MAEA: Problemas 1,2 y 3

Code

Alberto Rincón Borreguero

Problema 1

Se clasificó a 177 personas casadas según su estatus de fumador, variable B, con valores de No Fumador, b1, Poco Fumador, b2, (< 6 cigarrillos/día), Fumador Moderado, b3 (≥ 6 y < 15 cigarrillos/día) y Gran Fumador, b4 (≥ 15 cigarrillos/día), y el de su pareja, variable A, con valores No Fumador, a1, Poco Fumador, a2 (< 6 cigarrillos/día), Fumador Moderado, a3 (≥ 6 y < 15 cigarrillos/día) y Gran Fumador, a4 (≥ 15 cigarrillos/día). Los resultados aparecen recogidos en la siguiente tabla:

Hide

```
X           <- matrix(c(42,12,18,2,18,22,6,8,4,8,10,12,0,2,6,7), ncol = 4)
colnames(X) <- c("No Fumador","Poco Fumador","Fumador Moderado","Gran Fumador")
rownames(X) <- c("Pareja No fumadora","Pareja Poco Fumadora","Pareja Fumadora Moderad
a","Pareja Gran Fumadora")
X
```

```
                       No Fumador Poco Fumador Fumador Moderado Gran Fumador
Pareja No fumadora             42           18                4            0
Pareja Poco Fumadora           12           22                8            2
Pareja Fumadora Moderada       18            6               10            6
Pareja Gran Fumadora            2            8               12            7
```

Test de independencia de caracteres χ^2 con hipótesis nula h_0 : $Independecia\ entre\ las\ variables\ B\ y\ A.$

Hide

```
chi2 = chisq.test(X)
chi2
```

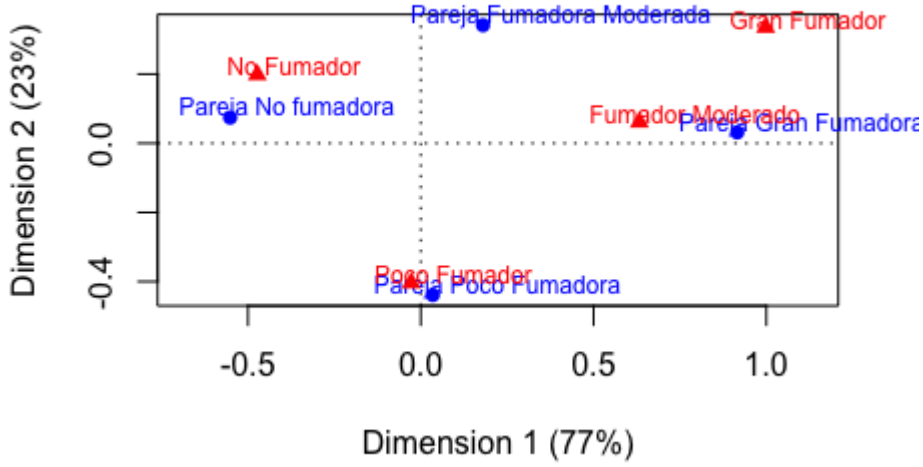
```

    Pearson's Chi-squared test

data:  X
X-squared = 58.661, df = 9, p-value = 2.427e-09
```

Dado que el p-valor es prácticamente cero, 0.0000000024, se rechaza la hipótesis nula de independecia entre ambas variables.

Se procede a realizar un análisis de correspondencias para comprobar si exista alguna relación entre los valores observados de las dos variables.

Hide

```
library(ca)
correspondencias <- ca(X)
plot(correspondencias)
```



El gráfico bi-dimensional obtenido establece que las personas entrevistadas que se declararon No Fumadoras o Poco Fumadoras tienen parejas con el mismo hábito. Por otra parte, cuando el entrevistado se declara como Fumador Moderado o Gran Fumador, su pareja, por lo general, es Gran Fumadora. Se aprecia que la observación de la variable A, Pareja Fumadora Moderada, no tiene una relación estrecha con ninguna observacion de la variable B, siendo Poco Fumador la más alejada de todas.

Problema 2

Hide

```
injerto <- read.csv("datos/examen/injerto.txt", sep=" ")
head(injerto)
```

	pnr <int>	rcpage <int>	donage <int>	type <int>	preg <int>	index <dbl>	gvhd <int>	time <int>	dead <int>
1	1	27	23	2	0	2.7e-01	0	95	1
2	2	13	18	2	0	3.1e-01	0	1385	0
3	3	19	19	1	0	3.9e-01	0	465	1
4	4	21	22	2	0	4.8e-01	0	810	1
5	5	28	38	2	0	4.9e-01	0	1497	0
6	6	22	20	2	0	5.0e-01	0	1181	1

6 rows

(a) Dado que las covariables a considerar en un modelo de regresión deben ser independientes, contrastar mediante un test de Spearman de independencia, si pueden considerarse independientes las covariables rcpage y donage.

Hide

```
corr
```

```
    Spearman's rank correlation rho

data:  injerto$rcpage and injerto$donage
S = 2324.1, p-value = 3.985e-07
alternative hypothesis: true rho is not equal to 0
sample estimates:
      rho
0.7245021
```

El test de spearman se interpreta de la siguiente forma: Magnitudes de rho cercanas a 1 indican mayor correlación mientras que los valores cercanos a cero indican una menor correlación. En este caso, el valor **0.72** indica que las variables rcpage y donage **no son independientes**.

(b) Analizar mediante una Regresión Logística qué variables son significativas para predecir la probabilidad p de presentar la enfermedad de Injerto contra Huésped, variable gvhd, de entre las 3 covariables siguientes: index, donage, preg.

Hide

```
summary(regression)
```

```
Call:
glm(formula = gvhd ~ index + donage + preg, family = binomial(link = "logit"),
    data = injerto)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0716  -0.4978  -0.2732   0.6925   1.9978

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.88275    2.22347  -2.646  0.00815 **
index        0.88989    0.37068   2.401  0.01637 *
donage       0.11925    0.06261   1.905  0.05682 .
preg         1.55904    1.01886   1.530  0.12597
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51.049  on 36  degrees of freedom
Residual deviance: 29.848  on 33  degrees of freedom
AIC: 37.848

Number of Fisher Scoring iterations: 5
```

Se observa que la variable $index$ con p-valor de 0.01 es la más significativa en la predicción.

(c) Determinar la estimación de p en función de las variables que resulten significativas.

Hide

```
summary(regression)
```

```
Call:
glm(formula = gvhd ~ index, family = binomial(link = "logit"),
    data = injerto)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9188  -0.7462  -0.5665   0.8256   1.6821

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.9885     0.7479  -2.659  0.00784 **
index         0.7747     0.2921   2.652  0.00799 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 51.049  on 36  degrees of freedom
Residual deviance: 39.211  on 35  degrees of freedom
AIC: 43.211


Number of Fisher Scoring iterations: 5
```

d. ¿Qué probabilidad de presentar la Enfermedad de injerto contra huésped tiene un individuo con índice reacciones de linfocitos igual a 2’5, cuyo donante que nunca ha estado embarazada y tiene una edad de 55 años?

Hide

```
predict.glm(object = regression, data.frame(index=2.5, preg=0 ,donage=55), type = "re
sponse")
```

```
        1
0.4870662
```

La probabilidad de presentar la enfermedad es del **48.71%**

Problema 3

Se desea realizar una Regresiónn no Lineal ajustando una función tipo sigmoide a los siguientes pares de datos,

x	y
19	65
25	61
38	56
47	28
53	12

x	y
69	10

utilizando la correspondiente función de autoarranque. Determinar la función sigmoide ajustada.

Hide

```
pb3.datos <- data.frame(x = c(19,25,38,47,53,69),
                        y = c(65,61,56,28,12,10))
```

Sea la función sigmoide:

$$\eta(x,\theta) = \theta_1 + \frac{\theta_2 - \theta_1}{1 + e^{\theta_3(x-\theta_4)}}$$

Se realiza la regresión no lineal, con su correspondiente función de arranque SSfpl, en la siguiente linea.

Hide

```
model <- nls(y~ SSfpl(-x,b1,b2,b3,b4), data=pb3.datos)
```

Obtenemos información acerca del modelo generado mediante la función summary.

Hide

```
summary(model)
```

```
Formula: y ~ SSfpl(-x, b1, b2, b3, b4)

Parameters:
   Estimate Std. Error t value Pr(>|t|)
b1   9.1077     2.1178   4.301 0.050044 .
b2  62.9236     1.6828  37.392 0.000714 ***
b3 -44.7386     0.7732 -57.865 0.000299 ***
b4   3.3635     0.6024   5.584 0.030610 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.335 on 2 degrees of freedom

Number of iterations to convergence: 0
Achieved convergence tolerance: 9.216e-06
```

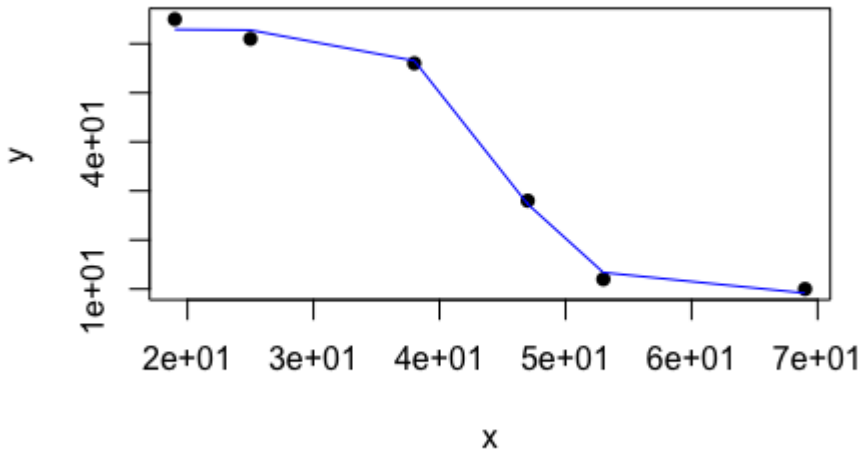
La suma de los errores residuales es muy pequeña, como se aprecia a continuación. Por tanto podemos decir que el modelo se ajusta adecuadamente a los datos.

Hide

```
sum(resid(model))
```

```
[1] 7.105e-15
```

Por último, mostramos el ajuste del modelo (linea azul) para los datos de la columna y del *dataframe*.


y
1e+01
4e+01
2e+01
3e+01
4e+01
5e+01
6e+01
7e+01
x