Modelo Lineal Generalizado

Práctica 5

Ejercicio 1. Consideremos el conjunto de datos del archivo **postdoc** (Allison,1999). Los datos corresponden 557 hombres doctorados en bioquímica en 106 universidades norteamericanas en un lapso entre 1950 y 1960. El siguiente cuadro describe las variables consideradas.

Variable	**Nombre**
Edad a la que se doctoró	AGE
1 si está casado, 0 si no está casado	MAR
Medida del prestigio de la universidad donde hizo el Ph.D.	DOC
Medida de la selectividad de la institución de grado	UND
Número de citas a artículos publicados	CITS

Nos interesa estudiar la relación entre CITS y las demás variables.

a) Realice un histograma para la variable CITS como para detectar si la variable es asimétrica o no y si una regresión Poisson tiene sentido.

b) Realice un ajuste para CITS considerando en su modelo las variables AGE, MAR (tomando como base a los no casados), DOC, UND y constante. Analice la significación de cada una de las variables al 5%. Interprete el coeficiente de AGE.

c) Calcule los residuos de Pearson y de la deviance y haga un summary de los mismos. ¿Nota algo raro?

d) Compare la deviance y el estadístico de Pearson con sus grados de libertad tomando el cociente. Un cociente de esa magnitud, ¿qué le sugiere? ¿Le parece que el ajuste es bueno? ¿Le parece que puede haber sobredispersión?

e) Lleve a cabo un nuevo ajuste en el que contemple un coeficiente de escala distinto de 1. Analice la significación de cada una de las variables. Compare estos resultados con los obrtenidos en los items anteriores.

f) Analice los residuos de este nuevo modelo.

g) La siguiente salida corresponde al ajuste de los datos con stata usando la distribución binomial negativa. Compare la relación entre los estadísticos X^2 y D y sus grados de libertad. Analice la significación de las variables al 5%. ¿El ajuste es bueno?

   Compare con los análisis anteriores.

```
. glm cits age mar doc und, family(nbinom)

Iteration 1 : deviance = 1107.9744
Iteration 2 : deviance =  996.7248
Iteration 3 : deviance =  981.2391
Iteration 4 : deviance =  981.0663
Iteration 5 : deviance =  981.0659
Iteration 6 : deviance =  981.0659
Iteration 7 : deviance =  981.0659

Residual df  =       549                                 No. of obs =       554
Pearson X2   =  1686.264                                 Deviance   =  981.0659
Dispersion   =  3.071519                                 Dispersion =  1.787005

Negative Binomial (k=1) distribution, log link
------------------------------------------------------------------------------
    cits |      Coef.   Std. Err.       z     P>|z|       [95% Conf. Interval]
---------+--------------------------------------------------------------------
     age |   -.035637   .0132867     -2.682   0.007      -.0616785   -.0095955
     mar |   .2190039   .1414675      1.548   0.122      -.0582673    .4962752
     doc |   .0007312   .0004779      1.530   0.126      -.0002055    .0016678
     und |   .1426933   .0316167      4.513   0.000       .0807256     .204661
   _cons |   1.068791   .4809021      2.222   0.026       .1262407    2.011342
------------------------------------------------------------------------------

.
```

Ejercicio 2. Los datos que siguen corresponden al registro de nuevos casos de melanoma informados en Estados Unidos durante 1969-1991 en hombres blancos clasificados por edad y región (Koch, Imrey et al., 1995). La última columna es el tamaño de la población registrada en el censo de EE.UU.

Región	Edad	Casos	Pop
Norte	0-35	61	2880262
	35-44	76	564535
	45-54	98	592983
	55-64	104	450740
	65-74	63	270908
	75+	80	161850
Sur	0-35	64	1074246
	35-44	75	220407
	45-54	68	198119
	55-64	63	134084
	65-74	45	70708
	75+	27	34233

En el archivo koch.txt se encuentran los datos.

a) Grafique el log de la tasa observada (número de casos sobre tamaño de la población) versus la edad. ¿Qué ve en este gráfico?

b) Ajuste un modelo de Poisson incorporando un interecpt, una variable dummy para la región, variables dummies para distinguir los grupos de edad (tome como base el grupo de menor edad) y el offset que le parezca adecuado. (Recuerde que para incorporar el offset debe usar el procedimeto **glim( )**)

¿ El ajuste parece adecuado?

c) Chequee si la función de varianza es adecuada. ¿Qué puede estar ocurriendo? ¿Le parece que el problema puede ser solucionado utilizando un parámetro de escala o cambiando a un modelo binomial-negativo?