Atividade Prática

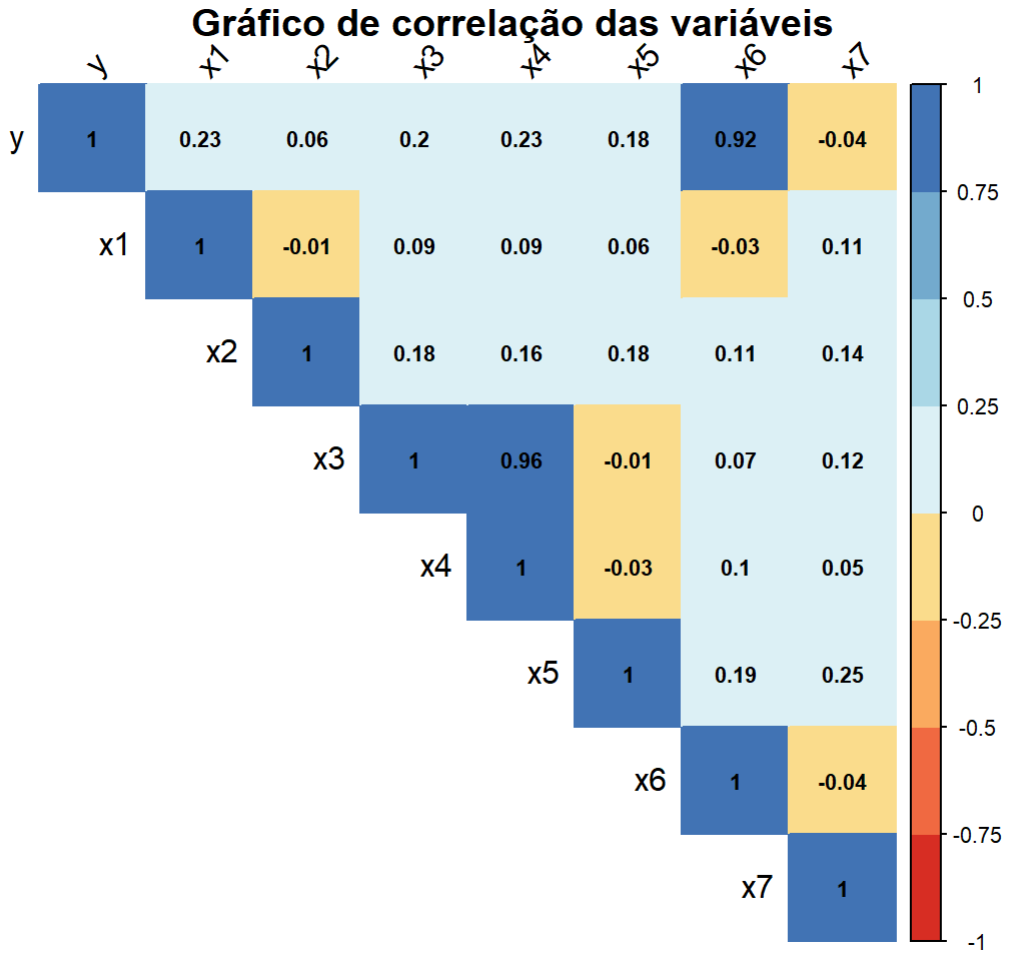
AUTHOR
Leonardo Prior Migliorini

Análise descritiva dos dados

Nosso banco de dados tem 107 observações e 8 variáveis, sendo uma delas a variável de interesse e as demais, covariáveis para incluirmos nos ajustes dos modelos. Primeiramente, vamos verificar quais os tipos de covariáveis temos em nosso banco de dados.

```
       y                x1               x2               x3        
 Min.   : 43.90   Min.   :0.0000   Min.   : 24.32   Min.   :0.0860  
 1st Qu.: 85.96   1st Qu.:0.0000   1st Qu.: 44.51   1st Qu.:0.3090  
 Median :105.88   Median :1.0000   Median : 55.05   Median :0.4270  
 Mean   :111.00   Mean   :0.6168   Mean   : 52.80   Mean   :0.4466  
 3rd Qu.:127.98   3rd Qu.:1.0000   3rd Qu.: 59.79   3rd Qu.:0.5615  
 Max.   :259.91   Max.   :1.0000   Max.   :102.59   Max.   :0.9240  
       x4               x5               x6               x7        
 Min.   :0.822   Min.   : 3.00   Min.   : 0.000   Min.   : 74.92  
 1st Qu.:3.163   1st Qu.: 8.00   1st Qu.: 0.840   1st Qu.: 93.78  
 Median :4.422   Median :11.00   Median : 2.490   Median :100.89  
 Mean   :4.517   Mean   :10.47   Mean   : 3.613   Mean   :100.44  
 3rd Qu.:5.495   3rd Qu.:13.00   3rd Qu.: 5.415   3rd Qu.:106.93  
 Max.   :9.068   Max.   :21.00   Max.   :18.767   Max.   :146.85  
```

Podemos observar que a covariável x_1 é do tipo *dummy*, enquanto que as demais são todas contínuas positivas.



Através do *corrplot* acima, nota-se que a covariável x_6 é a que apresenta a maior correlação com a variável de interesse. Percebe-se também um provável problema de mutlicolinearidade aproximada entre as covariáveis x_3 e x_4 pela sua alta correlação de $0,96$.

Modelo incialmente ajustado

Inicialmente consideraremos um modelo ajustado com todas as covariáveis disponíveis em nosso banco de dados. O ajuste é dado abaixo:

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.076  -6.026  -0.214   7.436  21.794 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 74.46224   10.18764   7.309 7.02e-11 ***
x1          18.99352    2.03183   9.348 2.89e-15 ***
x2          -0.15909    0.08117  -1.960   0.0528 .  
x3          16.22563   20.72954   0.783   0.4357    
x4           1.12853    1.99299   0.566   0.5725    
x5           0.06316    0.28217   0.224   0.8233    
x6          10.02365    0.30531  32.831  < 2e-16 ***
```

```
x7          -0.15932    0.10050  -1.585   0.1161
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 10.08 on 99 degrees of freedom
Multiple R-squared:  0.928, Adjusted R-squared:  0.9229
F-statistic: 182.3 on 7 and 99 DF,  p-value: < 2.2e-16
```

Podemos observar que algumas das covariáveis consideradas não foram significativas no modelo inicial, portanto, utilizaremos a função `step` para nos indicar um modelo melhor. A saída da função, assim como o modelo sugerido são dados abaixo:

```
Start:  AIC=502.08
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7

       Df Sum of Sq    RSS    AIC
- x5    1         5  10058 500.13
- x4    1        33  10086 500.43
- x3    1        62  10115 500.74
<none>               10053 502.08
- x7    1       255  10308 502.76
- x2    1       390  10443 504.15
- x1    1      8874  18927 567.78
- x6    1    109454 119507 764.96

Step:  AIC=500.13
y ~ x1 + x2 + x3 + x4 + x6 + x7

       Df Sum of Sq    RSS    AIC
- x4    1        32  10090 498.47
- x3    1        62  10121 498.80
<none>               10058 500.13
- x7    1       252  10310 500.78
- x2    1       385  10444 502.16
- x1    1      8925  18983 566.10
- x6    1    114481 124539 767.37

Step:  AIC=498.47
y ~ x1 + x2 + x3 + x6 + x7

       Df Sum of Sq    RSS    AIC
<none>               10090 498.47
- x7    1       309  10399 499.70
- x2    1       391  10481 500.54
- x3    1      2230  12320 517.84
- x1    1      9005  19095 564.73
- x6    1    115653 125743 766.40


Call:
lm(formula = y ~ x1 + x2 + x3 + x6 + x7, data = dados)

Coefficients:
(Intercept)           x1           x2           x3           x6           x7
    75.7074      19.0782      -0.1575      27.3825      10.0518      -0.1664
```

Podemos ver que as covariáveis x_4 e x_5 foram desconsideradas no novo ajuste. Seguimos então, ajustando esse novo modelo e posteriormente realizando a análise de diagnósticos para verificar a adequação do ajuste.

Modelo reajustado

```
Call:
lm(formula = y ~ x1 + x2 + x3 + x6 + x7, data = dados)

Residuals:
     Min       1Q   Median       3Q      Max
-21.8872  -6.3075  -0.2461   7.7121  21.9382

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 75.70742    9.85719   7.680 1.05e-11 ***
x1          19.07824    2.00942   9.494 1.17e-15 ***
x2          -0.15752    0.07966  -1.977   0.0507 .
x3          27.38254    5.79553   4.725 7.46e-06 ***
x6          10.05183    0.29543  34.025  < 2e-16 ***
x7          -0.16636    0.09461  -1.758   0.0817 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.995 on 101 degrees of freedom
Multiple R-squared:  0.9277,    Adjusted R-squared:  0.9241
F-statistic: 259.3 on 5 and 101 DF,  p-value: < 2.2e-16
```

Note que algumas covariáveis não são muito significativas para o modelo, entretanto, prosseguiremos para a análise de diagnósticos a fim de verificar se as suposições do modelo foram satisfeitas e, consequentemente, se os resultados dos testes de hipóteses não sofreram nenhuma distorção devido a desvios de normalidade. Além disso, iremos verificar se não há pontos possivelmente influentes no modelo.

Análise de diagnóstico

Suposições do modelo

Vamos começar testando as suposições do modelo para verificar se os resultados dos testes de hipóteses são confiáveis.

Testanto [S0]:

```
    RESET test

data:  fit2
RESET = 0.20546, df1 = 2, df2 = 99, p-value = 0.8146
```

Testanto [S1]:

```
    One Sample t-test

data:  resid(fit2)
t = -5.7785e-16, df = 106, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.869971  1.869971
sample estimates:
    mean of x
-5.450274e-16
```

Testanto [S2]:

```
    studentized Breusch-Pagan test

data:  fit2
BP = 9.9212, df = 5, p-value = 0.0775
```

Testanto [S3]:

```
    Durbin-Watson test

data:  fit2
DW = 2.1593, p-value = 0.7949
alternative hypothesis: true autocorrelation is greater than 0
```

Testanto [S4]:

```
      x1       x2       x3       x6       x7
1.022151 1.061339 1.055285 1.019242 1.045772
```

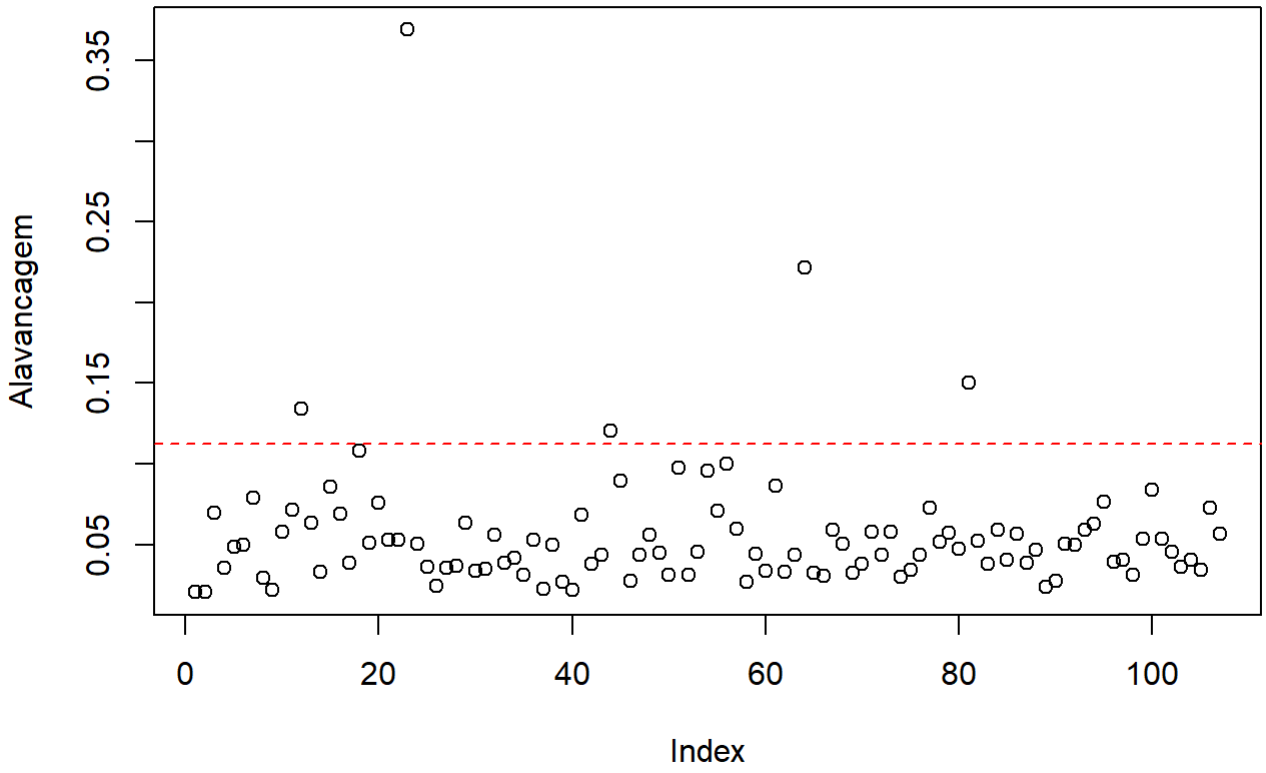
Testanto [S5]:

```
    Jarque Bera Test

data:  resid(fit2)
X-squared = 0.79299, df = 2, p-value = 0.6727
```

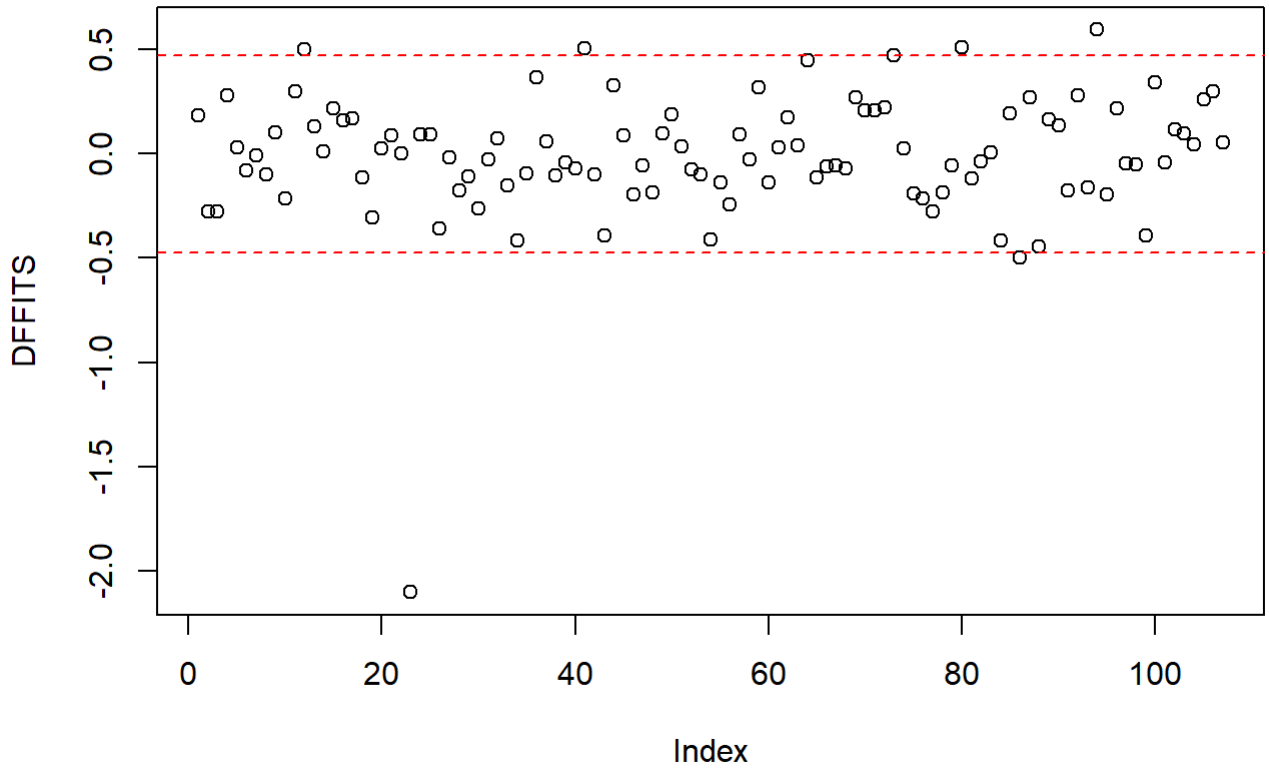
Note que todos os testes de hipóstese obtiveram um $P-valor$ maior que $\alpha = 0,05$. Logo, todos os testes não rejeitaram H_0, ou seja, as suposições do modelo foram satisfeitas. Em relação aos VIFs, como todos se encontram muito próximos de 1, conclui-se que o modelo não tem problemas de multicolinearidade aproximada, ou seja, $[S_4]$ também é satisfeita.

Alavancagem



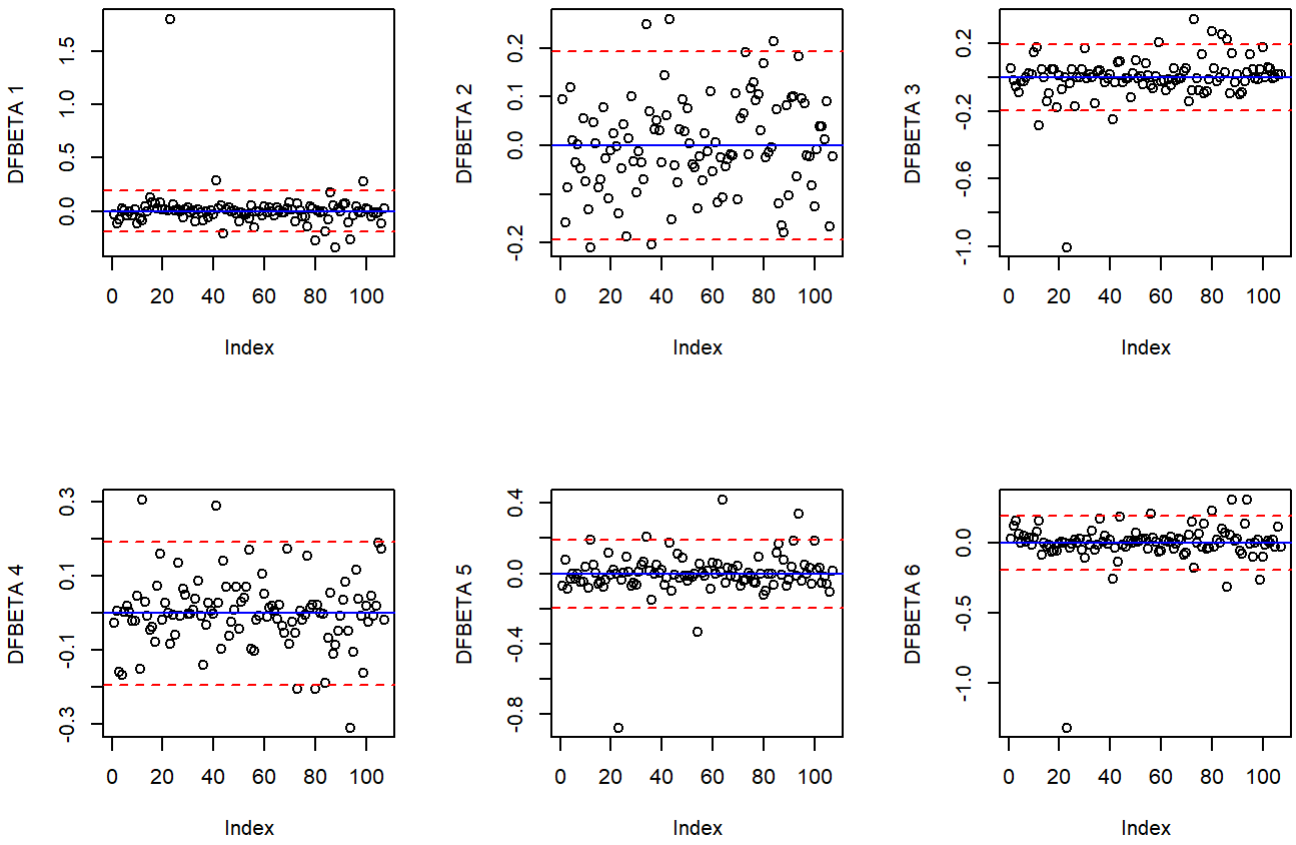
No gráfico acima, podemos notar alguns pontos de alavanca, sendo a obrservação 23 a mais discrepante das demais.

DFFITS



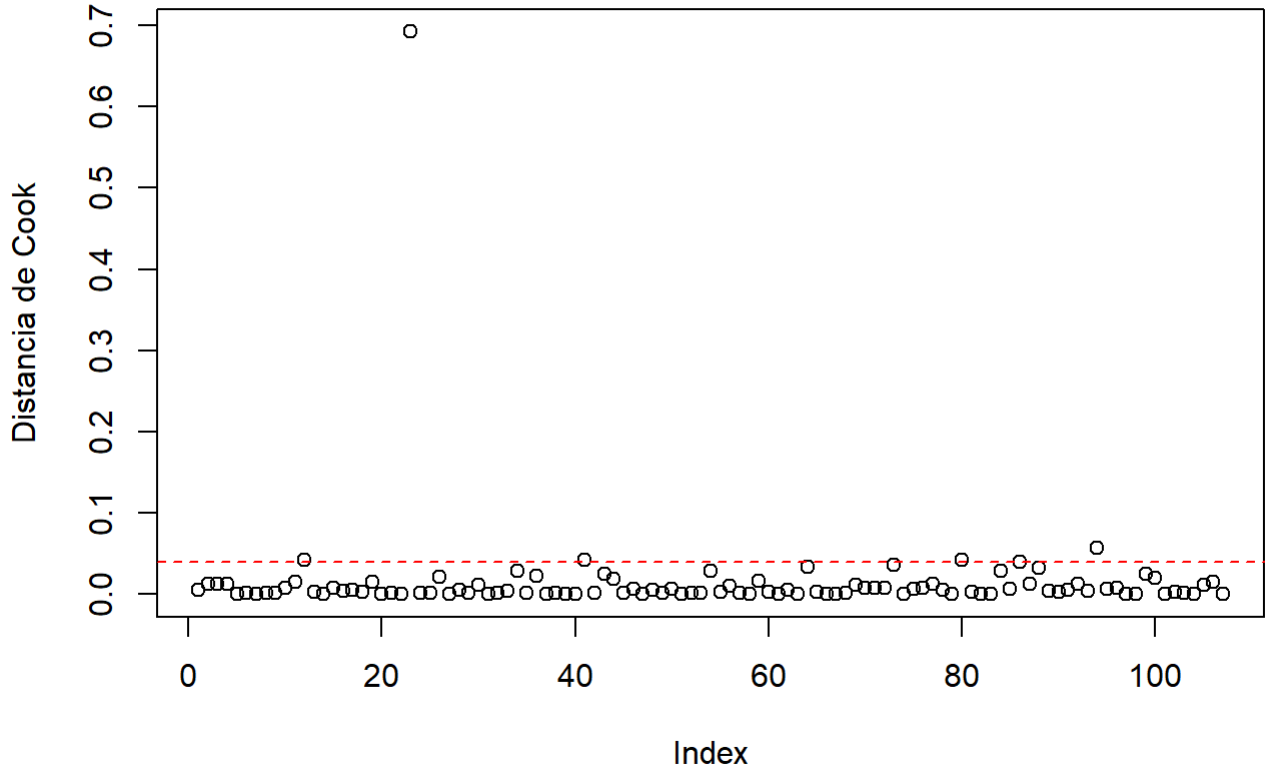
Podemos ver que a observação 23 também tem alta influência sobre seu próprio valor ajustado, nesse caso, também se encontrando bem dispersa das demais observações.

DFBETAS



Podemos notar que a observação 23 também é muito influente sobre $\beta 1$, $\beta 3$, $\beta 5$ e $\beta 6$. Para os demais parâmetros de regressão, não observamos pontos com comportamento muito influente.

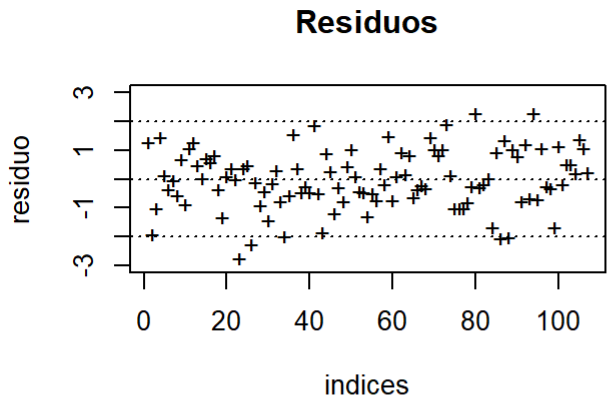
Distância de Cook

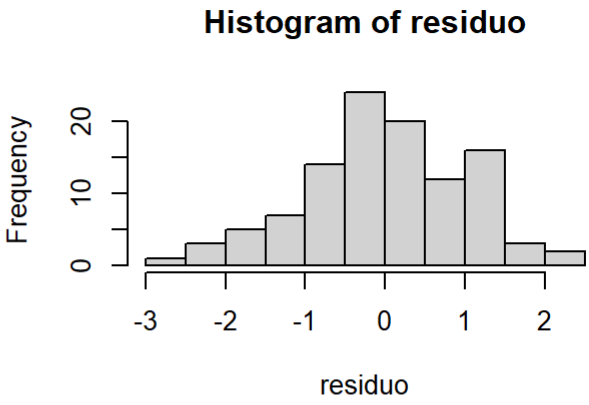


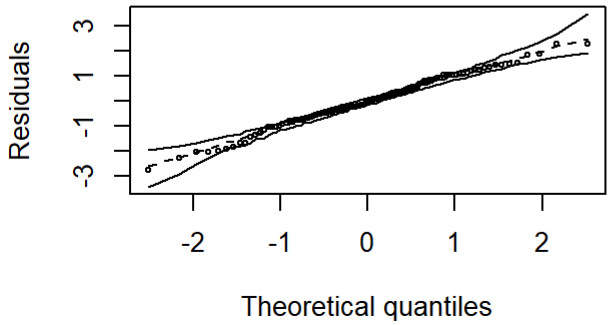
Mais uma vez, a observação 23 se destaca como um ponto de influência sobre o ajuste geral do modelo.

Resíduos

```
Gaussian model (lm object)
```







Nos resíduos, não notamos nenhuma medida muito discrepante, mas podemos notar que a observação 23 é a mais discrepante no primeiro gráfico, quase ultrapassando o intervalo de -3 a 3.

Novo Ajuste

Nesse caso, devido à alta influência da observação 23 em diversos aspectos do modelo, vamos optar por removê-la de nosso conjunto de dados e, então, reajustar os modelos do início utilizando a função `step`. O modelo recomendado é dado abaixo:

```
Start:  AIC=490.87
y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7

       Df Sum of Sq    RSS    AIC
- x4    1         6   9358 488.93
- x7    1        22   9374 489.12
- x5    1        23   9375 489.13
- x2    1        98   9450 489.98
- x3    1       126   9478 490.29
<none>              9352 490.87
- x1    1      9112  18464 560.97
- x6    1    105043 114395 754.30

Step:  AIC=488.93
y ~ x1 + x2 + x3 + x5 + x6 + x7

       Df Sum of Sq    RSS    AIC
```

```
- x5    1        23   9380 487.19
- x7    1        26   9384 487.23
- x2    1        97   9455 488.03
<none>                9358 488.93
- x3    1      2328  11685 510.48
- x1    1      9172  18530 559.35
- x6    1    106811 116168 753.93

Step:  AIC=487.19
y ~ x1 + x2 + x3 + x6 + x7

       Df Sum of Sq    RSS    AIC
- x7    1        19   9399 485.40
- x2    1        89   9469 486.19
<none>                9380 487.19
- x3    1      2305  11685 508.48
- x1    1      9240  18620 557.87
- x6    1    110278 119658 755.07

Step:  AIC=485.4
y ~ x1 + x2 + x3 + x6

       Df Sum of Sq    RSS    AIC
- x2    1        86   9485 484.36
<none>                9399 485.40
- x3    1      2287  11686 506.49
- x1    1      9229  18628 555.91
- x6    1    114445 123844 756.71

Step:  AIC=484.36
y ~ x1 + x3 + x6

       Df Sum of Sq    RSS    AIC
<none>                9485 484.36
- x3    1      2209  11694 504.56
- x1    1      9359  18844 555.14
- x6    1    114488 123973 754.82


Call:
lm(formula = y ~ x1 + x3 + x6, data = dados_new)

Coefficients:
(Intercept)           x1           x3           x6
      50.06        19.39        26.94        10.33
```

O modelo selecionado considerou apenas as covariáveis x_1, x_3 e x_6. O ajuste deste modelo é dado abaixo:

```
Call:
lm(formula = y ~ x1 + x3 + x6, data = dados_new)

Residuals:
     Min       1Q   Median       3Q      Max
-23.3746  -5.0895  -0.1982   7.0186  22.5797
```

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  50.0609     2.9751  16.827  < 2e-16 ***
x1           19.3892     1.9326  10.033  < 2e-16 ***
x3           26.9352     5.5266   4.874 4.03e-06 ***
x6           10.3288     0.2944  35.089  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.643 on 102 degrees of freedom
Multiple R-squared:  0.9294,    Adjusted R-squared:  0.9273
F-statistic: 447.7 on 3 and 102 DF,  p-value: < 2.2e-16
```

Partimos agora para a análise de diagnóstico desse novo ajuste.

Análise de diagnóstico

Suposições do modelo

Começaremos novamente testando as suposições do modelo.

Testanto [S0]:

```
    RESET test

data:  fit4
RESET = 0.0086259, df1 = 2, df2 = 100, p-value = 0.9914
```

Testanto [S1]:

```
    One Sample t-test

data:  resid(fit4)
t = 2.5885e-17, df = 105, p-value = 1
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 -1.83041  1.83041
sample estimates:
   mean of x
2.389541e-17
```

Testanto [S2]:

```
    studentized Breusch-Pagan test

data:  fit4
BP = 2.9632, df = 3, p-value = 0.3973
```

Testanto [S3]:

```
    Durbin-Watson test

data:  fit4
DW = 1.997, p-value = 0.4747
alternative hypothesis: true autocorrelation is greater than 0
```

Testanto [S4]:

```
      x1       x3       x6
1.009861 1.007971 1.004456
```

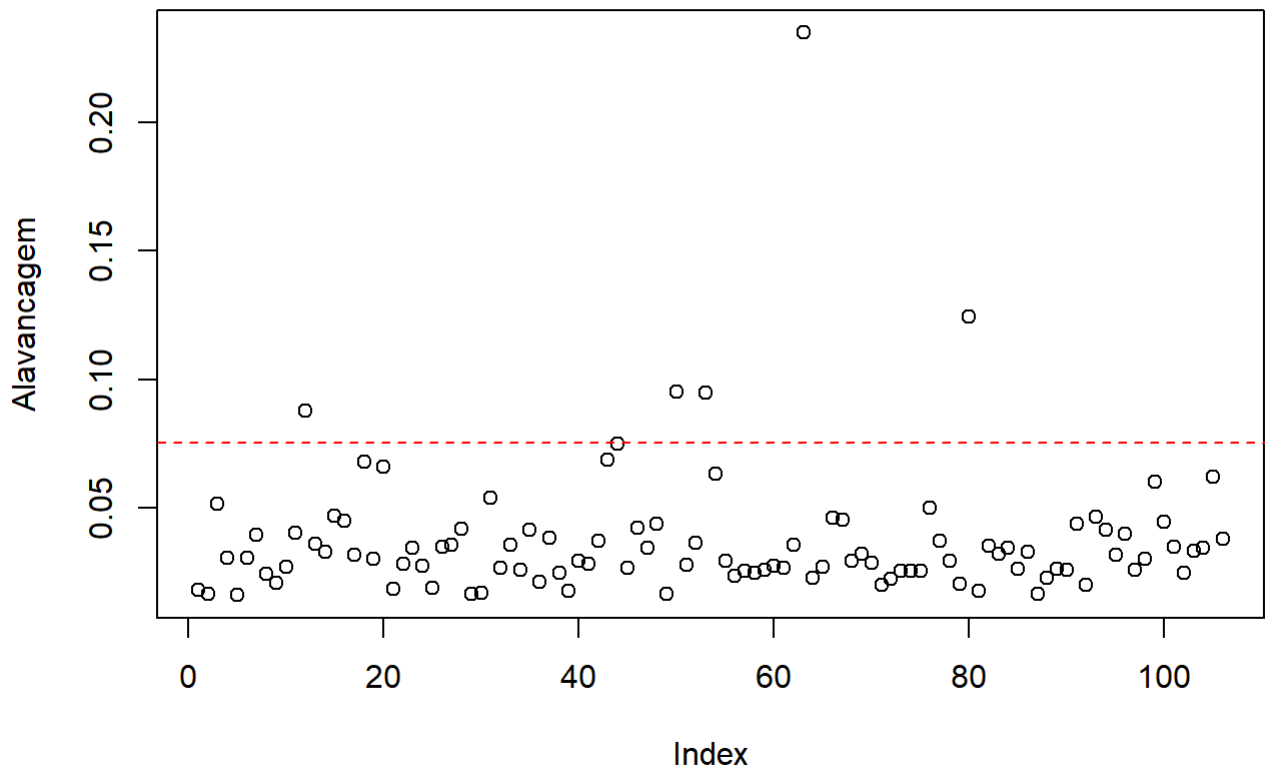
Testanto [S5]:

```
    Jarque Bera Test

data:  resid(fit4)
X-squared = 0.90357, df = 2, p-value = 0.6365
```

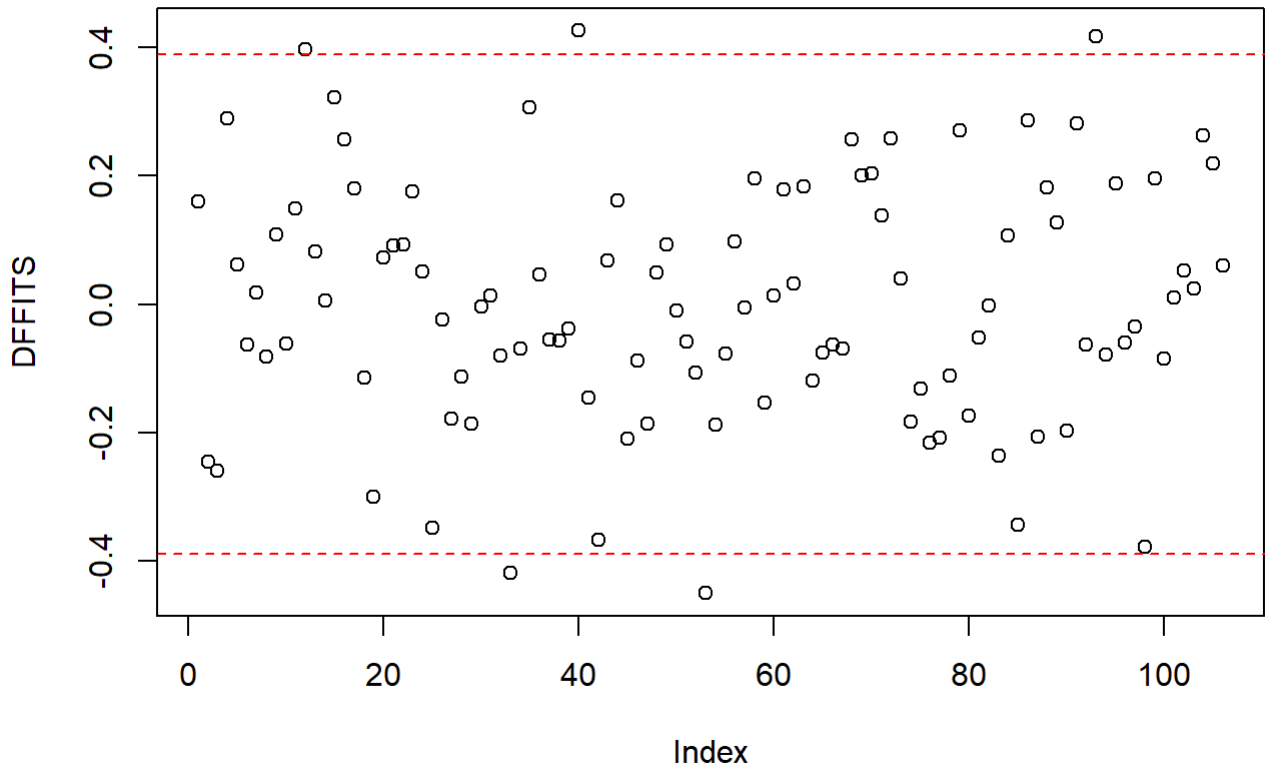
Mais uma vez, todos os testes de hipóstese obtiveram um $P - valor$ maior que $\alpha = 0,05$. Portanto, as suposições do modelo foram satisfeitas. Novamente os VIFs se encontram extremamente próximos de 1 logo, $[S_4]$ também é satisfeita.

Alavancagem



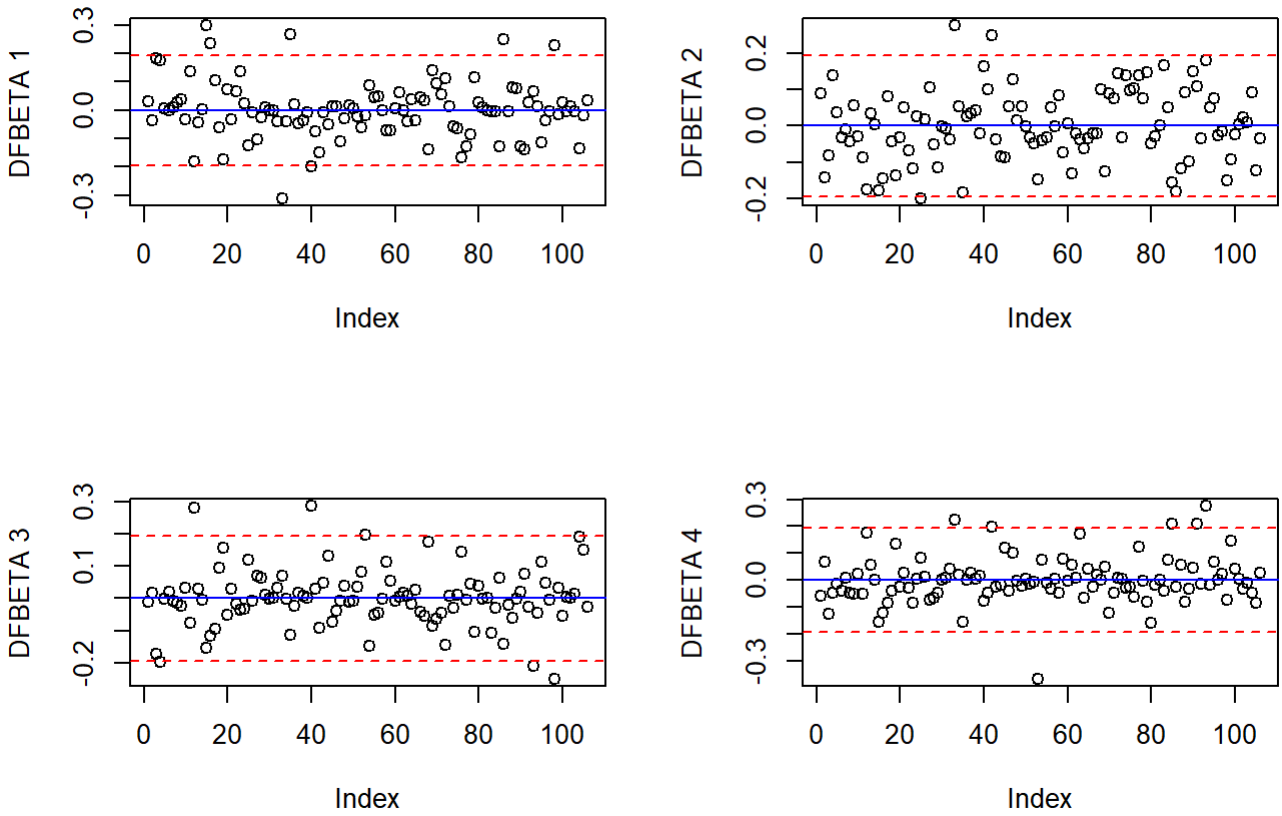
Novamente, temos a presença de alguns pontos de alavanca, sendo a observação 63 a mais discrepante e única preocupante até o momento.

DFFITS



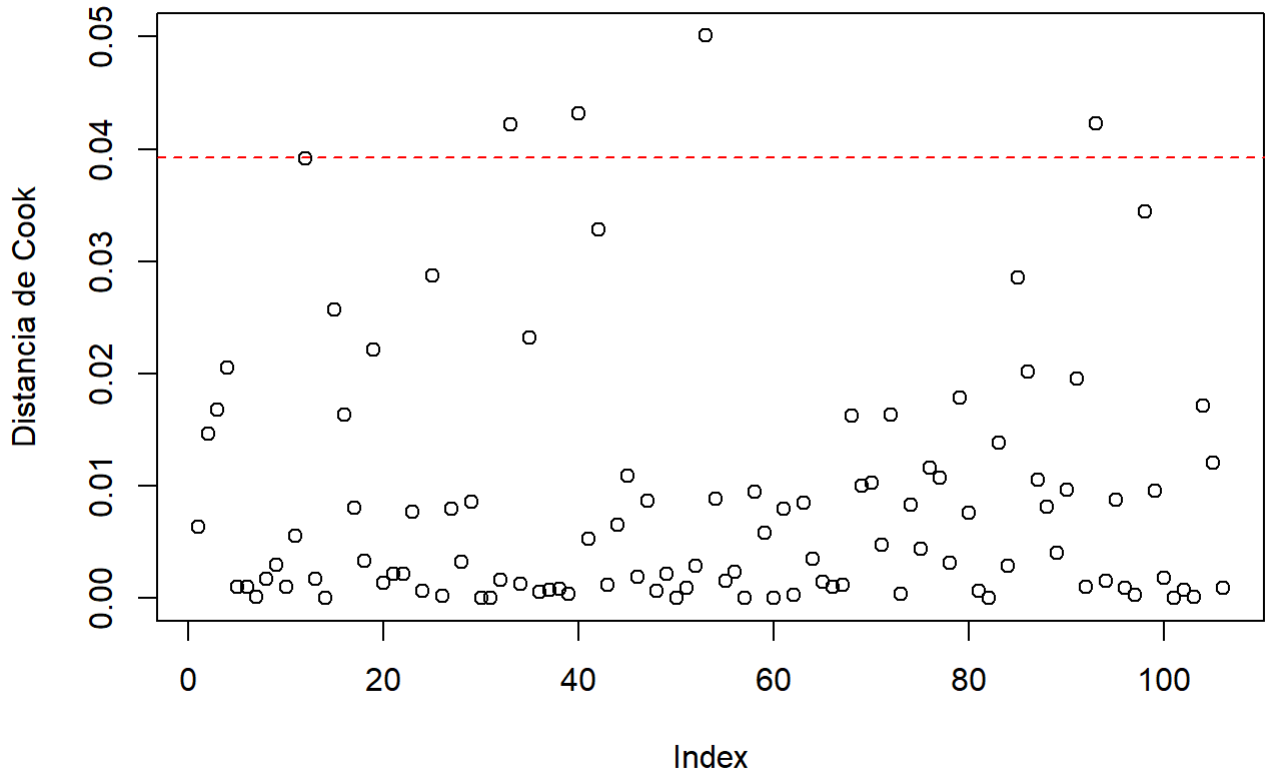
Aqui, não temos nenhum indício significativo de observações influentes sobre seu próprio valor ajustado.

DFBETAS



Também não temos nenhuma observação muito influente em relação aos parâmetros de regressão.

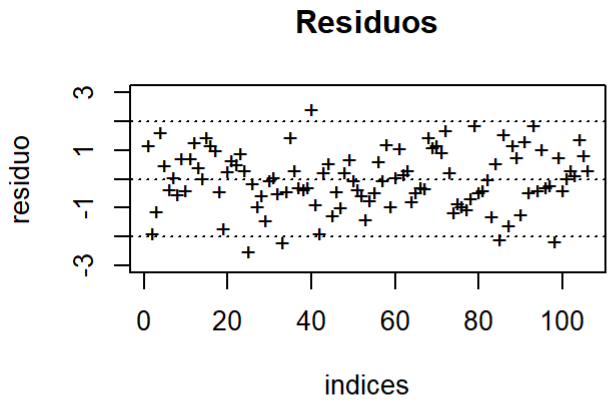
Distância de Cook

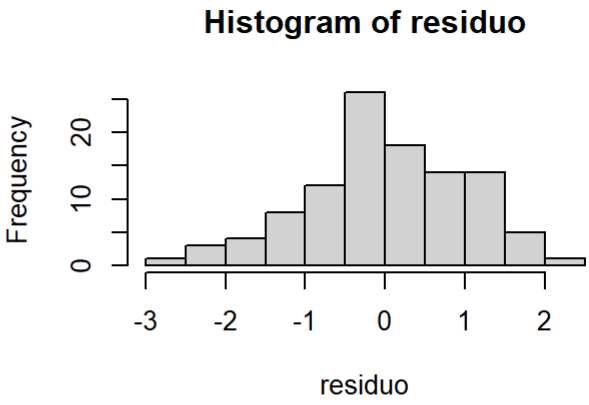


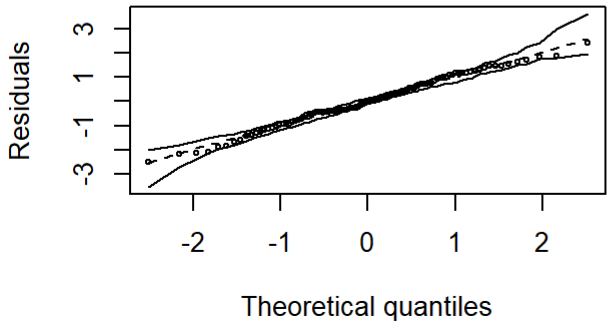
Não há pontos de muita influência sobre o ajuste em geral do modelo.

Resíduos

```
Gaussian model (lm object)
```







No primeiro gráfico não observamos nenhum valor fora do esperado, ou seja, nenhum valor discrepante. Podemos ver que os resíduos aparentam apresentar normalidade pelo histograma, e a maioria dos pontos se encontram dentro das bandas de confiança do invelope simulado.

Assim, podemos concluir que o modelo atual está bem ajustado e não necessita de mais reajustes. Logo, este será o nosso modelo final.

Predição de alguns valores

Por fim, faremos a predição para alguns valores hipotéticos gerados aleatoriamente ao fixarmos uma *seed*. Os valores gerados são mostrados abaixo:

```
# A tibble: 10 × 3
      x1    x3    x6
   <dbl> <dbl> <dbl>
 1     0 0.382  6.63
 2     0 0.734  2.29
 3     0 0.590  3.48
 4     1 0.523  2.42
 5     1 0.901  5.72
 6     0 0.767 10.4
 7     0 0.444  2.76
 8     0 0.548 11.2
 9     0 0.229  4.19
10     0 0.382  1.45
```

Os valores preditos pelo modelo ajustado são dados abaixo:

```
       1        2        3        4        5        6        7        8
128.84526 93.46131 101.86487 108.51321 152.84406 178.42344 90.51140 180.54182
       9       10
 99.53992 75.37362
```