Modelos Lineares Generalizados - Prova 1

Mariana Costa Freitas

2024-11-26

Questão 4

a) Ajuste um modelo comparando a ligação canônica com pelo menos mais uma função de ligação. Qual é o modelo mais adequado? Explique detalhadamente suas análises.

Como, nesse caso, queremos modelar a variável resposta `Count`, que representa a contagem do número de mortes por câncer em três regiões do Canadá, vamos usar o modelo Poisson, que trata de contagens, cuja função de ligação canônica é $\theta = log(\mu)$. Para fazer a comparação com outra função de ligação, vamos usar $\theta_2 = \frac{1}{\mu^2}$.

A seguir, vamos elaborar os modelos com essas duas funções de ligação usando a função `glm()` e alterando as funções de ligação usando o argumento `link`.

```
#Carregando os pacotes utilizados
library(GLMsData)

#Carregando os dados
data(ccancer)

# Ajuste do modelo com a ligação log (canônica)
modelo_log <- glm(Count ~ Gender + Region + Site + Population,
                  data = ccancer,
                  family = poisson(link = "log"))
summary(modelo_log)
```

```
##
## Call:
## glm(formula = Count ~ Gender + Region + Site + Population, family = poisson(link = "log"),
##     data = ccancer)
##
## Coefficients: (1 not defined because of singularities)
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)     3.82199    0.04225  90.461   <2e-16 ***
## GenderM         0.15700    0.01283  12.236   <2e-16 ***
## RegionOntario   2.97522    0.03976  74.838   <2e-16 ***
## RegionQuebec    2.78287    0.03996  69.642   <2e-16 ***
## SiteColorectal  0.21294    0.02227   9.561   <2e-16 ***
## SiteLung        1.16922    0.01896  61.661   <2e-16 ***
## SitePancreas   -0.61279    0.02794 -21.932   <2e-16 ***
## SiteProstate   -0.34557    0.02573 -13.431   <2e-16 ***
## Population           NA         NA      NA       NA
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 35187.5  on 29  degrees of freedom
## Residual deviance:  9375.8  on 22  degrees of freedom
## AIC: 9582.6
##
## Number of Fisher Scoring iterations: 6
```

```r
# Ajuste do modelo com outra ligação
modelo_inversa_2 <- glm(Count ~ Gender + Region + Site + Population,
                        data = ccancer,
                        family = poisson(link = "1/mu^2"))
summary(modelo_inversa_2)
```

```
##
## Call:
## glm(formula = Count ~ Gender + Region + Site + Population, family = poisson(link = "1/mu^2"),
##     data = ccancer)
##
## Coefficients: (1 not defined because of singularities)
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     2.241e-04  1.737e-05  12.897   <2e-16 ***
## GenderM        -1.096e-07  6.003e-09 -18.265   <2e-16 ***
## RegionOntario  -2.226e-04  1.737e-05 -12.815   <2e-16 ***
## RegionQuebec   -2.226e-04  1.737e-05 -12.814   <2e-16 ***
## SiteColorectal -5.540e-07  5.313e-08 -10.426   <2e-16 ***
## SiteLung       -1.240e-06  4.717e-08 -26.297   <2e-16 ***
## SitePancreas    2.927e-06  2.017e-07  14.516   <2e-16 ***
## SiteProstate    1.070e-06  1.076e-07   9.944   <2e-16 ***
## Population             NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 35187.5  on 29  degrees of freedom
## Residual deviance:  9550.3  on 22  degrees of freedom
## AIC: 9757.1
##
## Number of Fisher Scoring iterations: 8
```

A partir desses outputs, podemos observar as estimativas para o intercepto e para os coeficientes associados a cada variável explicativa. Porém, é importante notar que a variável `Population` foi removida em ambos os modelos, pois apresenta colinearidade, ou seja, essa é altamente correlacionada com outras variáveis explicativas ou mesmo com a variável resposta. Além disso, todas as outras variáveis tem um p-valor pequeno, indicando então que são importantes no modelo.

Para definir o modelo mais adequado, precisamos analisar o ajuste do modelo (residual deviance) para cada um e também as medidas do AIC e BIC, que quanto mais baixos geralmente indicam o modelo melhor ajustado. Já temos as medidas do ajuste dos modelos e do AIC, que foram apresentados no output acima e ambos são mais baixos no modelo que usa a função de ligação canônica, mostrando que esse é o mais

adequado. Abaixo, apresentamos as medidas do BIC para ambos os modelos, o que reforça que o modelo que melhor se ajusta aos dados é o modelo de função de ligação canônica.

```
BIC(modelo_log, modelo_inversa_2)
```

```
##                   df      BIC
## modelo_log         8 9593.853
## modelo_inversa_2   8 9768.350
```

b) Considerando o melhor modelo obtido no item anterior, realize a análise dos desvios e interprete detalhadamente cada passo.

A análise dos desvios nos permite avaliar o quão bem o modelo ajusta os dados. A função desvio indica a diferença entre o modelo ajustado e um modelo saturado. Quanto menor o desvio, melhor o ajuste. Abaixo, vamos obter essa medida usando a função `deviance()`.

```
# Obtendo o desvio do modelo
deviance(modelo_log)
```

```
## [1] 9375.846
```

Para nos aprofundar melhor, podemos ainda usar a função `anova()` para avaliar a contribuição de cada variável explicativa para a diminuição do desvio.

O modelo `fit0`abaixo representa o modelo apenas com o intercepto, `fit1`, apresenta apenas a variável `Gender` como explicativa, `fit2`inclui `Region`e `fit3`inclui `Site`.

```
fit0 <- glm(Count ~ 1,
                  data = ccancer,
                  family = poisson(link = "log"))

fit1 <- glm(Count ~ Gender,
                  data = ccancer,
                  family = poisson(link = "log"))

fit2 <- glm(Count ~ Gender + Region,
                  data = ccancer,
                  family = poisson(link = "log"))

fit3 <- glm(Count ~ Gender + Region + Site,
                  data = ccancer,
                  family = poisson(link = "log"))
```

Agora, vamos usar `anova()` para avaliar os efeitos das variáveis.

```
anova(fit0, fit1)
```

```
## Analysis of Deviance Table
##
## Model 1: Count ~ 1
## Model 2: Count ~ Gender
```

```
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        29      35187
## 2        28      35037  1   150.17 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Podemos observar que a adição da variável `Gender` causou uma diminuição de 150 (35187 - 35037) no desvio e diminui também em um grau de liberdade.

```
anova(fit1, fit2)
```

```
## Analysis of Deviance Table
##
## Model 1: Count ~ Gender
## Model 2: Count ~ Gender + Region
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        28      35037
## 2        26      20179  2    14858 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

A adição da variável `Region`dado que `Gender`já estava presente no modelo representou uma grande diminuição de 14858 no desvio e de 2 graus de liberdade.

```
anova(fit2, fit3)
```

```
## Analysis of Deviance Table
##
## Model 1: Count ~ Gender + Region
## Model 2: Count ~ Gender + Region + Site
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1        26    20178.9
## 2        22     9375.8  4    10803 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Já a adição da variável `Site`dado que `Gender`e `Region`já estavam presentes, ocasionou uma diminuição ainda maior de 10803.1 no desvio e 4 graus de liberdade.

c) Interprete o modelo final ajustado e explique o critério de seleção usado.

Vamos interpretar os coeficientes do modelo escolhido abaixo:

```
modelo_log$coefficients
```

```
##    (Intercept)        GenderM  RegionOntario   RegionQuebec SiteColorectal
##      3.8219926      0.1570000      2.9752226      2.7828740      0.2129408
##       SiteLung   SitePancreas   SiteProstate     Population
##      1.1692195     -0.6127880     -0.3455670             NA
```

Cada coeficiente indica a mudança na taxa de mortes. Assim, a partir dos coeficientes,podemos observar que homens têm uma taxa de mortalidade maior que mulheres, Ontário e Quebec têm taxas de mortalidade por câncer mais altas que Newfoundland, o câncer de pulmão tem a maior taxa de mortalidade, seguido de câncer colorretal e o câncer de pâncreas tem a menor taxa de mortalidade.

Para selecionar o melhor modelo, usamos as medidas de AIC, BIC e de desvio, que se mostraram serem mais baixas para o modelo que usa função de ligação logarítmica.