

DESCRICAO DO PROJETO

Lauro C. M. de Paula

27/10/2017

Projeto 3 - Prevendo Despesas Hospitalares

Para esta análise, vamos usar um conjunto de dados simulando despesas médicas hipotéticas para um conjunto de pacientes espalhados por 4 regiões do Brasil. Esse dataset possui 1.338 observações e 7 variáveis. Todo o projeto será descrito de acordo com suas etapas.

Etapa 1 - Coletando os Dados

```
# Coletando dados
```

```
despesas <- read.csv("http://datascienceacademy.com.br/blog/aluno/RFundamentos/Datasets/ML/despesas.csv")
```

Etapa 2 - Explorando os Dados

```
# Visualizando as variaveis
```

```
str(despesas)
```

```
## 'data.frame': 1338 obs. of 7 variables:
## $ idade : int 19 18 28 33 32 31 46 37 37 60 ...
## $ sexo : Factor w/ 2 levels "homem","mulher": 2 1 1 1 1 2 2 2 1 2 ...
## $ bmi : num 27.9 33.8 33 22.7 28.9 25.7 33.4 27.7 29.8 25.8 ...
## $ filhos : int 0 1 3 0 0 0 1 3 2 0 ...
## $ fumantes: Factor w/ 2 levels "nao","sim": 2 1 1 1 1 1 1 1 1 1 ...
## $ regioao : Factor w/ 4 levels "nordeste","norte",...: 3 4 4 1 1 4 4 1 2 1 ...
## $ gastos : num 16885 1726 4449 21984 3867 ...
```

```
# Medias de Tendencia Central da variavel gastos
```

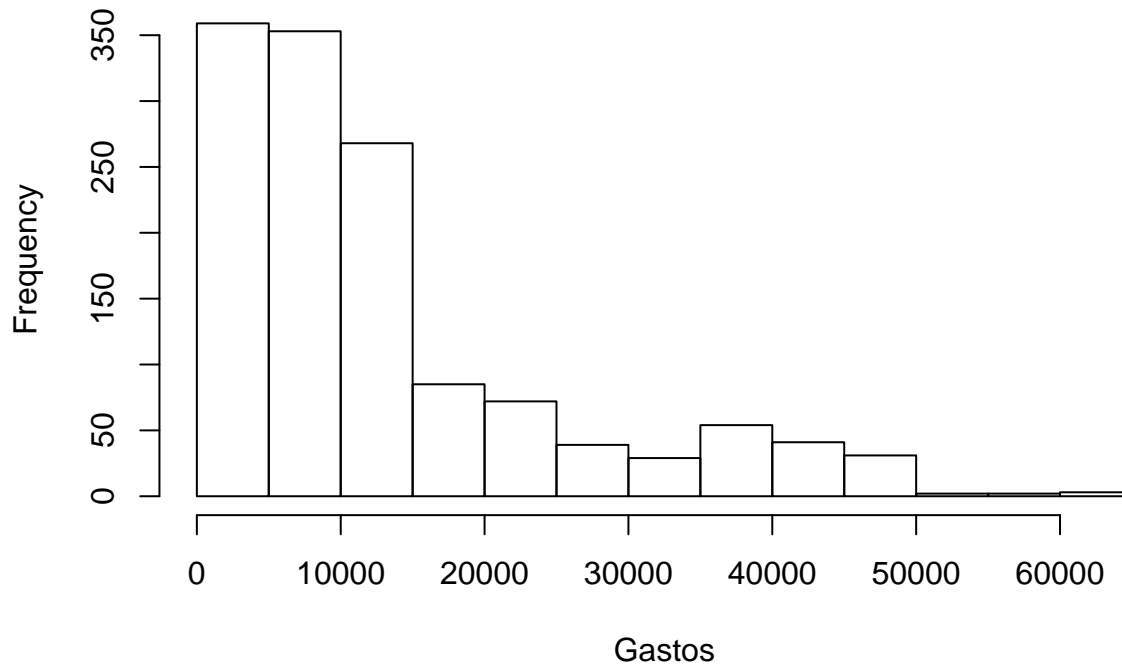
```
summary(despesas$gastos)
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.     Max.
##      1122   4740   9382   13270   16640   63770
```

```
# Construindo um histograma
```

```
hist(despesas$gastos, main = 'Histograma', xlab = 'Gastos')
```

Histograma



```
# Tabela de contingencia das regioes
table(despesas$regiao)
```

```
##
## nordeste    norte  sudeste    sul
##      325      324      325      364
```

```
# Explorando relacionamento entre as variaveis: Matriz de Correlacao
cor(despesas[c("idade", "bmi", "filhos", "gastos")])
```

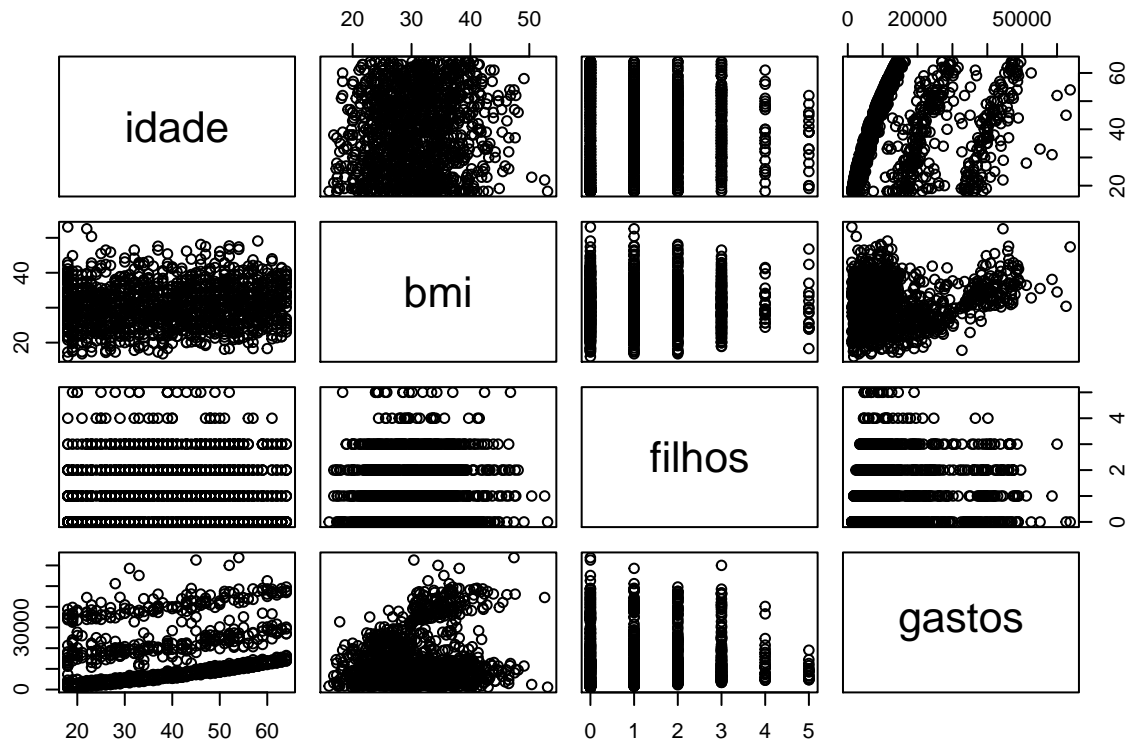
```
##          idade          bmi          filhos          gastos
## idade  1.0000000  0.10934101  0.04246900  0.29900819
## bmi    0.1093410  1.00000000  0.01264471  0.19857626
## filhos 0.0424690  0.01264471  1.00000000  0.06799823
## gastos 0.2990082  0.19857626  0.06799823  1.00000000
```

```
# Nenhuma das correlacoes na matriz é considerada forte, mas existem algumas associacoes interessantes.
# Por exemplo, a idade e o bmi (IMC) parecem ter uma correlacao positiva fraca, o que significa que,
# com o aumento da idade, a massa corporal tende a aumentar. Ha tambem uma correlacao positiva
# moderada entre a idade e os gastos, alem do numero de filhos e os gastos. Essas associacoes implicam
# que, a media que idade, massa corporal e numero de filhos aumentam, o custo esperado do seguro saude
```

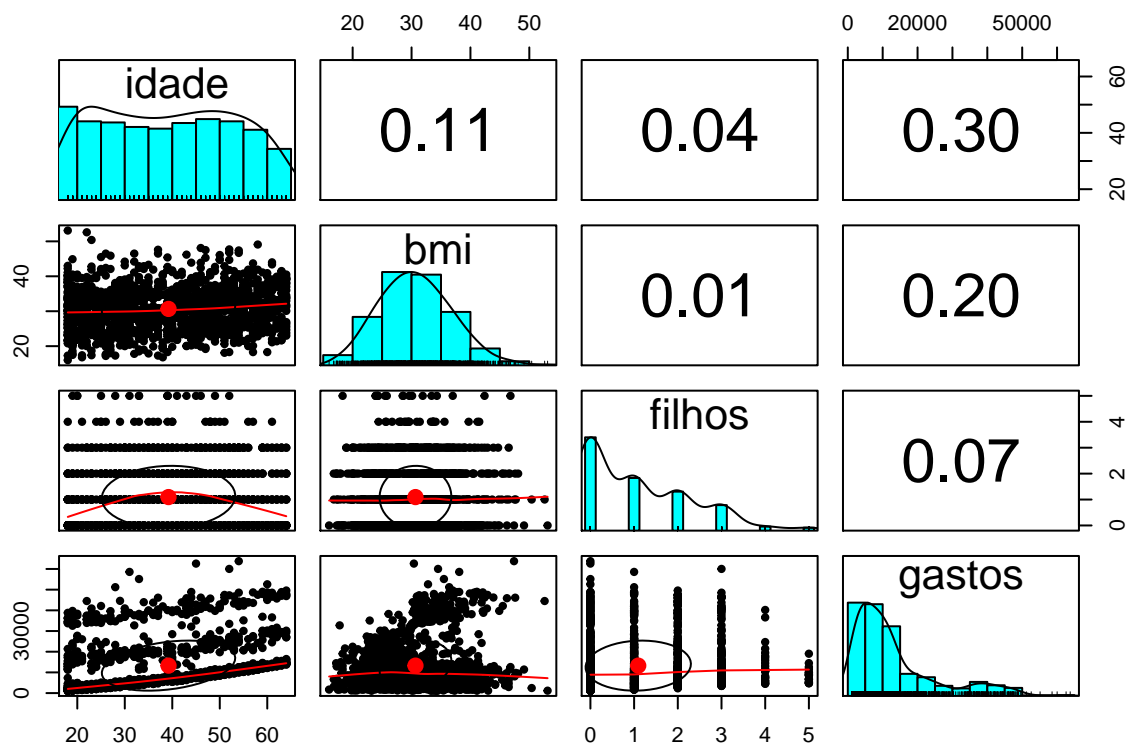
```
# Visualizando relacionamento entre as variaveis: Scatterplot
# Perceba que nao existe um claro relacionamento entre as variaveis
pairs(despesas[c("idade", "bmi", "filhos", "gastos")])
```

```
# Scatterplot Matrix
#install.packages("psych")
library(psych)
```

```
## Warning: package 'psych' was built under R version 3.4.2
```



```
# Este grafico fornece mais informacoes sobre o relacionamento entre as variaveis
pairs.panels(despesas[c("idade", "bmi", "filhos", "gastos")])
```



Etapa 3 - Treinando o modelo

```
modelo <- lm(gastos ~ idade + filhos + bmi + sexo + fumantes + regiao,
             data = despesas)
```

Similar ao item anterior

```
modelo <- lm(gastos ~ ., data = despesas)
```

Visualizando os coeficientes

```
modelo
```

```
##
```

```
## Call:
```

```
## lm(formula = gastos ~ ., data = despesas)
```

```
##
```

```
## Coefficients:
```

```
## (Intercept)      idade  sexomulher      bmi      filhos
## -12425.7      256.8      131.4      339.3      475.7
## fumantessim  regiaonorte regiaosudeste  regiaosul
## 23847.5      352.8      -606.5      -682.8
```

Prevendo despesas medicas

```
previsao <- predict(modelo)
```

```
class(previsao)
```

```
## [1] "numeric"
```

```
head(previsao)
```

```
##           1           2           3           4           5           6
## 25292.740  3458.281  6706.619  3751.868  5598.626  3704.606
```

Etapa 4 - Avaliando a Performance do Modelo

```
# Mais detalhes sobre o modelo
```

```
summary(modelo)
```

```
##
## Call:
## lm(formula = gastos ~ ., data = despesas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11302.7  -2850.9   -979.6   1383.9  29981.7
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -12425.7    1000.7  -12.418  < 2e-16 ***
## idade         256.8       11.9   21.586  < 2e-16 ***
## sexomulher    131.3       332.9    0.395  0.693255
## bmi           339.3       28.6   11.864  < 2e-16 ***
## filhos        475.7       137.8    3.452  0.000574 ***
## fumantessim  23847.5       413.1   57.723  < 2e-16 ***
## regiaonorte   352.8       476.3    0.741  0.458976
## regioasudeste -606.5       477.2   -1.271  0.203940
## regioasul     -682.8       478.9   -1.426  0.154211
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6062 on 1329 degrees of freedom
## Multiple R-squared:  0.7509, Adjusted R-squared:  0.7494
## F-statistic: 500.9 on 8 and 1329 DF,  p-value: < 2.2e-16
```

Etapa 5 - Otimizacao do Modelo

```
# Adicionando uma variavel com o dobro do valor das idades
```

```
despesas$idade2 <- despesas$idade ^ 2
```

```
# Adicionando um indicador para BMI >= 30
```

```
despesas$bmi30 <- ifelse(despesas$bmi >= 30, 1, 0)
```

```
# Criando o modelo final
```

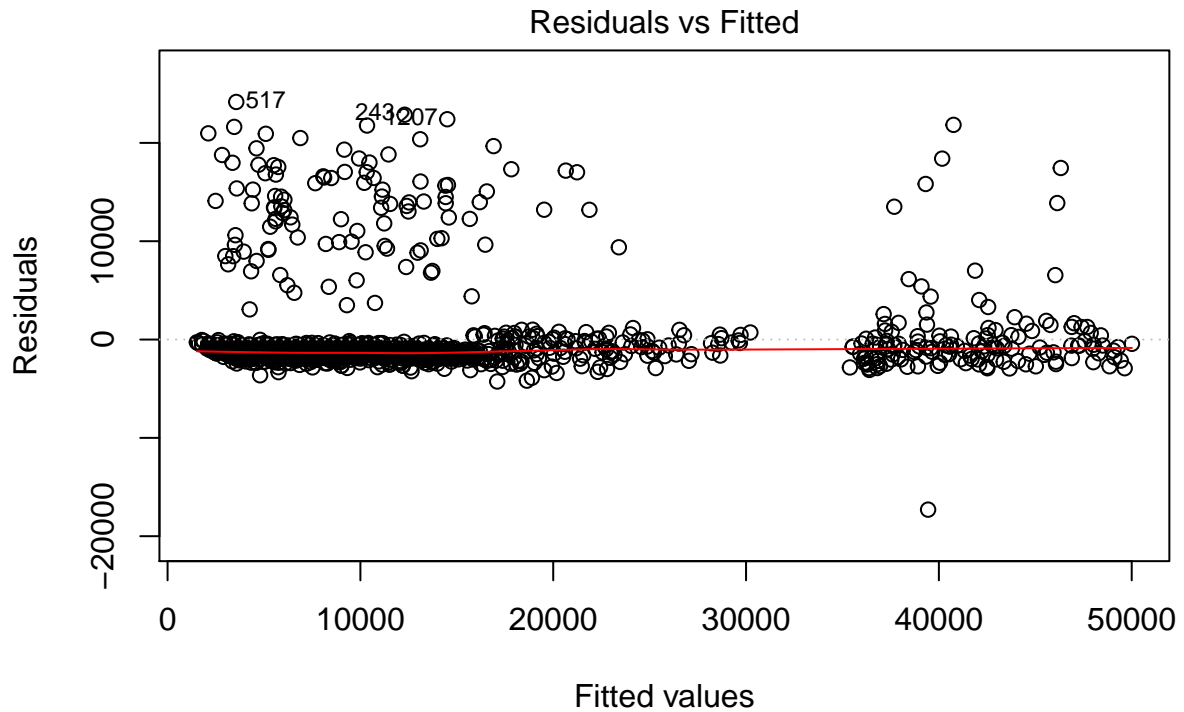
```
modelo_v2 <- lm(gastos ~ idade + idade2 + filhos + bmi + sexo +
               bmi30 * fumantes + regioao, data = despesas)
```

```
summary(modelo_v2)
```

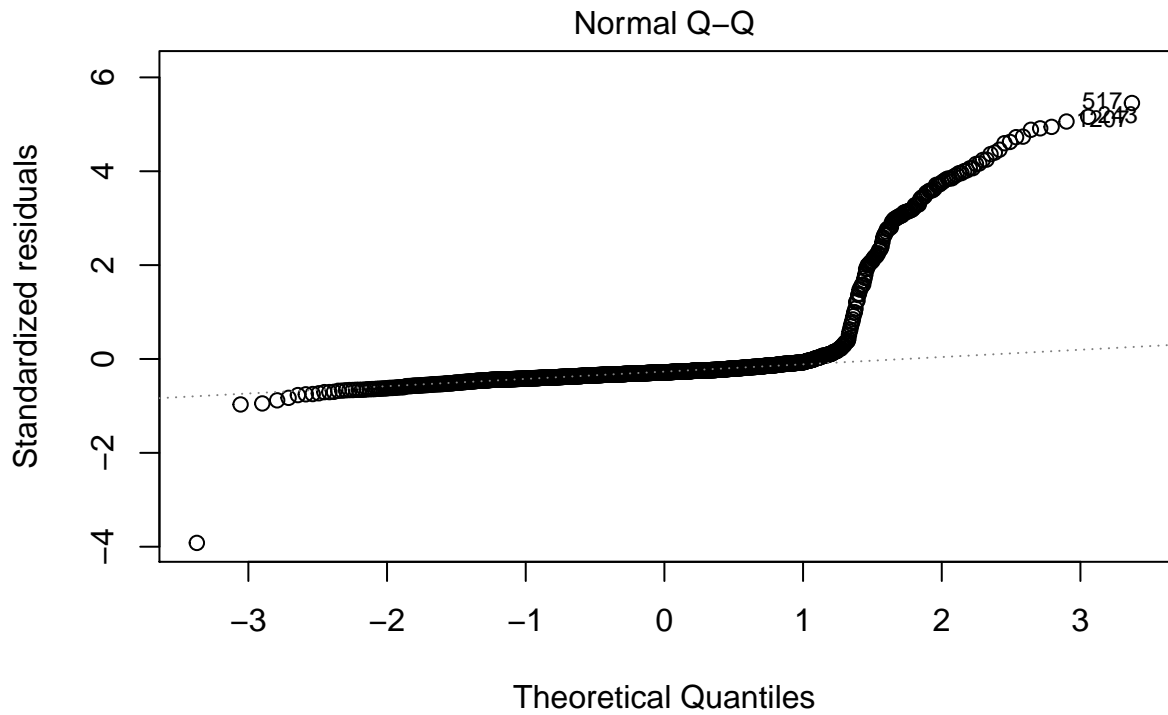
```
##
## Call:
```

```
## lm(formula = gastos ~ idade + idade2 + filhos + bmi + sexo +
##     bmi30 * fumantes + regiao, data = despesas)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17297.1  -1656.0  -1262.7   -727.8  24161.6
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -636.9298   1361.0589   -0.468  0.639886
## idade          -32.6181     59.8250   -0.545  0.585690
## idade2           3.7307      0.7463    4.999 6.54e-07 ***
## filhos          678.6017   105.8855    6.409 2.03e-10 ***
## bmi            119.7715     34.2796    3.494 0.000492 ***
## sexomulher      496.7690    244.3713    2.033 0.042267 *
## bmi30          -997.9355    422.9607   -2.359 0.018449 *
## fumantessim    13404.5952   439.9591   30.468 < 2e-16 ***
## regiaonorte     279.1661    349.2826    0.799 0.424285
## regiaosudeste  -942.9958    350.1754   -2.693 0.007172 **
## regiaosul      -548.8684    352.1950   -1.558 0.119372
## bmi30:fumantessim 19810.1534   604.6769   32.762 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4445 on 1326 degrees of freedom
## Multiple R-squared:  0.8664, Adjusted R-squared:  0.8653
## F-statistic: 781.7 on 11 and 1326 DF, p-value: < 2.2e-16

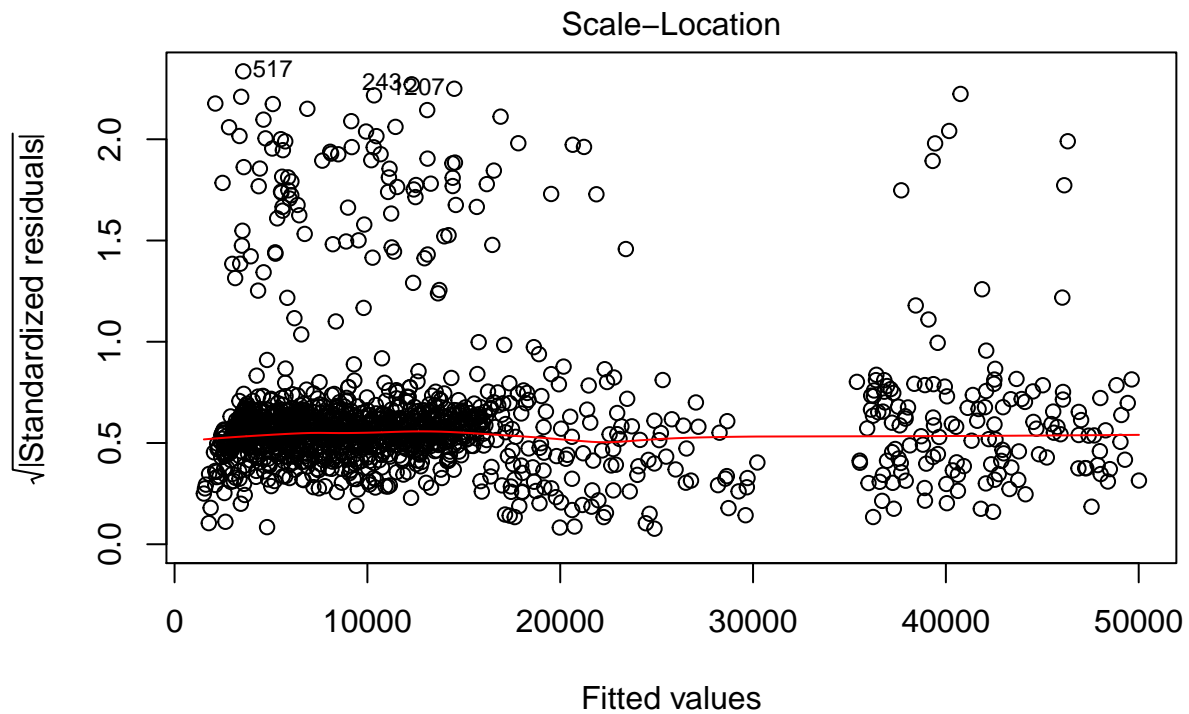
plot(modelo_v2)
```



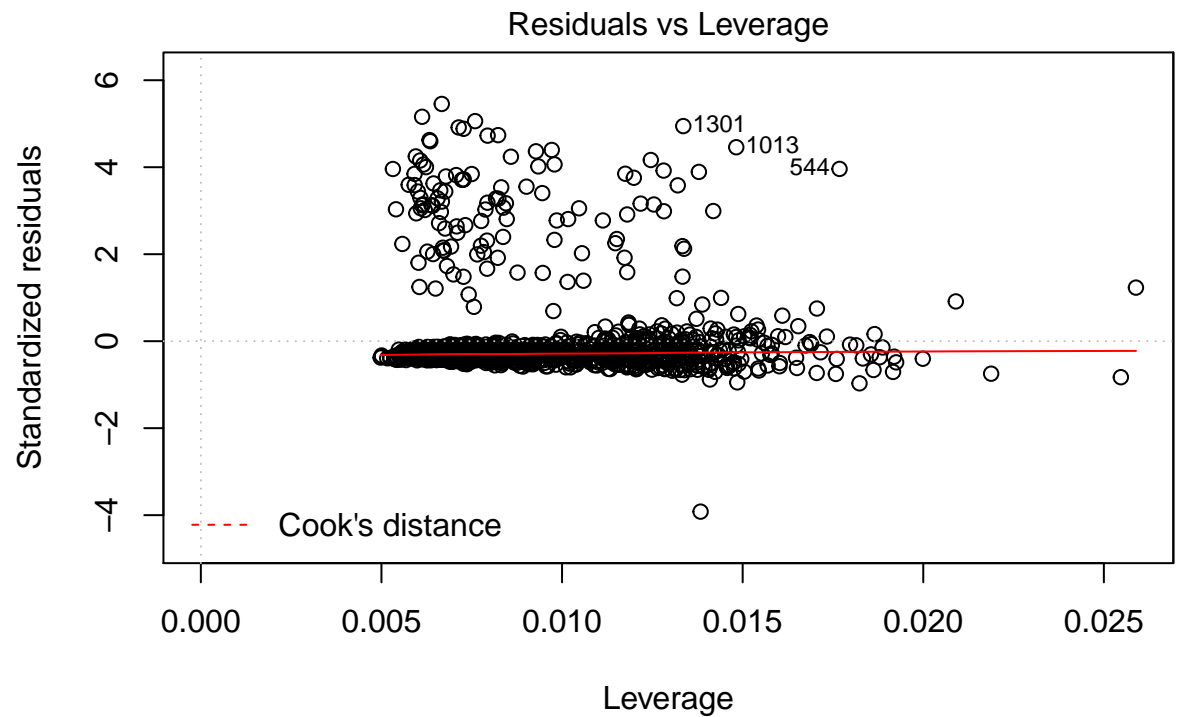
lm(gastos ~ idade + idade2 + filhos + bmi + sexo + bmi30 * fumantes + regia ...



lm(gastos ~ idade + idade2 + filhos + bmi + sexo + bmi30 * fumantes + regia ...



lm(gastos ~ idade + idade2 + filhos + bmi + sexo + bmi30 * fumantes + regia ...



$\text{lm}(\text{gastos} \sim \text{idade} + \text{idade2} + \text{filhos} + \text{bmi} + \text{sexo} + \text{bmi30} * \text{fumantes} + \text{regia} \dots)$

Fim

www.datascienceacademy.com.br