UNIVERSIDADE DE ÉVORA
ESCOLA DE CIÊNCIAS E TECNOLOGIA

Escola de Ciências e Tecnologia da Universidade de Évora

Mestrado em Modelação Estatística e Análise de Dados

Ano Letivo 2017/2018

2º Semestre

U.C.: Análise Categórica de Dados

Modelos Lineares Generalizados

Docente:

Dulce Pereira

Discente:

Ana Sapata n.º39504

1.

a) Inicialmente procedeu-se a um modelo linear generalizado (MLG) seguindo este uma distribuição normal, tendo-se obtido o seguinte resultado

```
> fit1<-glm(dados$client ~ dados$habit + dados$rendim + dados$idade +
+             dados$dist_conc + dados$dist_loja, family="gaussian", data=dados)
> summary(fit1)

Call:
glm(formula = dados$client ~ dados$habit + dados$rendim + dados$idade +
    dados$dist_conc + dados$dist_loja, family = "gaussian", data = dados)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.8067  -1.7834  -0.2094   1.8458  10.1409

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1.943e+01  2.163e+00   8.984 1.24e-14 ***
dados$habit      6.600e-03  1.521e-03   4.339 3.33e-05 ***
dados$rendim    -1.162e-04  2.283e-05  -5.088 1.61e-06 ***
dados$idade     -3.593e-02  1.879e-02  -1.912   0.0587 .
dados$dist_conc  1.904e+00  2.579e-01   7.382 3.99e-11 ***
dados$dist_loja -1.710e+00  1.739e-01  -9.837  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for gaussian family taken to be 10.32092)

    Null deviance: 4805.6  on 109  degrees of freedom
Residual deviance: 1073.4  on 104  degrees of freedom
AIC: 576.76

Number of Fisher Scoring iterations: 2
```

Fig. 1 Modelo Normal

Escrevendo-se então o modelo da seguinte forma

$$Client = 1.943e^{+01} + 6.600e^{-03}habit - 1.162e^{-04}rendim - 3.593e^{-02}idade + 1.904e^{+00}dist\_conc - 1.710e^{+00}dist\_loja$$

De seguida procedeu-se à validação dos pressupostos do mesmo:

- Homogeneidade das variâncias

$$H_0: As\ variâncias\ são\ todas\ iguais$$

$$H_1: Pelo\ menos\ uma\ das\ variancias\ difere$$

Pelo teste de Breusch Pagan obteve-se um $p - value = 0.0135 < 0.05 = \alpha$, rejeitando-se assim a hipótese nula. Portanto, com um nível de significância de 5% não existe evidencia estatística para afirmar que as variâncias são iguais.

Falhando o pressuposto da homogeneidade então este modelo já não poderá ser usado nos dados em análise.

b) De seguida ajustou-se um modelo de regressão de Poisson aos dados onde foram obtidos os seguintes dados

Variável	$\boldsymbol{\beta}$	$\boldsymbol{Exp(\beta)}$	**Std.Error**	**p-value**
Interceção	$2.942e^{+00}$	18.9620	$2.072e^{-01}$	$< 2e^{-16}$
Habit	$6.058e^{-04}$	1.0006	$1.421e^{-04}$	$2.02e^{-05}$
Rendim	$-1.169e^{-05}$	0.9999	$2.112e^{-06}$	$3.13e^{-08}$
Idade	$-3.726e^{-03}$	0.9963	$1.782e^{-03}$	0.0365
Dist_conc	$1.684e^{-01}$	1.1834	$2.577e^{-02}$	$6.39e^{-11}$
Dist_loja	$-1.288e^{-01}$	0.8792	$1.620e^{-02}$	$1.89e^{-15}$

Tabela 1 Modelo de Poisson

A incidência do número de clientes aumenta 1.1834 com o aumento da distância entre a área e a loja concorrente mais próxima, ou seja, a loja terá mais clientes quanto mais longe estiver a loja concorrente mais próxima.

Já a incidência do número de clientes diminui 0.8792 com o aumento da distância entre a zona e a loja, portanto quanto mais longe estiver a loja menos clientes terá.

As variáveis que mais explicam o numero esperado de clientes em cada zona são a distancia entre a zona e a loja e a distancia entre a área e a loja concorrente mais próxima, pois no modelo, são as que têm um valor de $\exp(\beta)$ mais diferentes de 1 e que também têm um p-value menor, pelo que seriam sempre as ultimas a ser retiradas do modelo.

c) No gráfico (Fig.2) referente à adequabilidade da função de ligação vê-se que existe alguma dispersão, mas os dados formam aproximadamente uma linha, pelo que a função é adequada.

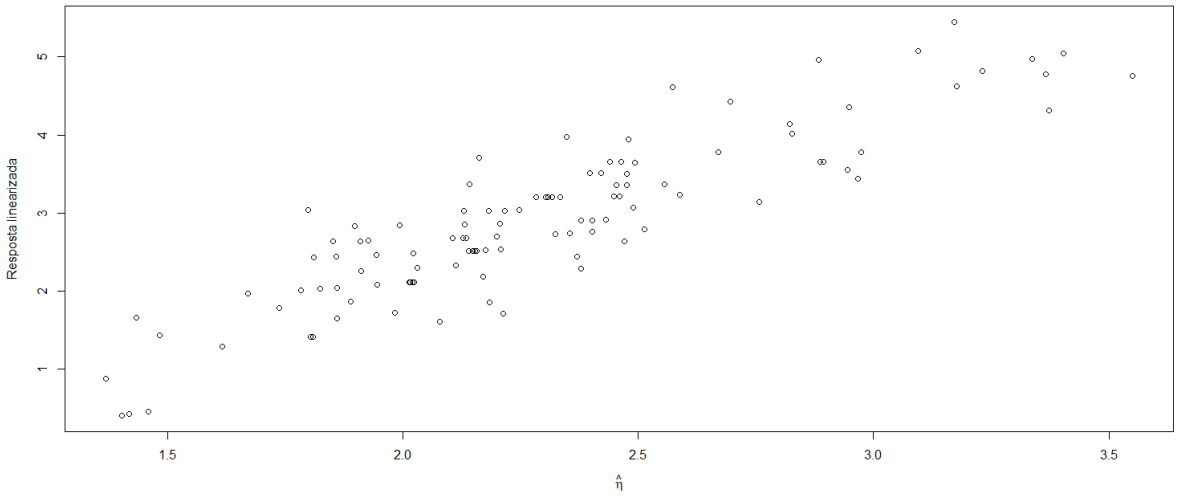


Fig. 2 Adequabilidade da função de ligação

No gráfico (Fig3) referentes aos resíduos, observa-se que estes estão dispersados aleatoriamente em torno do 0, o que significa que o modelo é adequado aos dados.

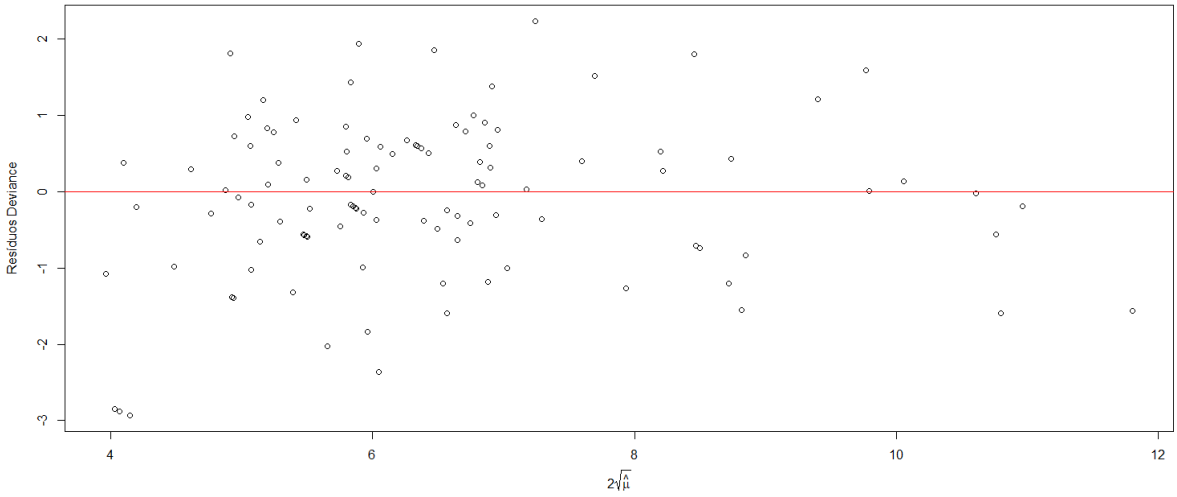


Fig. 3 Residuos

d) Numa área com as seguintes caraterísticas

- Número de domicílios: 500
- Rendimento médio anula: 38000€
- Idade média das habitações: 45 anos
- Distância da zona à loja concorrente: 5Km
- Distância da zona à loja: 7Km

São esperados cerca de 13 clientes. E com 95% de certeza são esperados entre 11 e 14 clientes.

2. Foram observados 33 pacientes, dos quais 17 tinham a características morfológica nos glóbulos brancos e nos restantes 16 a mesma estava ausente. Usando como covariaveis o log do WBC e o facto de o teste ter dado positivo ou negativo foram obtidos os seguintes coeficientes para um modelo gama.

Variável	β	$\exp(\beta)$	**Erro Padrão**	**p-value**
Interceção	5.4741	238.43	1.3732	0.0004
AG(Positivo)	1.0454	2.8445	0.3582	0.0066
logWBC	-0.2700	0.7634	0.1402	0.0637

Tabela 2 Modelo Gama

Pode-se concluir que o risco de do paciente ter leucemia é maior nos que tiveram teste positivo.

O modelo é então da forma

$$\lambda_i = e^{5.4741+1.0454AG(Positivo)-0.2700logWBC}$$

Para a variável AG(Positivo) tem-se um OR de 184.45%, portanto o risco de uma pessoa que obtém um resultado positivo no AG ter leucemia é 184vezes superior ao de uma pessoa que obtém resultado negativo no AG.

Para a variável logWBC tem-se um OR de -23.66%, o que significa que para pessoas com um logWBC mais elevado o risco de terem leucemia é 23.66% mais baixo do que pessoas com um valor de logWBC mais baixo.

Para a verificação dos pressupostos obteve-se os seguintes outputs do R

```
> bptest(mod1, studentize=F) #p-value=0.08025>0.05 não se rejeita H0

        Breusch-Pagan test

data:  mod1
BP = 5.0453, df = 2, p-value = 0.08025
```

Fig. 4 Breusch-Pagan Test

```
> ncvTest (lm(tab$Temp ~ tab$AG + tab$logWBC)) #p-value=0.04<0.05 rejeita-se H0
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 4.164719    Df = 1     p = 0.04127426
```

Fig. 5 ncvTest

```
> lillie.test(rs) #p-value=0.3912>0.05, não se rejeita H0

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  rs
D = 0.11022, p-value = 0.3912
```

Fig.6 Normality Test

```
> jarque.bera.test (rs) #p-value=0.601

        Jarque Bera Test

data:  rs
X-squared = 1.0183, df = 2, p-value = 0.601
```

Fig.7 Jarque Bera Test

```
> a #p-value=0.3856

        Anderson-Darling normality test

data:  rs
A = 0.37911, p-value = 0.3856
```

Fig.8 Anderson-Darling test

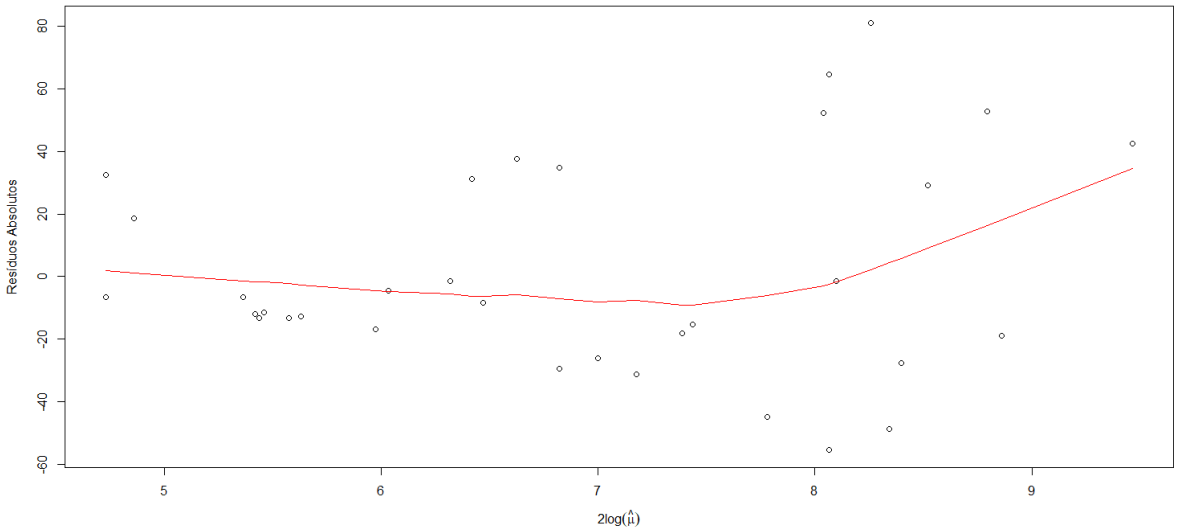
```
> mcor #não existe multicolinearidade
                  as.numeric.tab.AG. tab.logWBC
as.numeric.tab.AG.          1.0000000 -0.1277047
tab.logWBC                 -0.1277047  1.0000000
```

Fig.9 Multicolinearidade

```
> durbinWatsonTest(mod1) #p-value=0<.05 rejeita-se H0
 lag Autocorrelation D-W Statistic p-value
   1       0.3338939      1.251875       0
 Alternative hypothesis: rho != 0
```

Fig.10 DurbinWatsonTest

Para um nível de significância de 5% todos os pressupostos se verificam pelo que o modelo se pode aplicar. Quando à adequabilidade do mesmo obteve-se o seguinte gráfico que indica que o modelo é adequado.


Resíduos Absolutos
80
60
40
20
0
-20
-40
-60
5
6
7
8
9
2log($\hat{\mu}$)