Trabalho de dados Binários

Acidentes de carro

Laís Hoffmann, Simone Matsubara, Yasmin Fernandes, Willian Meira

2018-11-12

1. Base de Dados

1.1 Descrição dos dados

Os dados foram retirados do pacote "DAAG", sendo dados dos EUA, entre 1997-2002, de acidentes de carro relatados pela polícia nos quais há um evento prejudicial (pessoas ou propriedade) e do qual pelo menos um veículo foi rebocado. Os dados são restritos aos ocupantes do banco da frente, incluem apenas um subconjunto das variáveis registradas e são restritos de outras maneiras também.

A base original possui uma base de dados com 26.217 observações nas 15 variáveis a seguir.

1 - **veloc**: velocidades estimadas do impacto do acidente: 1-9km/h, 10-24, 25-39, 40-54, 55+
2 - **pesos**: Pesos de observação
3 - **sobrev**: Classificação se sobreviveu ao acidente: 1 = morreu ou 0 = sobreviveu
4 - **airbag**: Se o carro possui airbag: com ou sem airbag
5 - **cinto**: uso do cinto de segurança: com ou sem cinto
6 - **frontal**: impacto do acidente: 0 = não frontal, 1 = impacto frontal
7 - **sexo**: Sexo: 0 = Feminino ou 1 = Masculino
8 - **idade**: Idade dos ocupantes do veículo
9 - **anoaci**: Ano do acidente (1997-2002)
10 - **anovei**: Ano do veículo (1953-2003)
11 - **airbagcat**: Se Airbags foram acionados: deploy, nodeploy, unavail
12 - **ocupantes**: Posição do airbag acionado: driver, pass
13 - **abfunc**: Airbag acionados: 0: Se não possuia airbag ou não foi acionado, 1: Um ou mais airbags foram acionados
14 - **grav**: Gravidade do acidente: 0:none, 1 = Possível Lesão, 2:no incapacity, 3:incapacity, 4:killed; 5:unknown, 6:prior death
15 - **numcaso**: Número do caso.

No entanto, escolhemos analisar os dados do ano do acidente de 2002 e veículos de ano 2000 e retirar as variaveis weight, abcat e caseid.

2 Análise Descritiva

2.1 Medidas de Resumo

```
summary(dados[ , c(1:8,10)])
```

```
##    veloc     sobrev    airbag    cinto     frontal     sexo
##  01-09: 12   Nao: 23   Nao:  1   Nao:121   Nao:183   Fem :254
##  10-24:293   Sim:470   Sim:492   Sim:372   Sim:310   Masc:239
##  25-39:121
##  40-54: 46
##  55+  : 21
##
##
```

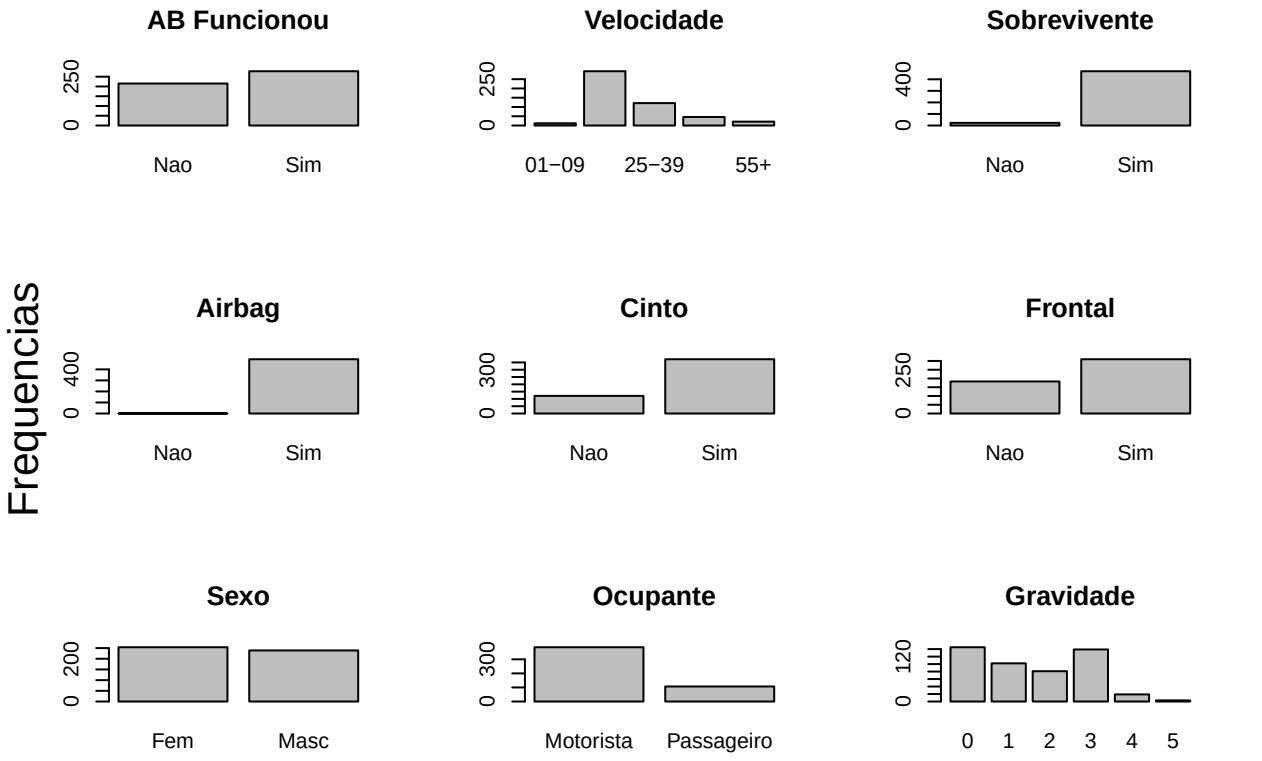
```
##      idade            ocupantes     grav
##  Min.   :16.00   Motorista :386   0   :145
##  1st Qu.:23.00   Passageiro:107   1   :102
##  Median :35.00                    2   : 81
##  Mean   :37.82                    3   :139
##  3rd Qu.:48.00                    4   : 19
##  Max.   :93.00                    5   :  3
##                                   NA's:  4
```

Nota-se na varável velocidade uma frequência maior na faixa 10-24 milhas.

A maioria estava com cinto de segurança e os acidentes foram a maioria frontais.

2.3 Histogramas

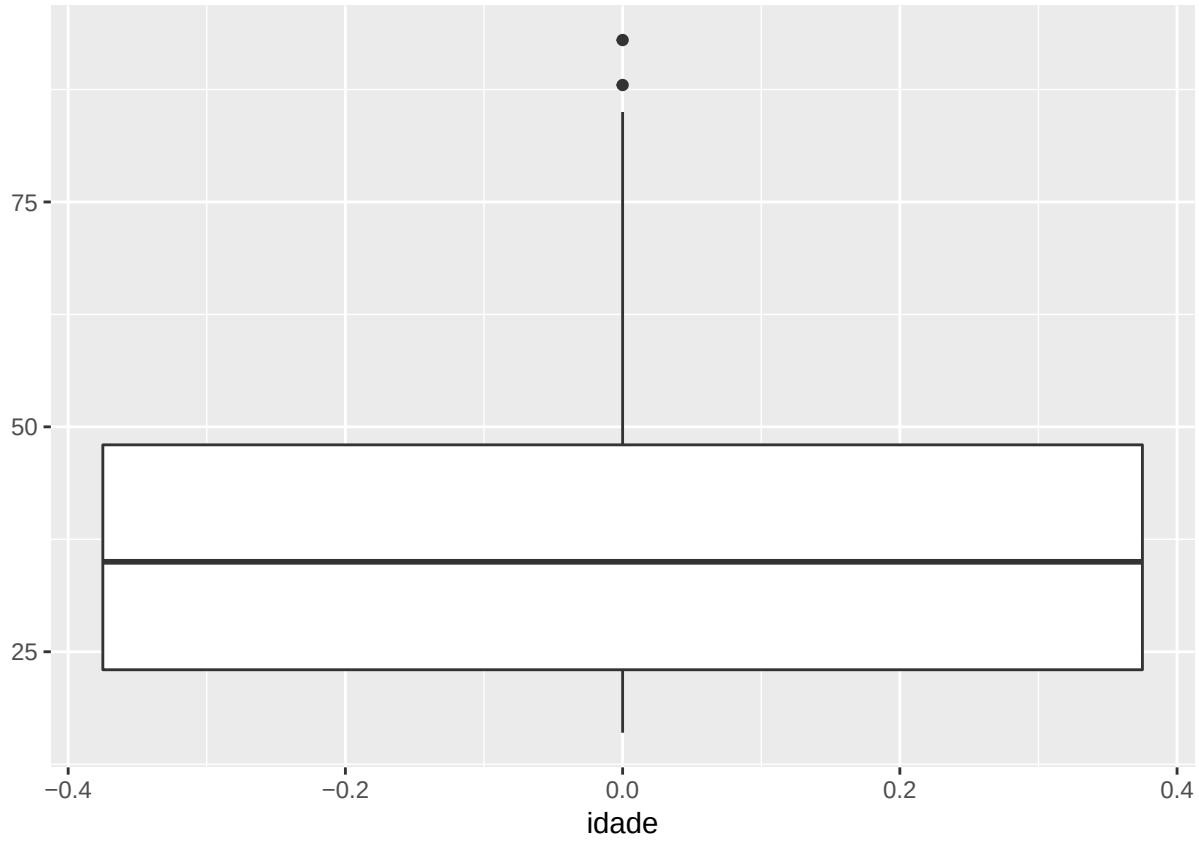
```
par(mfrow = c(3,3))
plot(dados$abfunc, xlab = '', ylab = '', main = 'AB Funcionou')
plot(dados$veloc, xlab = '', ylab = '', main = 'Velocidade')
plot(dados$sobrev, xlab = '', ylab = '', main = 'Sobrevivente')
plot(dados$airbag, xlab = '', ylab = '', main = 'Airbag')
plot(dados$cinto, xlab = '', ylab = '', main = 'Cinto')
plot(dados$frontal, xlab = '', ylab = '', main = 'Frontal')
plot(dados$sexo, xlab = '', ylab = '', main = 'Sexo')
plot(dados$ocupantes, xlab = '', ylab = '', main = 'Ocupante')
plot(dados$grav, xlab = '', ylab = '', main = 'Gravidade')
mtext(side=2,cex=1.3,line=-1.5,text="Frequencias",outer=TRUE)
```



2.4 Distribuição

```
g1<-ggplot(dados, aes(y=idade)) +
  geom_boxplot()+ xlab('idade')+ ylab('') +
  theme(legend.title=element_blank())


g1
```



###Intuitivamente sabemos que para nosso escopo a variável idade não é significativa para o nosso modelo porém para comprovar adiante faremos um teste para evidênciar a irrelevancia da variável no modelo.

3. Ajuste do Modelo de Regressão

3.1 Ligação Logito

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Logito. A expressão do modelo é dada por:

$ln(\frac{\pi_i}{1-\pi_i}) = \beta_0 + \beta_1 Veloc_i + \beta_2 Sobrev_i + \beta_3 Airbag_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 Grav_i$

No R, o modelo é declarado da seguinte forma:

```
ajuste1 <- glm(abfunc ~ .,family=binomial(link='logit'),data = dados)
```

3.2 Ligação Probito

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Probito. A expressão do modelo é dada por:

$\phi^{-1}(\pi_i) = \beta_0 + \beta_1 Veloc_i + \beta_2 Sobrev_i + \beta_3 Airbag_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 Grav_i$

No R, o modelo é declarado da seguinte forma:

```
ajuste2 <- glm(abfunc ~ .,family=binomial(link = 'probit'),data = dados)
```

3.3 Ligação Complemento log-log

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Complemento Log Log. A expressão do modelo é dada por:

$ln[-ln(1-\pi_i)] = \beta_0 + \beta_1 Veloc_i + \beta_2 Sobrev_i + \beta_3 Airbag_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 Grav_i$

No R, o modelo é declarado da seguinte forma:

```
ajuste3 <- glm(abfunc ~ .,family=binomial(link='cloglog'),data = dados)
```

3.4 Ligação Cauchy

Vamos ajustar um Modelo Linear Generalizado Binomial com função de ligação Cauchy. A expressão do modelo é dada por:

$tan[\pi_i(\mu_i - 0,5)] = \beta_0 + \beta_1 Veloc_i + \beta_2 Sobrev_i + \beta_3 Airbag_i + \beta_4 Cinto_i + \beta_5 Frontal_i + \beta_6 Sexo_i + \beta_7 Idade_i + \beta_8 Ocupantes_i + \beta_9 Grav_i$

No R, o modelo é declarado da seguinte forma:

```
ajuste4 <- glm(abfunc ~ .,family=binomial(link='cauchit'),data = dados)
```

4. Escolha do Modelo

XXXXXXX lineu Para seleção de modelos diversas medidas podem ser utilizadas, em especial vamos utilizar a verossimilhança dos modelos.

O critério de informação AIC pode também ser utilizado, porém o AIC penaliza o número de parâmetros do modelo. Como os modelos tem o mesmo número de parâmetros, o critério aponta para a mesma direção da verossimilhança pois todos são penalizados da mesma forma; para fins de ilustração, as duas quantidades são exibidas: XXXXXXX

```
##    ajuste      aic    logLik
## 1  logito 514.1660 -240.0830
## 2 probito 514.1476 -240.0738
## 3 cloglog 513.2723 -239.6361
## 4  cauchy 514.4334 -240.2167
```

O modelo que apresentou menor AIC e maior verossimilhança foi o modelo Binomial com função de ligação C Log-Log.

5. Análise do Modelo Ajustado Selecionado

5.1 Resumo do Modelo

```
summary(ajuste3)
```

```
##
## Call:
## glm(formula = abfunc ~ ., family = binomial(link = "cloglog"),
##     data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5970  -0.7381   0.1408   0.8342   2.5503
##
## Coefficients:
##                       Estimate Std. Error z value Pr(>|z|)
## (Intercept)         -17.117970 843.971464  -0.020 0.983818
## veloc10-24            1.657013   1.051727   1.576 0.115137
## veloc25-39            2.213254   1.056112   2.096 0.036112 *
## veloc40-54            3.083987   1.097654   2.810 0.004960 **
## veloc55+              2.614381   1.115507   2.344 0.019095 *
## sobrevSim            -0.924814   0.741988  -1.246 0.212618
## airbagSim            14.650899 843.970429   0.017 0.986150
## cintoSim             -0.289103   0.168727  -1.713 0.086631 .
## frontalSim            1.678297   0.179356   9.357  < 2e-16 ***
## sexoMasc             -0.056287   0.145436  -0.387 0.698737
## idade                -0.002542   0.004372  -0.581 0.560989
## ocupantesPassageiro  -0.017739   0.173337  -0.102 0.918490
## grav1                 0.624505   0.202880   3.078 0.002083 **
## grav2                 0.832691   0.229412   3.630 0.000284 ***
## grav3                 0.771844   0.203548   3.792 0.000149 ***
## grav4                -0.472641   0.823550  -0.574 0.566031
## grav5                -0.039578   0.777019  -0.051 0.959376
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 670.27  on 488  degrees of freedom
## Residual deviance: 479.27  on 472  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 513.27
##
## Number of Fisher Scoring iterations: 13
```

XXXXXX LIneu O resumo do modelo ajustado indica que as variáveis adesão marginal, nucléolos nus, cromatina branda, nucléolo normal e espessura do aglomerado estão associadas a uma maior probabilidade de tumor maligno, enquanto as demais variáveis não apresentam relação com a resposta.

XXXXXXX

5.2 Reajuste do Modelo

XXXXX Lineu Como as covariáveis são altamente correlacionadas, é válido inserir as covariáveis uma a uma no modelo para verificar sua significância na presença das outras, tal como o realizado pelo algoritmo stepwise.

Sendo assim, o novo modelo fica da seguinte forma:

XXXXX

XXXXXX

```
ajuste3.1 <- step(ajuste3, direction = "both")
```

```
summary(ajuste3.1)
```

```
##
## Call:
## glm(formula = abfunc ~ veloc + airbag + cinto + frontal + grav,
##     family = binomial(link = "cloglog"), data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.5133  -0.7352   0.1164   0.8654   2.5279
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -18.1312   843.9711  -0.021 0.982860
## veloc10-24    1.6646     1.0491   1.587 0.112586
## veloc25-39    2.2202     1.0524   2.110 0.034897 *
## veloc40-54    3.1220     1.0911   2.861 0.004219 **
## veloc55+      2.7701     1.1111   2.493 0.012663 *
## airbagSim    14.6278   843.9704   0.017 0.986172
## cintoSim     -0.2896     0.1652  -1.753 0.079642 .
## frontalSim    1.6614     0.1777   9.348  < 2e-16 ***
## grav1         0.6186     0.2002   3.089 0.002006 **
## grav2         0.8201     0.2249   3.647 0.000266 ***
## grav3         0.7718     0.1998   3.863 0.000112 ***
## grav4         0.3550     0.4228   0.840 0.401014
## grav5        -0.1153     0.7652  -0.151 0.880218
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 670.27  on 488  degrees of freedom
## Residual deviance: 481.18  on 476  degrees of freedom
##   (4 observations deleted due to missingness)
## AIC: 507.18
##
## Number of Fisher Scoring iterations: 13
```

```
selec2 <- data.frame(ajuste=c('aj3', 'aj3.1'),
                     aic=c(AIC(ajuste3), AIC(ajuste3.1)),
                     logLik=c(logLik(ajuste3),logLik(ajuste3.1)),
                     Dev=c(deviance(ajuste3),deviance(ajuste3.1)))
```

```
selec2
```

```
##   ajuste      aic    logLik      Dev
## 1    aj3 513.2723 -239.6361 479.2723
## 2  aj3.1 507.1838 -240.5919 481.1838
```

O resumo do novo modelo ajustado:

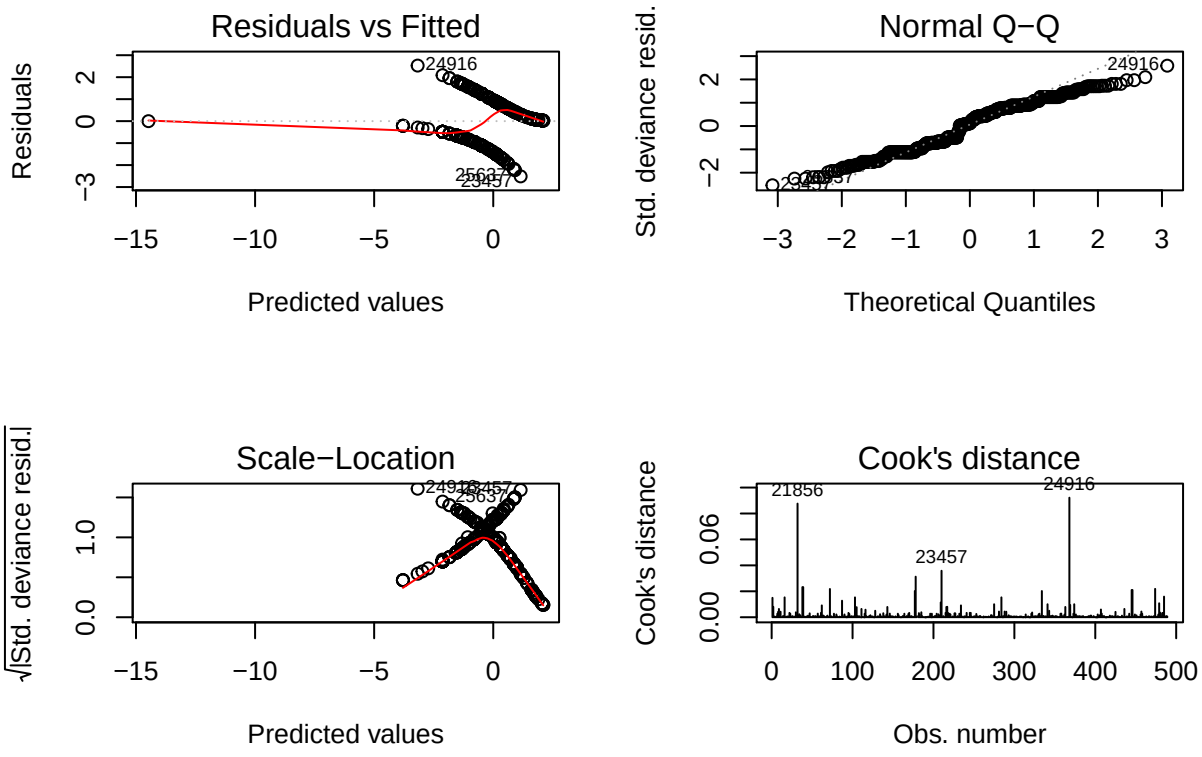
```
anova(ajuste3, ajuste3.1, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: abfunc ~ veloc + sobrev + airbag + cinto + frontal + sexo + idade +
##     ocupantes + grav
## Model 2: abfunc ~ veloc + airbag + cinto + frontal + grav
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1       472     479.27
## 2       476     481.18 -4  -1.9115    0.752
```

5.3 Análise de Resíduos

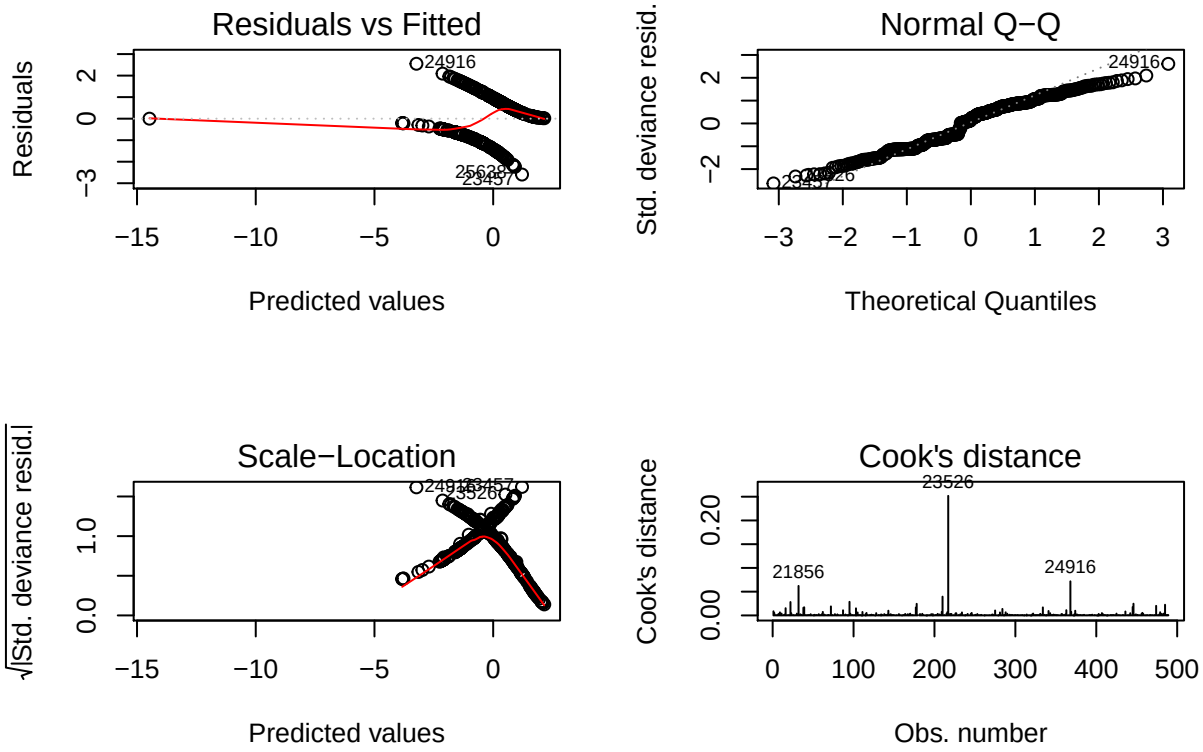
```
par(mfrow=c(2,2))
plot(ajuste3.1, 1:4)
```

```
## Warning: not plotting observations with leverage one:
##   180
```



```
par(mfrow=c(2,2))
plot(ajuste3, 1:4)
```

```
## Warning: not plotting observations with leverage one:
##   180
```

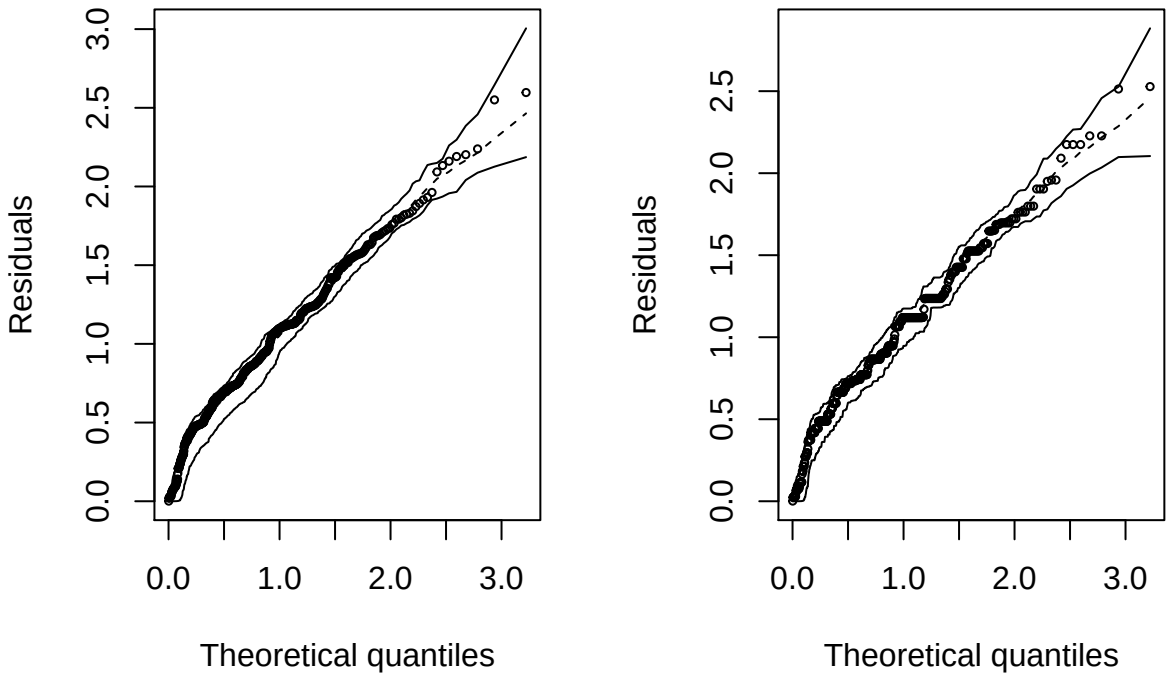



```
par(mfrow=c(1,2))
hnp(ajuste3)
```

```
## Binomial model
```

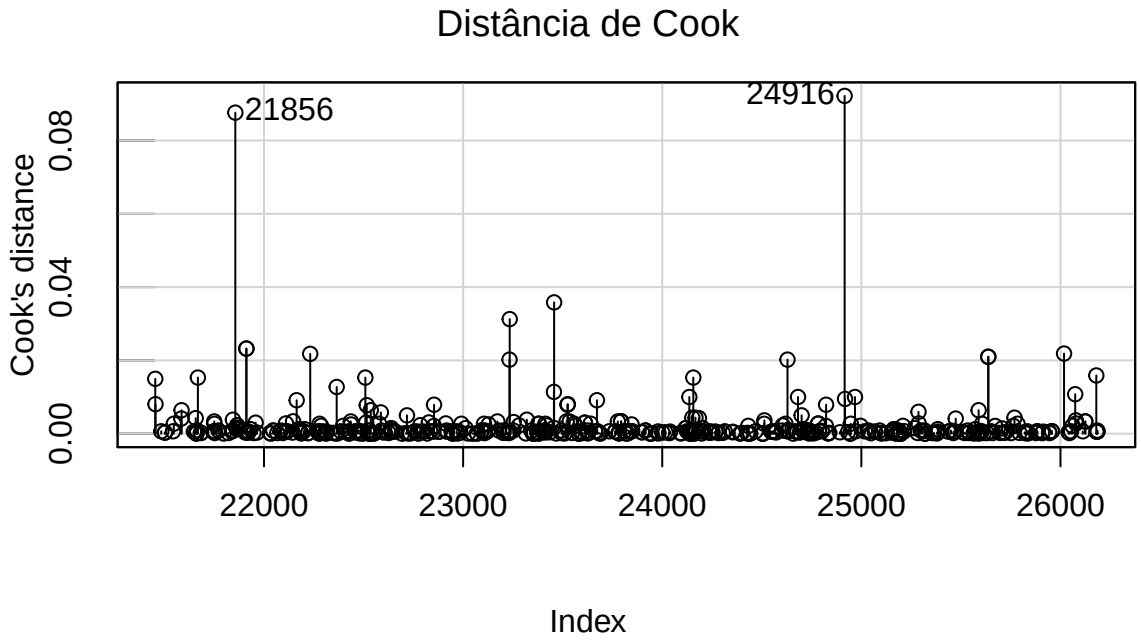
```
hnp(ajuste3.1)
```

```
## Binomial model
```

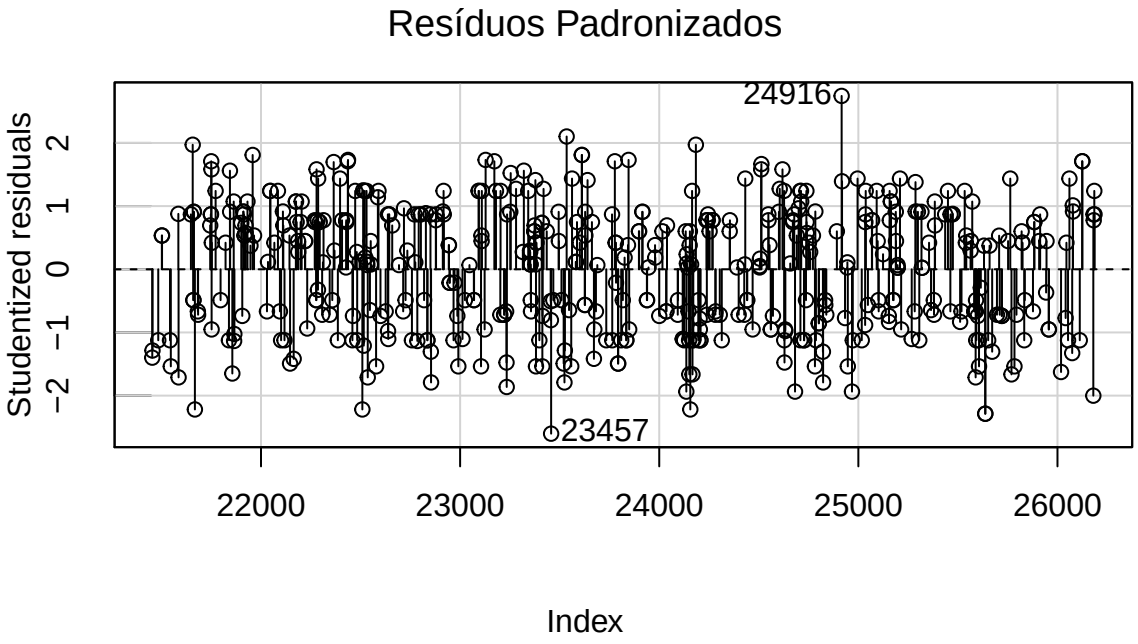


5.4 Medidas de Influencia

```
influenceIndexPlot(ajuste3.1, vars=c("Cook"), main="Distância de Cook")
```




```
influenceIndexPlot(ajuste3.1, vars=c("Studentized"), main="Resíduos Padronizados")
```



5.5 Resíduos Quantílicos Aleatoriazados

5.6 Gráfico Normal de Probabilidades com Envelope Simulado

XXXXX Lineu O gráfico de resíduos simulados permite verificar a adequação do modelo ajustado mesmo que os resíduos não tenham uma aproximação adequada com a distribuição Normal. Neste tipo de gráfico espera-se, para um modelo bem ajustado, os pontos (resíduos) dispersos aleatoriamente entre os limites do envelope.

Deve-se ficar atento à presença de pontos fora dos limites do envelope ou ainda a pontos dentro dos limites porém apresentando padrões sistemáticos.

Vamos utilizar a função envelope implementada pelo professor Cesar Augusto Taconeli :

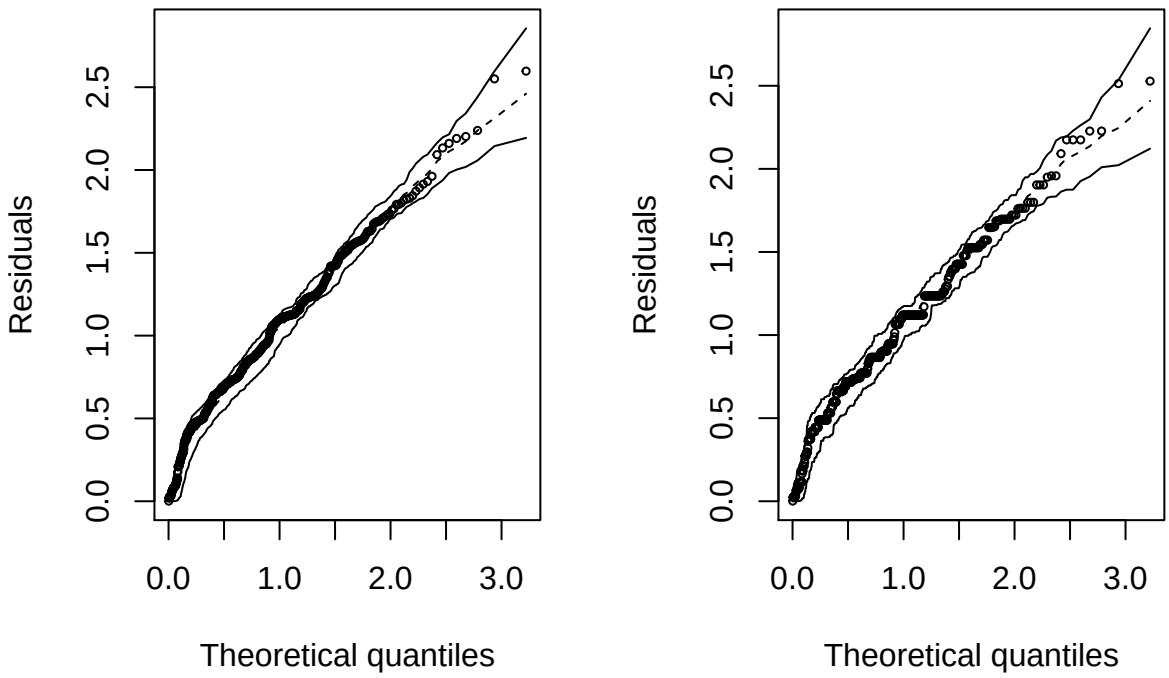
XXXXX

```
par(mfrow=c(1,2))
hnp(ajuste3)
```

```
## Binomial model
```

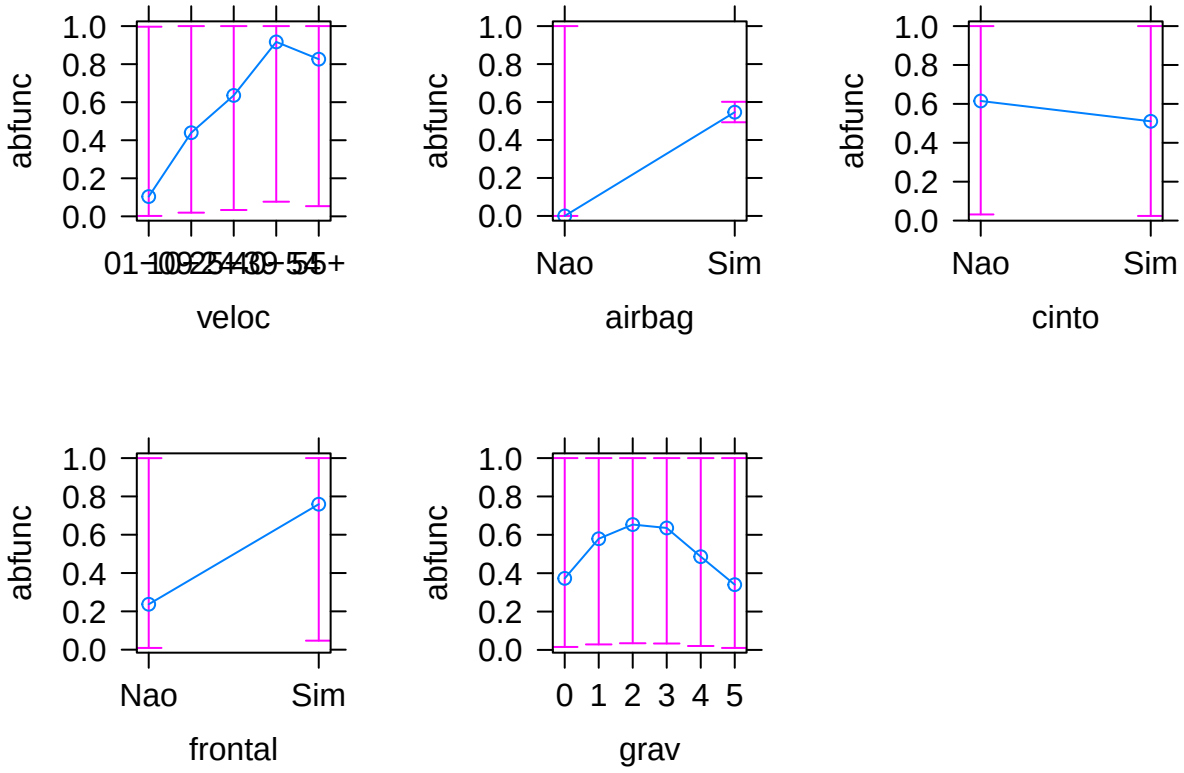
```
hnp(ajuste3.1)
```

```
## Binomial model
```



5.7 Gráficos de Efeitos

```
plot(allEffects(ajuste3.1), type = 'response', main = '')
```



6. PREDIÇÃO

7. AVALIAÇÃO DO PODER PREDITIVO DO MODELO

7.1 Divisão da Base de dados

7.2 Ponto de Corte

7.3 Sensibilidade e Especificidade

7.4 Curva ROC

7.5 Outra Alternativa de validação

8. REFERÊNCIAS