

Trabalho de dados Binários

Acidentes de carro

Laís Hoffmann, Simone Matsubara, Yasmin Fernandes, Willian Meira

2018-11-11

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 possuía 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 variáveis weight, abcat e caseid.

2 Análise Descritiva

2.1 Medidas de Resumo

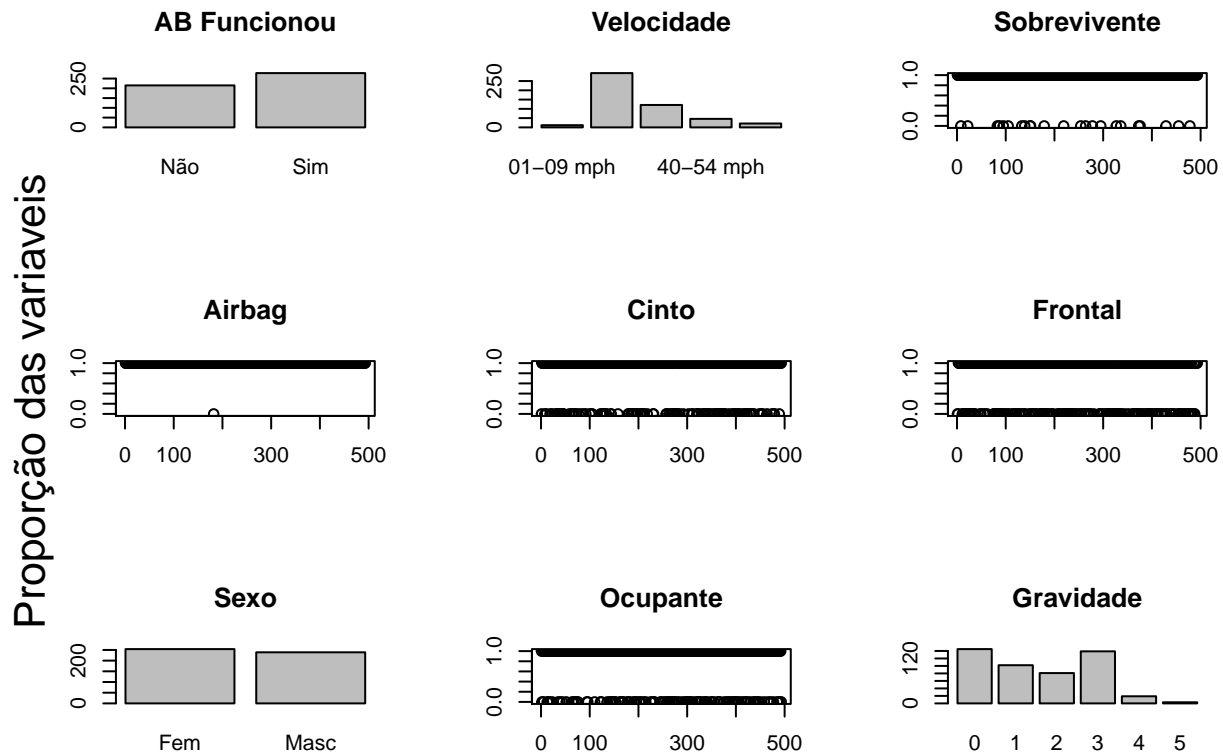
```
summary(dados[, c(1:8,10)])
```

##	veloc	sobrev	airbag	cinto
## 01-09 mph: 12	Min. :0.0000	Min. :0.000	Min. :0.0000	Min. :0.0000
## 10-24 mph:293	1st Qu.:1.0000	1st Qu.:1.000	1st Qu.:1.0000	1st Qu.:1.0000
## 25-39 mph:121	Median :1.0000	Median :1.000	Median :1.0000	Median :1.0000
## 40-54 mph: 46	Mean :0.9533	Mean :0.998	Mean :0.998	Mean :0.7546
## 55+ mph : 21	3rd Qu.:1.0000	3rd Qu.:1.000	3rd Qu.:1.0000	3rd Qu.:1.0000
##	Max. :1.0000	Max. :1.000	Max. :1.000	Max. :1.0000
##				

```
##      frontal      sexo      idade      ocupantes      grav
## Min.   :0.0000   Fem :254   Min.   :16.00   Min.   :0.000   0   :145
## 1st Qu.:0.0000   Masc:239   1st Qu.:23.00   1st Qu.:1.000   1   :102
## Median :1.0000           Median :35.00   Median :1.000   2   : 81
## Mean   :0.6288           Mean   :37.82   Mean   :0.783   3   :139
## 3rd Qu.:1.0000           3rd Qu.:48.00   3rd Qu.:1.000   4   : 19
## Max.   :1.0000           Max.   :93.00   Max.   :1.000   5   :  3
##                                     NA's:  4
```

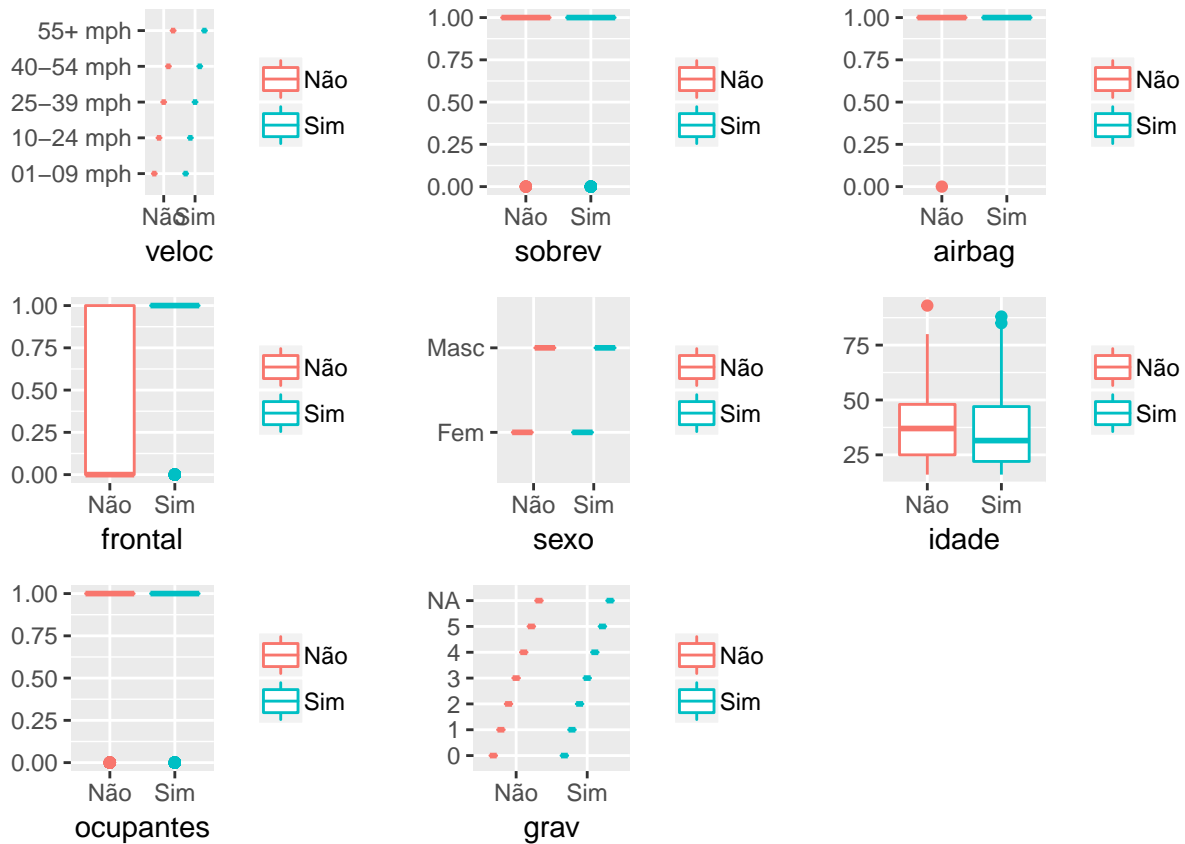
2.3 Histogramas

```
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="Proporção das variaveis",outer=TRUE)
```



2.4 Distribuição

```
grid.arrange(g1, g2, g3, g4, g5, g6, g7, g8, ncol=3, nrow=3)
```



2.5 Análise de correlações entre covariáveis

2.6 Gráficos de Dispersão

```
dis
```

```
## NULL
```

3. AJUSTE DO MODELO DE REGRESSÃO

3.1 Ligação Logito

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

3.2 Ligação Probit

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

3.3 Ligação Complemento log-log

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

3.4 Ligação Cauchy

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

4. ESCOLHA DO MODELO

```
selec <- data.frame(ajuste=c('logito', 'probito', 'cloglog', 'cauchy'),
                    aic=c(AIC(ajuste1), AIC(ajuste2), AIC(ajuste3), AIC(ajuste4)),
                    logLik=c(logLik(ajuste1), logLik(ajuste2), logLik(ajuste3), logLik(ajuste4)))

selec

##      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(ajuste4)

##
## Call:
## glm(formula = abfunc ~ ., family = binomial(link = "cauchit"),
##      data = dados)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.2170  -0.7038   0.4328   0.7398   2.5201
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.794e+05  4.762e+07  -0.004   0.9970
## veloc.L      4.958e+00  3.475e+00   1.427   0.1536
## veloc.Q     -3.154e+00  2.923e+00  -1.079   0.2805
## veloc.C      5.688e-01  1.765e+00   0.322   0.7473
## veloc^4     -8.876e-01  7.550e-01  -1.176   0.2398
## sobrev      -2.501e+00  2.218e+00  -1.127   0.2596
## airbag       1.794e+05  4.762e+07   0.004   0.9970
## cinto       -7.841e-01  3.298e-01  -2.377   0.0174 *
## frontal     3.124e+00  5.108e-01   6.116 9.61e-10 ***
## sexoMasc    -1.285e-01  2.512e-01  -0.511   0.6090
## idade       2.668e-03  7.695e-03   0.347   0.7288
## ocupantes   2.477e-01  3.258e-01   0.760   0.4472
## grav.L      -1.672e+00  1.122e+00  -1.490   0.1362
## grav.Q      -2.051e+00  8.045e-01  -2.549   0.0108 *
## grav.C       1.334e+00  1.291e+00   1.033   0.3014
## grav^4       2.100e+00  1.368e+00   1.535   0.1249
## grav^5       9.909e-01  8.447e-01   1.173   0.2407
## ---
## 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: 480.43 on 472 degrees of freedom
## (4 observations deleted due to missingness)
## AIC: 514.43
##
## Number of Fisher Scoring iterations: 17
```

5.2 Reajuste do Modelo

```
ajuste4.1 <- step(ajuste4, direction = "both")
```

```
## Start: AIC=514.43
## abfunc ~ veloc + sobrev + airbag + cinto + frontal + sexo + idade +
## ocupantes + grav
##
## Df Deviance AIC
## - idade 1 480.53 512.53
## - sexo 1 480.69 512.69
## - ocupantes 1 481.09 513.09
## <none> 480.43 514.43
## - airbag 1 482.53 514.53
## - sobrev 1 482.77 514.77
## - cinto 1 485.56 517.56
## - grav 5 512.76 536.76
## - veloc 4 511.37 537.37
## - frontal 1 588.47 620.47
##
## Step: AIC=512.53
## abfunc ~ veloc + sobrev + airbag + cinto + frontal + sexo + ocupantes +
## grav
##
## Df Deviance AIC
## - sexo 1 480.83 510.83
## - ocupantes 1 481.21 511.21
## <none> 480.53 512.53
## - airbag 1 482.66 512.66
## - sobrev 1 483.27 513.27
## + idade 1 480.43 514.43
## - cinto 1 485.70 515.70
## - grav 5 513.00 535.00
## - veloc 4 512.03 536.03
## - frontal 1 590.35 620.35
##
## Step: AIC=510.83
## abfunc ~ veloc + sobrev + airbag + cinto + frontal + ocupantes +
## grav
##
## Df Deviance AIC
## - ocupantes 1 481.42 509.42
## <none> 480.83 510.83
## - airbag 1 482.91 510.91
## - sobrev 1 483.46 511.46
```

```
## + sexo      1  480.53 512.53
## + idade     1  480.69 512.69
## - cinto     1  485.78 513.78
## - grav      5  514.00 534.00
## - veloc     4  512.16 534.16
## - frontal   1  590.86 618.86
##
## Step: AIC=509.42
## abfunc ~ veloc + sobreviv + airbag + cinto + frontal + grav
##
##           Df Deviance    AIC
## <none>          481.42 509.42
## - airbag      1  483.77 509.77
## - sobreviv    1  484.11 510.11
## + ocupantes   1  480.83 510.83
## + sexo        1  481.21 511.21
## + idade       1  481.26 511.26
## - cinto       1  486.27 512.27
## - grav        5  514.00 532.00
## - veloc       4  512.64 532.64
## - frontal     1  590.96 616.96
```

```
summary(ajuste4.1)
```

```
##
## Call:
## glm(formula = abfunc ~ veloc + sobreviv + airbag + cinto + frontal +
##       grav, family = binomial(link = "cauchit"), data = dados)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.1893  -0.6785   0.4402   0.7507   2.5068
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.794e+05  4.762e+07  -0.004  0.9970
## veloc.L      4.797e+00  3.407e+00   1.408  0.1592
## veloc.Q     -3.030e+00  2.869e+00  -1.056  0.2909
## veloc.C      5.366e-01  1.736e+00   0.309  0.7572
## veloc^4     -8.666e-01  7.454e-01  -1.163  0.2450
## sobreviv    -2.645e+00  2.290e+00  -1.155  0.2480
## airbag       1.794e+05  4.762e+07   0.004  0.9970
## cinto       -7.093e-01  3.192e-01  -2.222  0.0263 *
## frontal      3.066e+00  4.985e-01   6.151 7.71e-10 ***
## grav.L      -1.631e+00  1.139e+00  -1.432  0.1520
## grav.Q     -1.940e+00  7.924e-01  -2.448  0.0144 *
## grav.C      1.413e+00  1.320e+00   1.070  0.2847
## grav^4      2.123e+00  1.406e+00   1.510  0.1311
## grav^5      1.018e+00  8.640e-01   1.179  0.2386
## ---
## 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: 481.42 on 475 degrees of freedom
## (4 observations deleted due to missingness)
## AIC: 509.42
##
## Number of Fisher Scoring iterations: 17
```

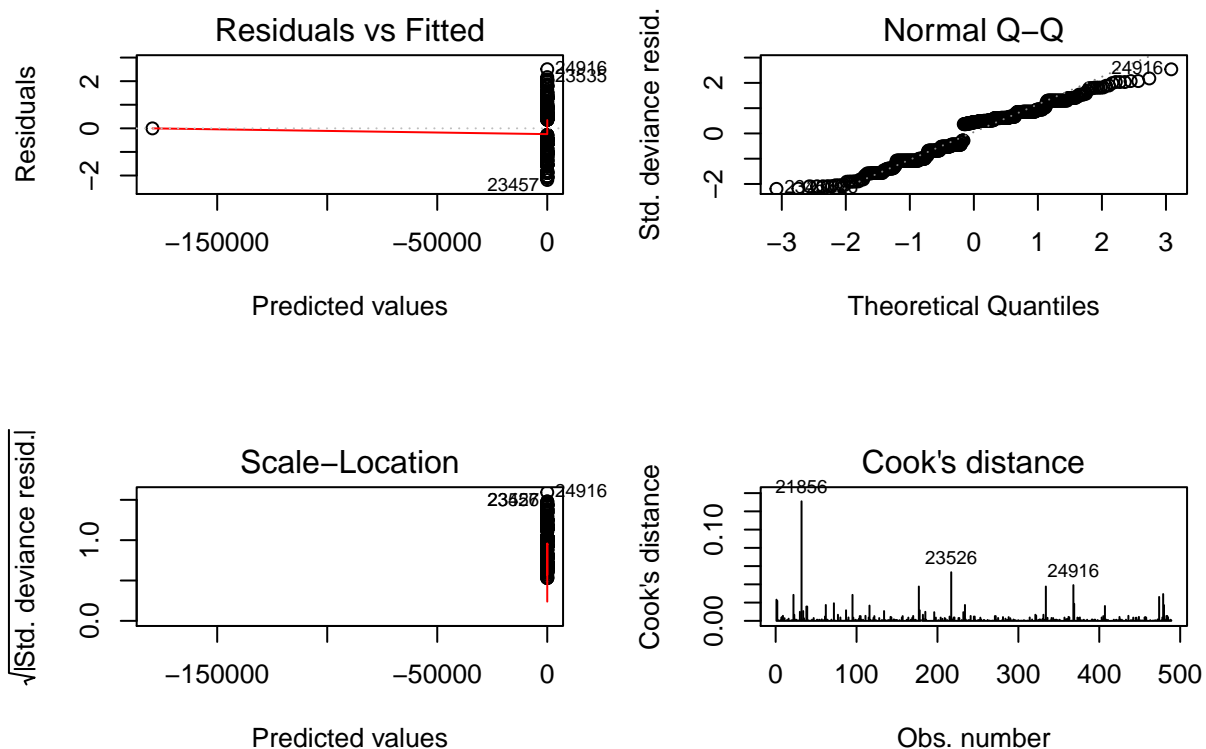
5.3 Análise de Resíduos

```
anova(ajuste4, ajuste4.1, test = 'Chisq')
```

```
## Analysis of Deviance Table
##
## Model 1: abfunc ~ veloc + sobrev + airbag + cinto + frontal + sexo + idade +
## ocupantes + grav
## Model 2: abfunc ~ veloc + sobrev + airbag + cinto + frontal + grav
## Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1 472 480.43
## 2 475 481.42 -3 -0.98817 0.8041
```

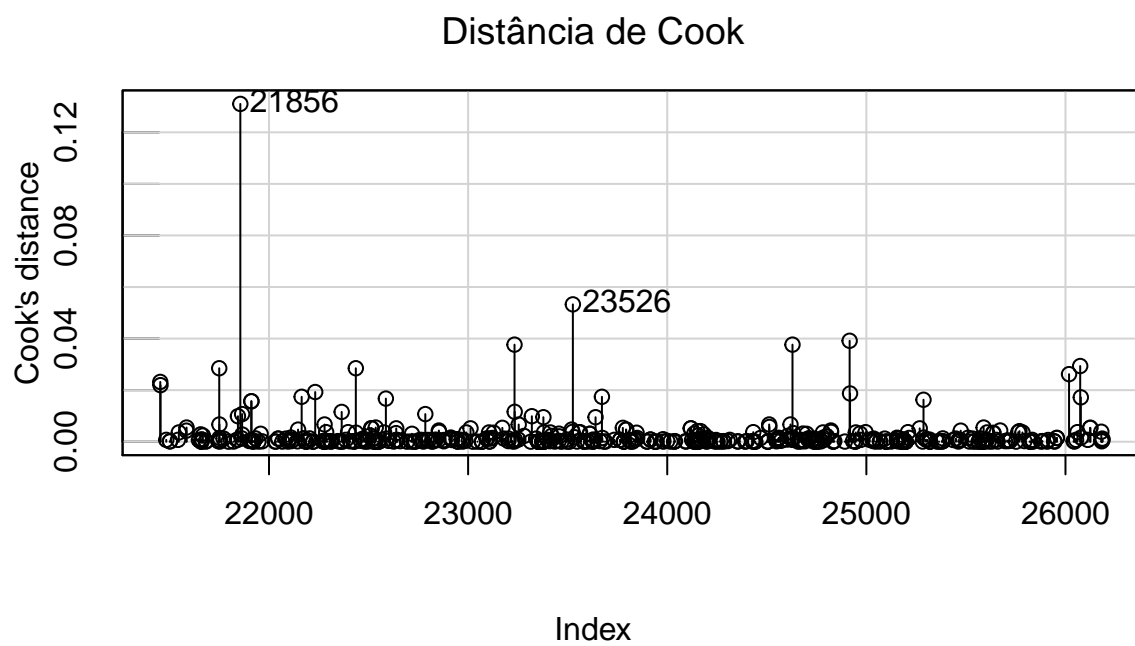
```
par(mfrow=c(2,2))
plot(ajuste4.1, 1:4)
```

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

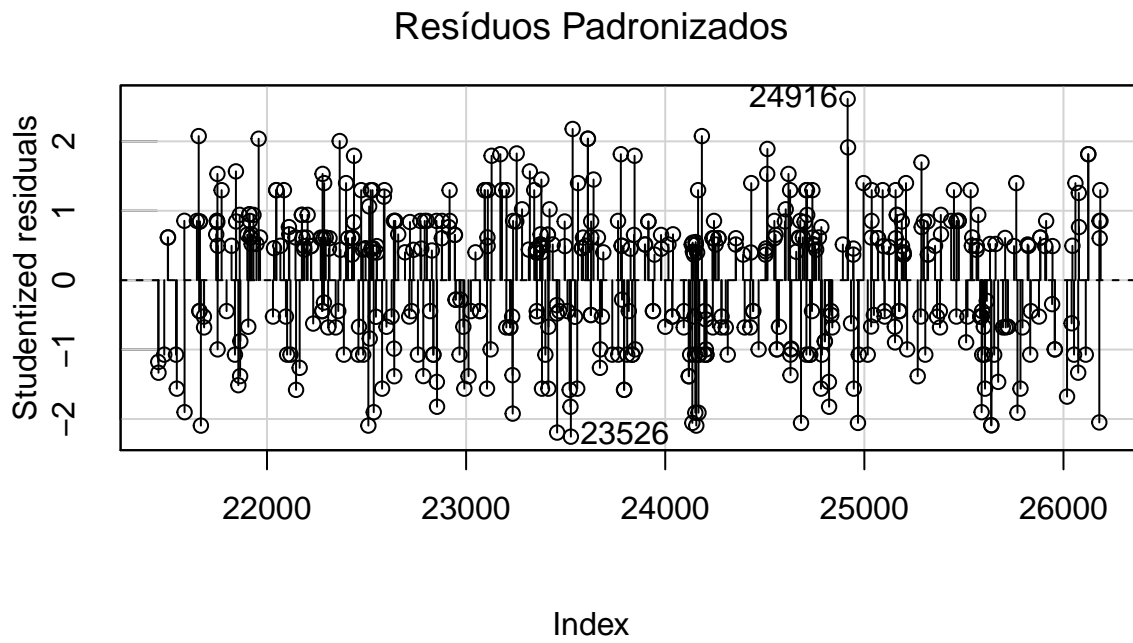


5.4 Medidas de Influência

```
influenceIndexPlot(ajuste4.1, vars=c("Cook"), main="Distância de Cook")
```



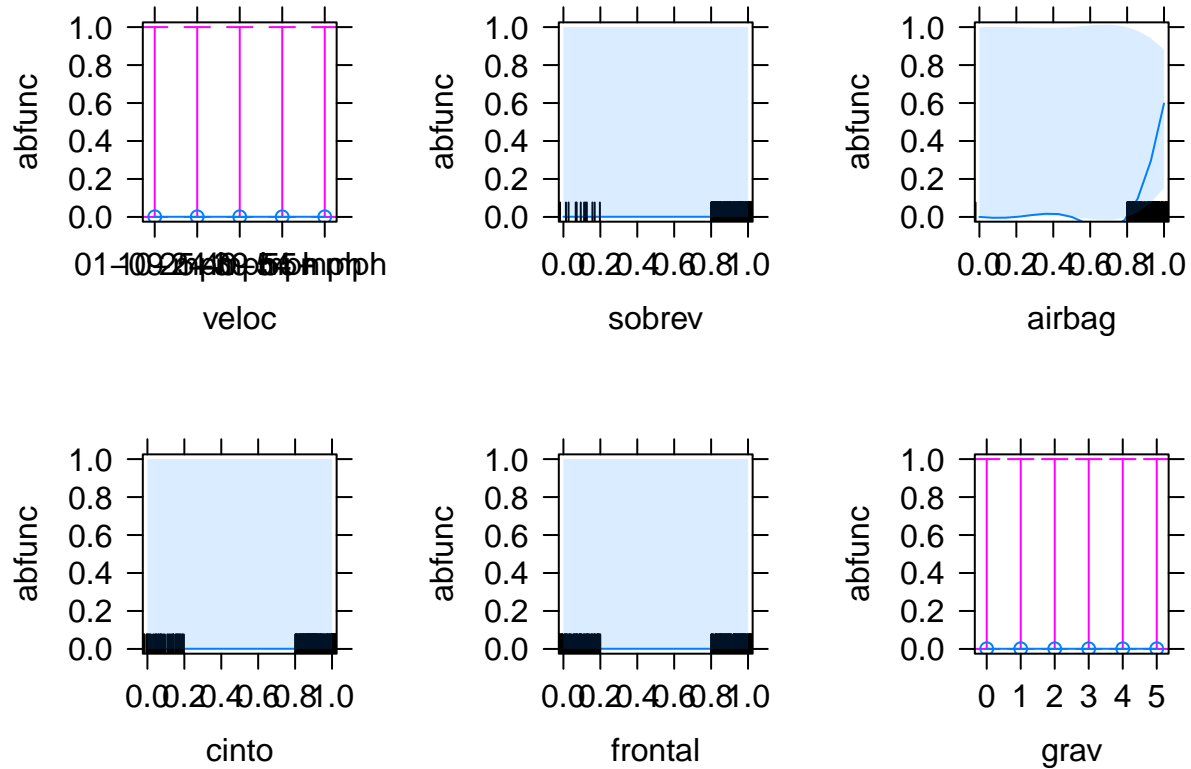
```
influenceIndexPlot(ajuste4.1, vars=c("Studentized"), main="Resíduos Padronizados")
```

5.5 Resíduos Quantílicos Aleatorizados

5.7 Gráficos de Efeitos

```
plot(allEffects(ajuste4.1), 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