trabalho Regressão parte 02

6 de dezembro de 2016

```
## Loading tidyverse: ggplot2
## Loading tidyverse: tibble
## Loading tidyverse: tidyr
## Loading tidyverse: readr
## Loading tidyverse: purrr
## Loading tidyverse: dplyr
```

```
## Conflicts with tidy packages ----------------------------------------------
```

```
## arrange():   dplyr, plyr
## compact():   purrr, plyr
## count():     dplyr, plyr
## failwith():  dplyr, plyr
## filter():    dplyr, stats
## id():        dplyr, plyr
## lag():       dplyr, stats
## mutate():    dplyr, plyr
## rename():    dplyr, plyr
## summarise(): dplyr, plyr
## summarize(): dplyr, plyr
```

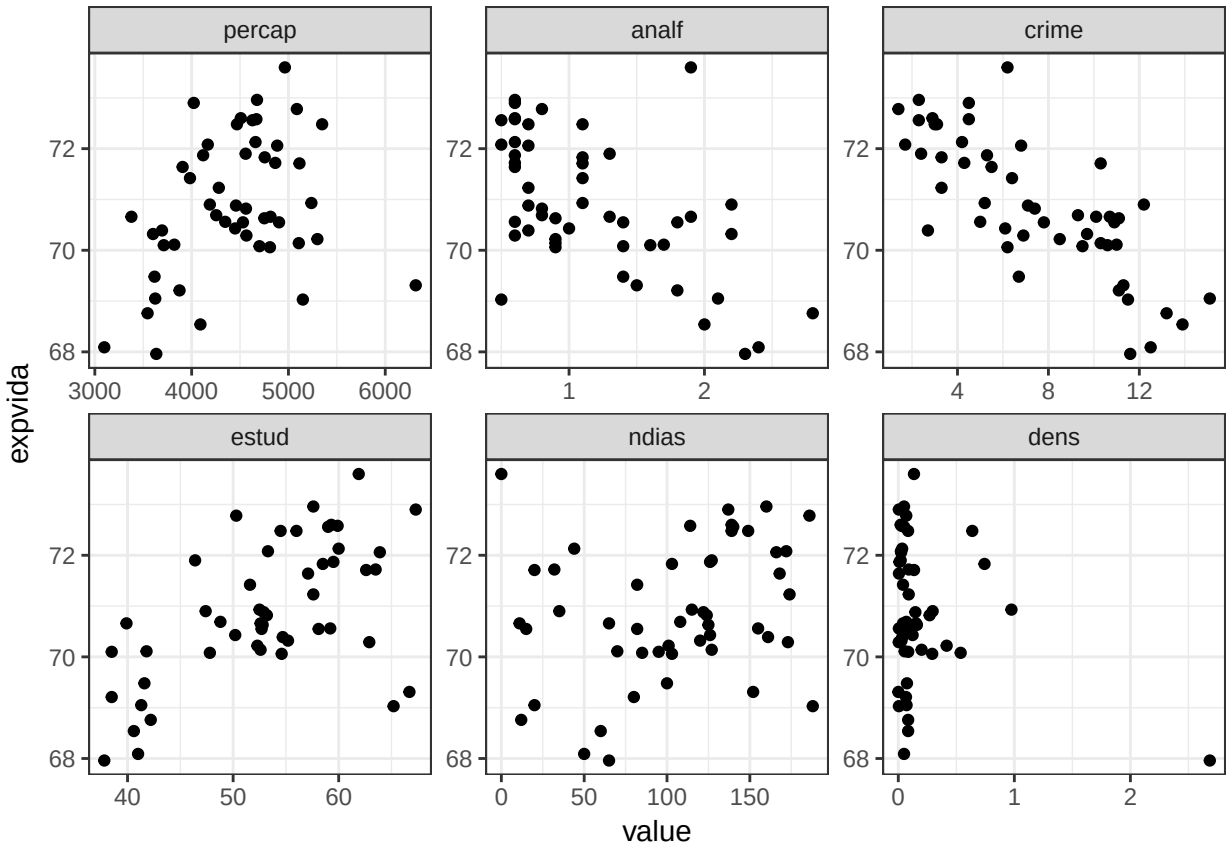
2.Análise Descritiva

```
##
## Attaching package: 'reshape2'
```

```
## The following object is masked from 'package:tidyr':
##
##     smiths
```

variable	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
percap	3098.0000	3993.0000	4519.000	4436.000	4814.0000	6315.000
analf	0.5000	0.6250	0.950	1.170	1.5750	2.800
crime	1.4000	4.3500	6.850	7.378	10.6800	15.100
estud	37.8000	48.0500	53.250	53.110	59.1500	67.300
ndias	0.0000	66.2500	114.500	104.500	139.8000	188.000
dens	0.0006	0.0277	0.069	0.188	0.1443	2.684
expvida	67.9600	70.1200	70.680	70.880	71.8900	73.600

Dado a natureza quantitativa dos dados, é de grande interesse que identifiquemos as relações entre as covariáveis explicativas e a variável resposta, a fim de tornar essas relações visuais, foi construido graficos de dispersão entre a variável resposta e entre todas as covariáveis de interesse:



A partir da FiguraXX acima, podemos identificar visualmente a relação supracitada, a qual identificamos que todas as covariáveis parecem ter uma relação linear significativa, em especial, identificamos uma "forte"(mudar isso kkk) relação linear negativa da variável resposta com as covariáveis "analf" e "crime", assim como uma significativa relação linear positiva da variável resposta com as covariáveis "percap","estud","ndias" e "dens" (a menos a um ponto).

Análise Inferêncial

Dada a existência de relações lineares entre a variável resposta e todas as covariáveis explicativas e visando poder comparar os coeficiêntes de diferentes tipos de unidades de dados optamos por iniciar a análise com o seguinte modelo:

Modelo:

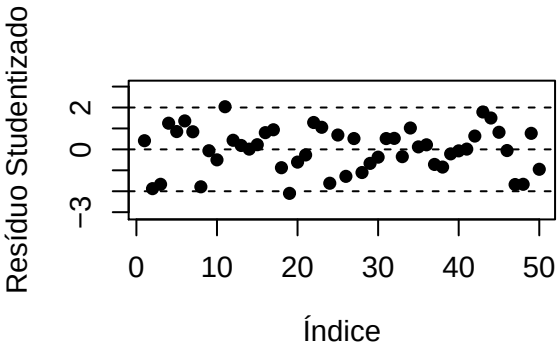
$$Y_i = \beta_0 + \sum_{j=1}^{6} \beta_j \left(\frac{x_{ji} - \bar{x_j}}{s_j} \right) + \varepsilon_i \begin{cases} i = 1, \ldots, 50 \\ j = 1, \ldots, 6 \end{cases}$$

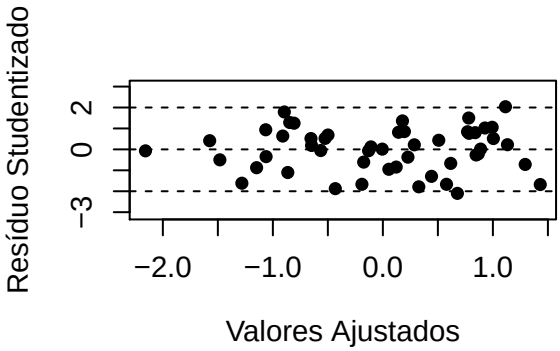
em que $\varepsilon_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$, $x_j = \frac{1}{50} \sum_{i=1}^{50} x_j i$ e $s_j = \sqrt{\frac{1}{50} \sum_{i=1}^{50} (x_{ji} - \bar{x_j})^2}$

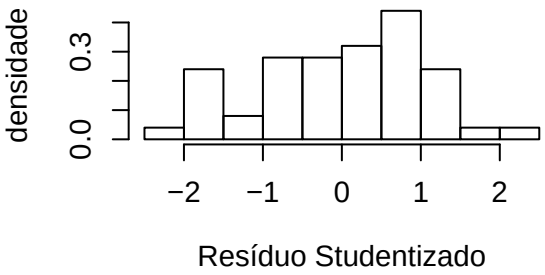
Y_i : Expectativa de vida em anos (1969-70).
β_0 : Expectativa de vida esperada em anos (1969-70) para valores de covariáveis iguais às suas respectivas médias.
$\frac{\beta_1}{s_1}$: Incremento na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "renda percapita (em 1974 em USD)" em uma unidade, mantendo-se as demais covariáveis fixas.
$\frac{\beta_2}{s_2}$: Incremento na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "proporção de analfabetos (em 1970)" em uma unidade, mantendo-se as demais covariáveis fixas.

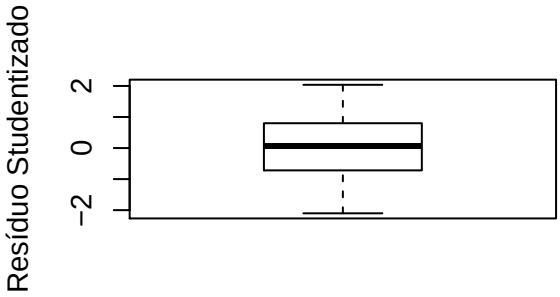
$\frac{\beta_3}{s_3}$: Incremento na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "taxa de criminalidade (por 100000 habitantes 1976)" em uma unidade, mantendo-se as demais covariáveis fixas.
$\frac{\beta_4}{s_4}$: Incremento na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "porcentagem de estudantes que concluem o segundo grau (1970)" em uma unidade, mantendo-se as demais covariáveis fixas.
$\frac{\beta_5}{s_5}$: Incremento na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor de "número de dias do ano com temperatura abaixo de zero grau Celsus na cidade mais importante do estado" em uma unidade, mantendo-se as demais covariáveis fixas.
$\frac{\beta_6}{s_6}$: Incremento na expectativa de vida em anos (1969-70) esperada quando se aumenta o valor da "densidade da população estimada em julho de 1975 por área do estado em milhas quadradas" em uma unidade, mantendo-se as demais covariáveis fixas.

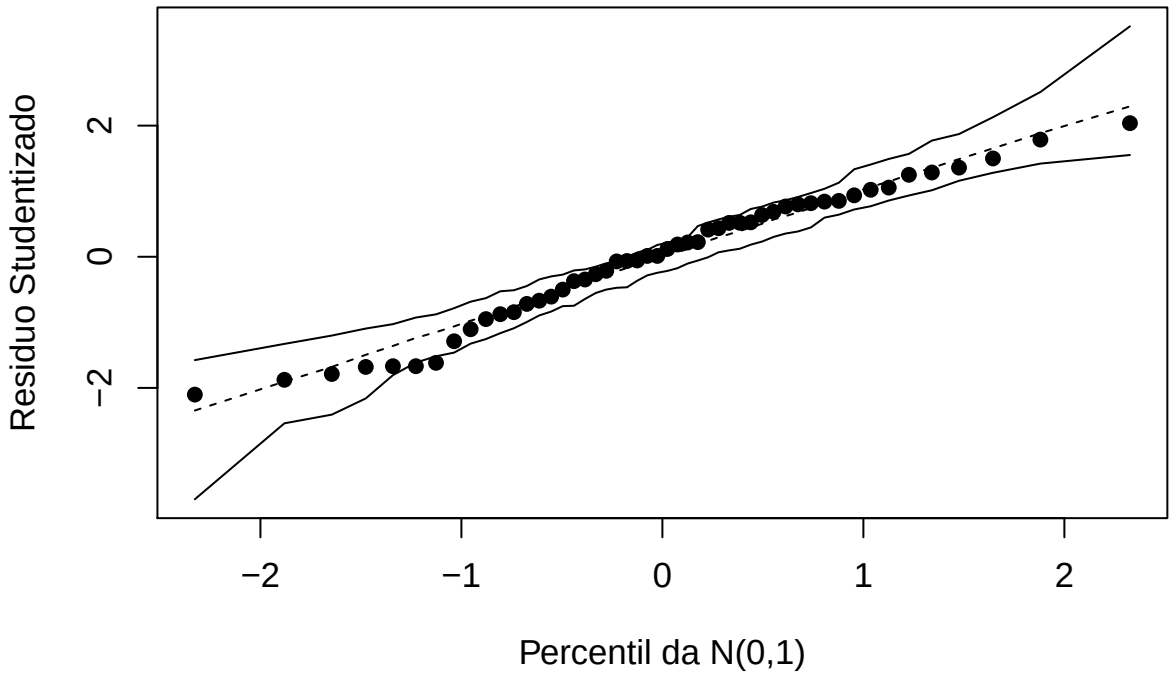
```
##
## Call:
## lm(formula = expvida ~ percap + analf + crime + estud + ndias +
##     dens, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.04132 -0.35980  0.03567  0.41991  0.91309
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.623e-15  7.740e-02   0.000  1.00000
## percap       1.007e-01  1.046e-01   0.963  0.34090
## analf       -2.398e-02  1.497e-01  -0.160  0.87352
## crime       -7.910e-01  1.137e-01  -6.953  1.5e-08 ***
## estud        1.829e-01  1.246e-01   1.468  0.14946
## ndias       -2.910e-01  1.076e-01  -2.706  0.00973 **
## dens        -1.571e-01  8.481e-02  -1.852  0.07093 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5473 on 43 degrees of freedom
## Multiple R-squared:  0.7372, Adjusted R-squared:  0.7005
## F-statistic:  20.1 on 6 and 43 DF,  p-value: 4.943e-11
```


Resíduo Studentizado
−3
0
2
0
10
20
30
40
50
Índice


Resíduo Studentizado
−3
0
2
−2.0
−1.0
0.0
1.0
Valores Ajustados


densidade
0.0
0.3
−2
−1
0
1
2
Resíduo Studentizado


Resíduo Studentizado
−2
0
2



Seleção dos modelos

Dada a natureza dos dados

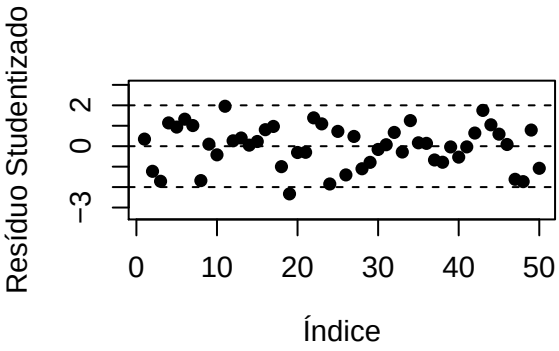
```
## Start:  AIC=0.99
## expvida ~ 1
##
##          Df Sum of Sq    RSS     AIC
## + crime   1   29.8763 19.124 -44.055
## + analf   1   16.9690 32.031 -18.266
## + estud   1   16.6098 32.390 -17.708
## + percap  1    5.6729 43.327  -3.162
## + ndias   1    3.3653 45.635  -0.568
## + dens    1    3.3323 45.668  -0.532
## <none>                49.000   0.990
##
## Step:  AIC=-44.05
## expvida ~ crime
##
##          Df Sum of Sq    RSS     AIC
## + estud   1    2.6032 16.521 -49.371
## + ndias   1    1.7395 17.384 -46.823
## + dens    1    1.5034 17.620 -46.149
## + percap  1    1.3345 17.789 -45.671
## <none>                19.124 -44.055
```

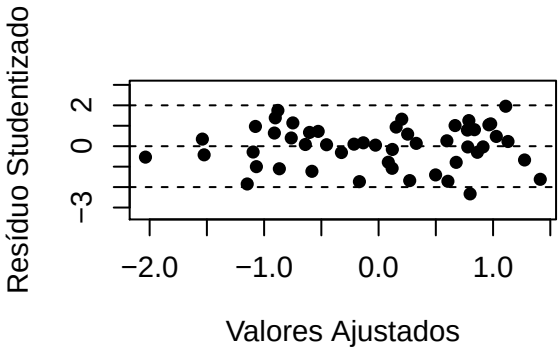
```
## + analf   1    0.1516 18.972 -42.453
## - crime   1   29.8763 49.000   0.990
##
## Step:  AIC=-49.37
## expvida ~ crime + estud
##
##          Df Sum of Sq    RSS     AIC
## + ndias   1    2.4410 14.080 -55.365
## + dens    1    0.7343 15.786 -49.644
## <none>                16.521 -49.371
## + analf   1    0.2452 16.275 -48.119
## + percap  1    0.0567 16.464 -47.543
## - estud   1    2.6032 19.124 -44.055
## - crime   1   15.8696 32.390 -17.708
##
## Step:  AIC=-55.37
## expvida ~ crime + estud + ndias
##
##          Df Sum of Sq    RSS     AIC
## + dens    1    0.8913 13.188 -56.635
## <none>                14.080 -55.365
## + percap  1    0.1012 13.978 -53.726
## + analf   1    0.0954 13.984 -53.705
## - ndias   1    2.4410 16.521 -49.371
## - estud   1    3.3047 17.384 -46.823
## - crime   1   18.1773 32.257 -15.915
##
## Step:  AIC=-56.63
## expvida ~ crime + estud + ndias + dens
##
##          Df Sum of Sq    RSS     AIC
## <none>                13.188 -56.635
## + percap  1    0.3020 12.886 -55.793
## - dens    1    0.8913 14.080 -55.365
## + analf   1    0.0319 13.156 -54.756
## - estud   1    2.3831 15.571 -50.329
## - ndias   1    2.5980 15.786 -49.644
## - crime   1   18.5026 31.691 -14.800

## Start:  AIC=-53.82
## expvida ~ percap + analf + crime + estud + ndias + dens
##
##          Df Sum of Sq    RSS     AIC
## - analf   1    0.0077 12.886 -55.793
## - percap  1    0.2778 13.156 -54.756
## <none>                12.879 -53.823
## - estud   1    0.6452 13.524 -53.379
## - dens    1    1.0270 13.906 -51.987
## - ndias   1    2.1928 15.071 -47.961
## - crime   1   14.4812 27.360 -18.148
##
## Step:  AIC=-55.79
## expvida ~ percap + crime + estud + ndias + dens
##
```

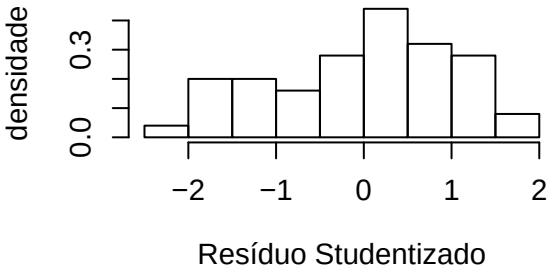
```
##          Df Sum of Sq    RSS     AIC
## - percap  1    0.3020 13.188 -56.635
## <none>                 12.886 -55.793
## - estud   1    0.7637 13.650 -54.914
## - dens    1    1.0921 13.978 -53.726
## - ndias   1    2.7034 15.590 -48.271
## - crime   1   18.8044 31.691 -12.800
##
## Step:  AIC=-56.63
## expvida ~ crime + estud + ndias + dens
##
##         Df Sum of Sq    RSS     AIC
## <none>               13.188 -56.635
## - dens   1    0.8913 14.080 -55.365
## - estud  1    2.3831 15.571 -50.329
## - ndias  1    2.5980 15.786 -49.644
## - crime  1   18.5026 31.691 -14.800
```

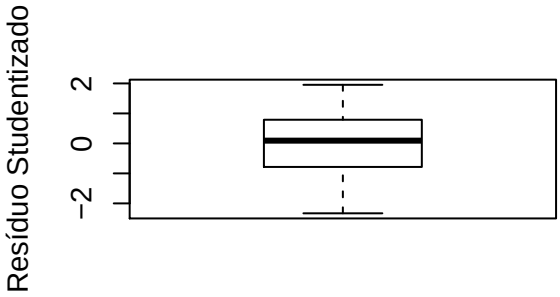
Com base no stepwise, chegamos ao modelo com maior poder preditivo(expvida ~ crime + estud + ndias + dens):

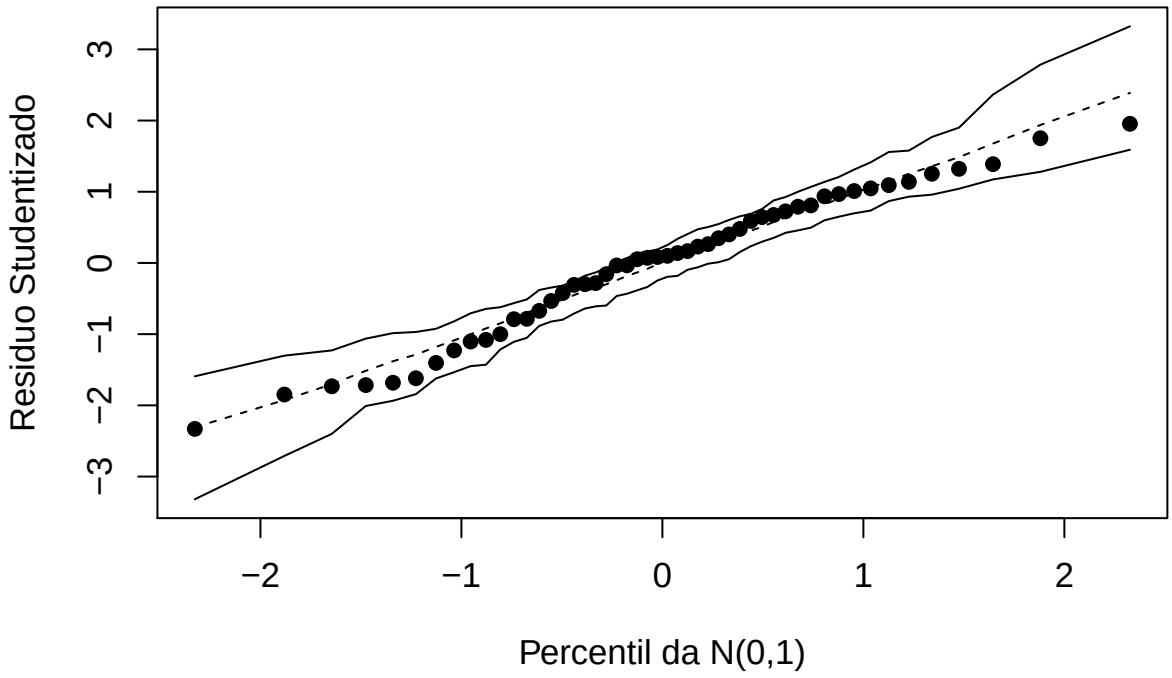
```
##
## Call:
## lm(formula = expvida ~ crime + estud + ndias + dens, data = dados)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.16349 -0.39785  0.04853  0.40958  0.91683
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.688e-15  7.656e-02   0.000  1.00000
## crime       -7.861e-01  9.894e-02  -7.946 4.14e-10 ***
## estud        2.626e-01  9.210e-02   2.852  0.00655 **
## ndias       -2.765e-01  9.288e-02  -2.977  0.00467 **
## dens        -1.402e-01  8.039e-02  -1.744  0.08800 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5414 on 45 degrees of freedom
## Multiple R-squared:  0.7309, Adjusted R-squared:  0.7069
## F-statistic: 30.55 on 4 and 45 DF,  p-value: 2.609e-12
```

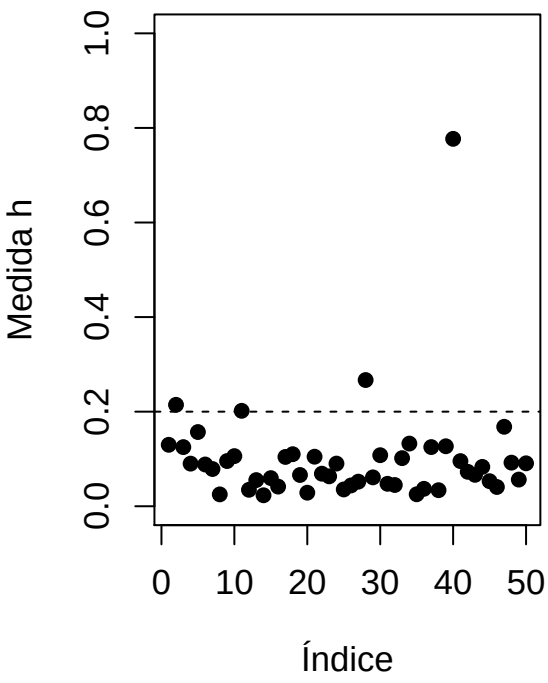

Resíduo Studentizado
−3
0
2
0
10
20
30
40
50
Índice

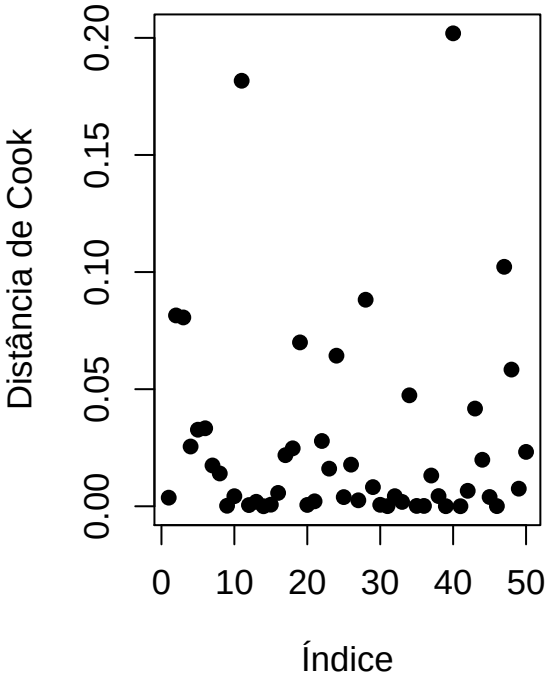

Resíduo Studentizado
−3
0
2
−2.0
−1.0
0.0
1.0
Valores Ajustados


densidade
0.0
0.3
−2
−1
0
1
2
Resíduo Studentizado


Resíduo Studentizado
−2
0
2


Residuo Studentizado
−3
−2
−1
0
1
2
3
−2
−1
0
1
2
Percentil da N(0,1)


Medida h
0.0
0.2
0.4
0.6
0.8
1.0
0
10
20
30
40
50
Índice


Distância de Cook
0.00
0.05
0.10
0.15
0.20
0
10
20
30
40
50
Índice