Quakes

Grupo: Americo Freitas, Arleks dos Santos e Luciano Ozorio

2022-05-08

Depois de chamar nosso conjunto de dados, é sempre uma boa prática realizar EDA, Análise Exploratória de Dados, antes de realizar análises rigorosas. O EDA nos permite confirmar que os dados corretos foram carregados no R, ao mesmo tempo em que nos dá uma prévia do conjunto de dados. Abaixo estão duas funções EDA comuns para ajudá-lo a ter uma ideia dos dados, sendo que ambas têm o conjunto de dados como o único argumento. A função head mostra as primeiras seis linhas e todas as colunas dos dados, enquanto a função de resumo fornece estatísticas de resumo para cada variável.

```
head(quakes)
```

```
##      lat   long depth mag stations
## 1 -20.42 181.62   562 4.8       41
## 2 -20.62 181.03   650 4.2       15
## 3 -26.00 184.10    42 5.4       43
## 4 -17.97 181.66   626 4.1       19
## 5 -20.42 181.96   649 4.0       11
## 6 -19.68 184.31   195 4.0       12
```

Summary

Além disso, o conjunto de dados de terremotos contém 1.000 observações de dados de terremotos que ocorreram perto da ilha tropical de Fiji. Cada observação tem a latitude, longitude, profundidade, magnitude e número de estações que relataram a atividade sísmica. Para este R-Guide, focaremos na magnitude, medida pela escala logarítmica de Richter, do terremoto e no número de estações que relataram cada terremoto.

```
summary(quakes)
```

```
##       lat              long           depth            mag
##  Min.   :-38.59   Min.   :165.7   Min.   : 40.0   Min.   :4.00
##  1st Qu.:-23.47   1st Qu.:179.6   1st Qu.: 99.0   1st Qu.:4.30
##  Median :-20.30   Median :181.4   Median :247.0   Median :4.60
##  Mean   :-20.64   Mean   :179.5   Mean   :311.4   Mean   :4.62
##  3rd Qu.:-17.64   3rd Qu.:183.2   3rd Qu.:543.0   3rd Qu.:4.90
##  Max.   :-10.72   Max.   :188.1   Max.   :680.0   Max.   :6.40
##     stations
##  Min.   : 10.00
##  1st Qu.: 18.00
##  Median : 27.00
##  Mean   : 33.42
##  3rd Qu.: 42.00
##  Max.   :132.00
```

```
cor(quakes)
```

```
##                   lat        long       depth         mag     stations
## lat       1.000000000 -0.36454404  0.03102583 -0.05046165 -0.002220645
```

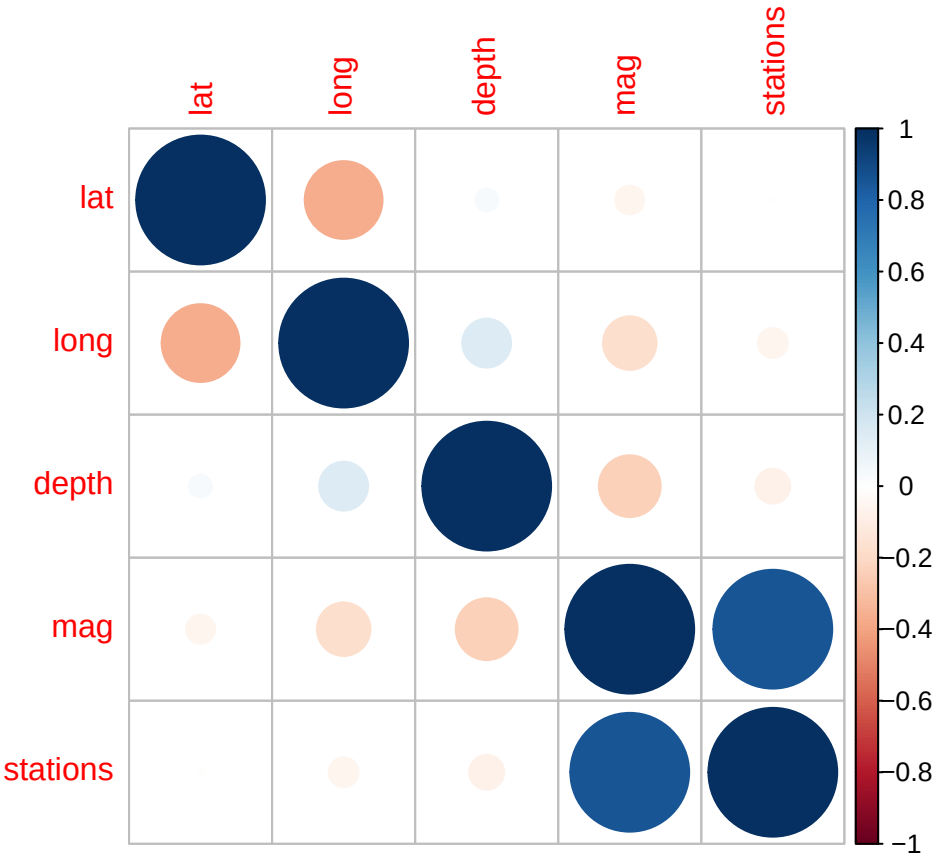
```
## long      -0.364544037  1.00000000  0.14444341 -0.17306726 -0.053512460
## depth      0.031025831  0.14444341  1.00000000 -0.23063770 -0.073515097
## mag       -0.050461651 -0.17306726 -0.23063770  1.00000000  0.851182422
## stations -0.002220645 -0.05351246 -0.07351510  0.85118242  1.000000000
```

Para melhorar a visualização dessa matriz de correlação com a função corrplot. Quanto maior o círculo maior a correlação entre as variaveis. *Além disso, quanto mais azul escuro, mais próxima a correlação fica de 1,* que significa que além de forte a correlação é positiva. *Equivalentemente quanto mais próximo de vermelho escuro, mais próxima a* correlação fica de -1, que significa que além de forte a correlação é negativa. *As variáveis explicativas mag e stations possuem alta correlação 0.85

```
library(corrplot)
```

```
## corrplot 0.92 loaded
```

```
corrplot(cor(quakes), method = "circle")
```



Nesta primeira aba

```
mod = lm(mag ~ ., data = quakes)
summary(mod)
```

```
##
## Call:
## lm(formula = mag ~ ., data = quakes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62156 -0.13401 -0.00419  0.12857  0.79298
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.731e+00  1.878e-01  30.514  < 2e-16 ***
## lat         -7.690e-03  1.308e-03  -5.879 5.63e-09 ***
## long        -9.452e-03  1.096e-03  -8.627  < 2e-16 ***
## depth       -2.726e-04  2.878e-05  -9.473  < 2e-16 ***
## stations     1.531e-02  2.795e-04  54.777  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1928 on 995 degrees of freedom
## Multiple R-squared:  0.7719, Adjusted R-squared:  0.7709
## F-statistic: 841.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

O método step indica que todas as variáveis são significativas.

```
mod1=step(mod, direction = "backward")
```

```
## Start:  AIC=-3287.54
## mag ~ lat + long + depth + stations
##
##            Df Sum of Sq     RSS     AIC
## <none>                  36.974 -3287.5
## - lat       1     1.284  38.258 -3255.4
## - long      1     2.765  39.739 -3217.4
## - depth     1     3.335  40.309 -3203.2
## - stations  1   111.500 148.474 -1899.3
```

```
summary(mod1)
```

```
##
## Call:
## lm(formula = mag ~ lat + long + depth + stations, data = quakes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.62156 -0.13401 -0.00419  0.12857  0.79298
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5.731e+00  1.878e-01  30.514  < 2e-16 ***
## lat         -7.690e-03  1.308e-03  -5.879 5.63e-09 ***
## long        -9.452e-03  1.096e-03  -8.627  < 2e-16 ***
## depth       -2.726e-04  2.878e-05  -9.473  < 2e-16 ***
## stations     1.531e-02  2.795e-04  54.777  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1928 on 995 degrees of freedom
## Multiple R-squared:  0.7719, Adjusted R-squared:  0.7709
## F-statistic: 841.6 on 4 and 995 DF,  p-value: < 2.2e-16
```

Todas as variáveis explicativas são significativas pois possuem p-value abaixo de 5% (teste t). O Valor de R quadrado ajustado é alto 77% (maior que 70%) o que significa que as variáveis explicativas explicam 77% da variável mag.

Análise de Resíduos

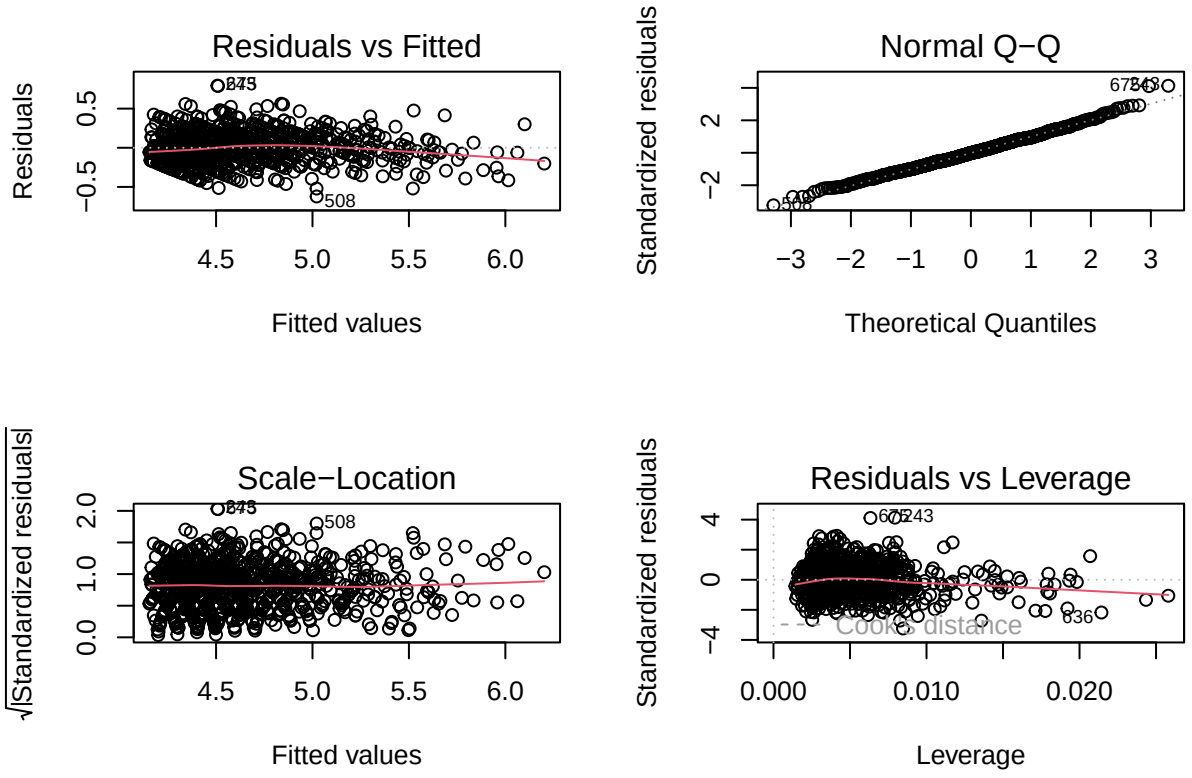
```
anares <- rstandard(mod)
par(mfrow=c(2,2))
aov(mod)
```

```
## Call:
##    aov(formula = mod)
##
## Terms:
##                       lat      long     depth  stations Residuals
## Sum of Squares    0.41268   6.85144   6.32537 111.50030  36.97406
## Deg. of Freedom         1         1         1         1       995
##
## Residual standard error: 0.1927689
## Estimated effects may be unbalanced
```

```
av=aov(mod)
av
```

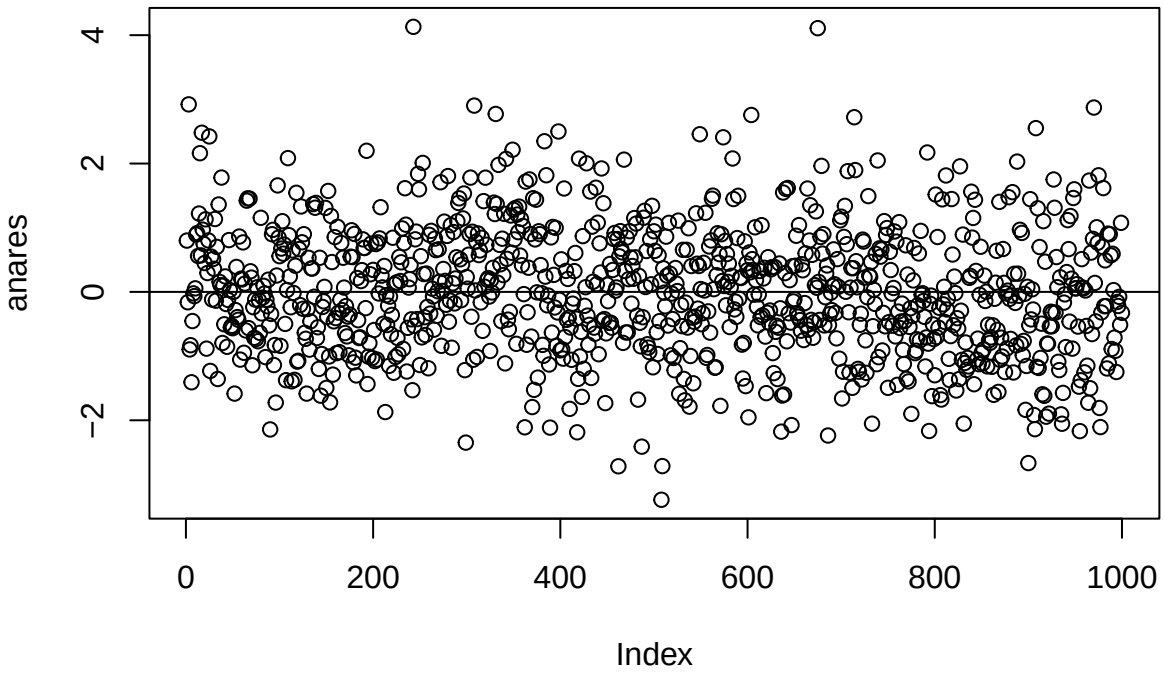
```
## Call:
##    aov(formula = mod)
##
## Terms:
##                       lat      long     depth  stations Residuals
## Sum of Squares    0.41268   6.85144   6.32537 111.50030  36.97406
## Deg. of Freedom         1         1         1         1       995
##
## Residual standard error: 0.1927689
## Estimated effects may be unbalanced
```

```
plot(av)
```



* O gráfico (Residual vs. Fitted) mostra indícios da presença de não-linearidades no modelo * O gráfico Q-Q dos resíduos padronizados, é usado para verificação da normalidade dos resíduos. No nosso caso, tomamos como hipótese nula a normalidade dos resíduos.

```
plot(anares)
abline(0,0)
```



###

Teste formais de Normalidade

```
library(nortest)
ad.test(anares)
```

```
##
##  Anderson-Darling normality test
##
## data:  anares
## A = 0.50785, p-value = 0.1992
```

A hipótese nula para o teste AD é que os dados seguem uma distribuição normal.Nesse caso, nosso valor p é 0.1992. Como isso não está abaixo do nosso nível de significância (digamos 0,05),não temos evidências suficientes para rejeitar a hipótese nula. É seguro dizer que nossos dados seguem uma distribuição normal.

```
shapiro.test(anares)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  anares
## W = 0.99617, p-value = 0.01446
```

O valor p do teste acaba sendo 0.01446 . Como esse valor é menor que 0,05, temos evidências suficientes para dizer que os dados da amostra não vêm de uma população com distribuição normal.

Teste formal de Homocedasticidade

```
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

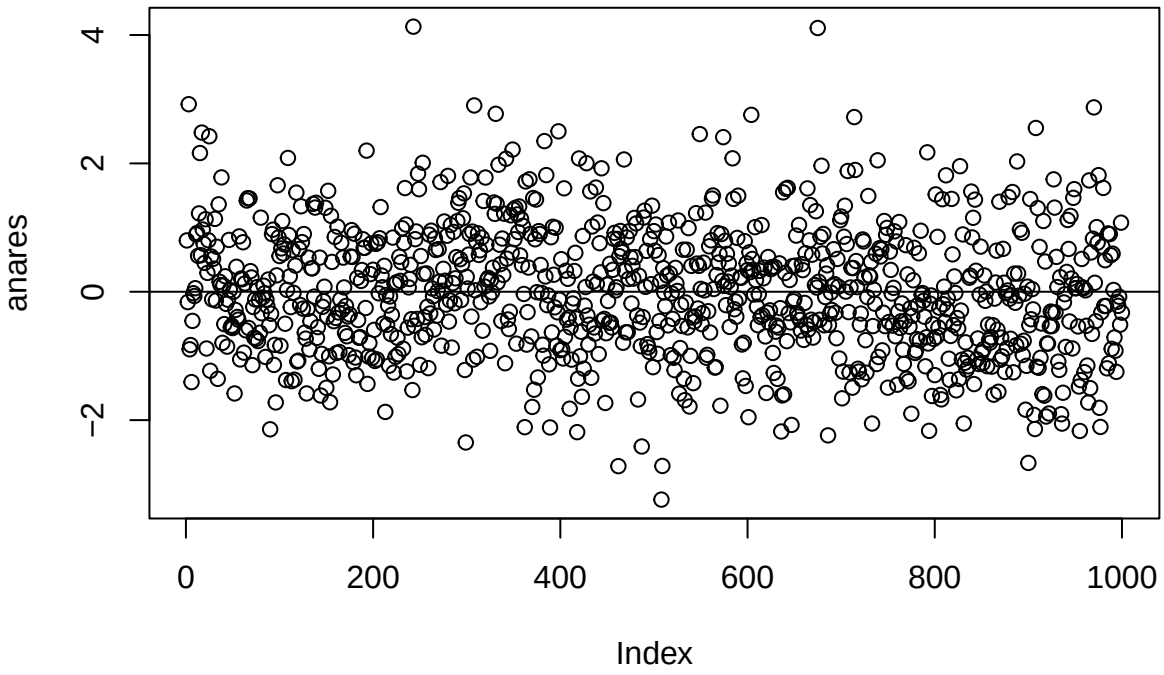
```
library(lmtest)
bptest(mod)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  mod
## BP = 6.7019, df = 4, p-value = 0.1525
```

#A estatística de teste é 6.7019 e o valor p correspondente é 0.1525 . Como o valor de p não é menor que 0,05, não rejeitamos a hipótese nula. Não temos evidências suficientes para dizer que a heterocedasticidade está presente no modelo de regressão.

Teste de Autocorrelação - Gráfico

```
plot(anares)
abline(0,0)
```



Teste formal de Autocorrelação

```
dwtest(mod)
```

```
##
##  Durbin-Watson test
##
## data:  mod
## DW = 1.9414, p-value = 0.1751
## alternative hypothesis: true autocorrelation is greater than 0
```

A partir da saída, podemos ver que a estatística de teste é 1.9414 e o valor p correspondente é 0.1751 . Como esse valor de p é maio que 0,05, não podemos rejeitar a hipótese nula e concluir que os resíduos nesse modelo de regressão não são autocorrelacionados.

```
pred_in = data.frame(lat=-20.62 , long = 181.03 , depth = 650 , stations = 15 )
predict(mod, pred_in, interval="confidence")
```

```
##        fit      lwr      upr
## 1 4.231062 4.207129 4.254995
```

A previsão de **mag** para os parametros passados com base no modelo, foi **4.231**. E foi estabelecido um intervalo de confiançã entre **4.207** e **4.254**.