     0.062701    -0.031651    -0.001479

Degrees of Freedom: 43 Total (i.e. Null);  40 Residual
Null Deviance:       74.98
Residual Deviance: 38.03  AIC: 142.3

En este caso en particular las tres opciones coinciden, pero esto no siempre es así.

Interpretación de los coeficientes, usaremos mine.step1.
> summary(mine.stp1)

Call:
glm(formula = frac ~ inb + extrp + time, family = poisson(link = "log"),
    data = mine)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

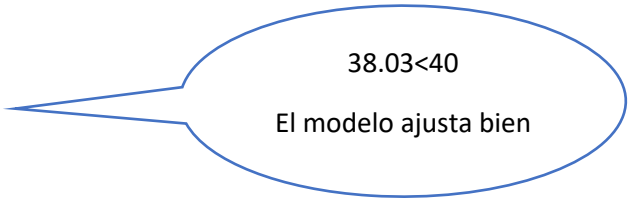
```
-1.7727  -0.9073  -0.0107   0.2716   2.1783

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -3.7206821  0.9788770  -3.801 0.000144 ***
inb         -0.0014793  0.0008244  -1.794 0.072757 .
extrp        0.0627011  0.0122711   5.110 3.23e-07 ***
time        -0.0316514  0.0163095  -1.941 0.052298 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 74.984  on 43  degrees of freedom
Residual deviance: 38.031  on 40  degrees of freedom
AIC: 142.3

Number of Fisher Scoring iterations: 5
```




```
> exp(10*mine.stp1$coefficients[2])##coef de inb

0.9853163
```

la media de las fracturas en una mina con inb=x+10 es aprox. el 98% del de una mina con inb=x
o
la media de las fracturas en una mina con inb=x+10 es poquito menor que la de una mina con inb=x

OJO, para un alfa=0.05 inb sería NO significativa.

```
> exp(10*mine.stp1$coefficients[3])##coef de extrp

1.872007
```

la media de las fracturas se incrementa en un 87% por cada incremento de 10 puntos porcentuales de extracción (extrp)

```
> exp(10*mine.stp1$coefficients[4])##coef de time

0.7286849
```

Como el coeficiente tiene signo negativo, la media de las grietas disminuye con la edad de la mina.
la media de las fracturas es aproximadamente el 70% de la media de una mina 10 años más joven.

El modelo se puede mejorar con interacciones

```
> mine.stp4<-step(modchico,scope = list( lower = modchico,upper =~inb*extrp*time  ),trace = FALSE,d
irection="forward")
> mine.stp4

Call:  glm(formula = frac ~ extrp + time + inb + extrp:inb, family = poisson,
    data = mine)

Coefficients:
(Intercept)        extrp         time          inb    extrp:inb
 -0.7891432    0.0278719   -0.0312947   -0.0240355    0.0002617

Degrees of Freedom: 43 Total (i.e. Null);  39 Residual
Null Deviance:       74.98
Residual Deviance: 33.63  AIC: 139.9
```

```
> anova(mine.stp1,mine.stp4)
Analysis of Deviance Table

Model 1: frac ~ inb + extrp + time
Model 2: frac ~ extrp + time + inb + extrp:inb
  Resid. Df Resid. Dev Df Deviance
1        40     38.031
2        39     33.635  1   4.3959
```

La diferencia de devianzas es 4.39 con 1 gl, entonces el modelo con la interacción extrp:inb es significativamente mejor que sin ella.

Observen también que el AIC del modelo mine.stp4 es menor que el de mine.stp1 (139.9< 142.3).
La interpretación de coeficientes es la que se dificulta.