TRABALHO FINAL DE MODELAGEM ESTATÍSTICA AVANÇADA

Professor: Paulo Maranhão
Turma 8

Rafael Rocha – A56660250

QUAKES

A descrição do dataset:

O conjunto de dados fornece as localizações de 1000 eventos sísmicos de MB> 4.0. Os eventos ocorreram em um cubo perto de Fiji desde 1964.

Atributos:

- **mag**: Magnitude Richter numérica (nossa variável resposta)
- **lat**: Latitude numérica do evento
- **long**: Longitude numérica longa
- **depth**: Profundidade numérica em KM.
- **stations**: Número de estações relatando

Visualização das 6 primeiras linhas dos dados desse dataset:

```
dados_quakes <- (quakes)
df <- dados_quakes
head(df)
```

```
     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
```

Checando as correlações:

```
cor(df)
```

```
                  lat        long       depth
lat       1.000000000 -0.36454404  0.03102583
long     -0.364544037  1.00000000  0.14444341
depth     0.031025831  0.14444341  1.00000000
mag      -0.050461651 -0.17306726 -0.23063770
stations -0.002220645 -0.05351246 -0.07351510
                 mag     stations
lat      -0.05046165 -0.002220645
long     -0.17306726 -0.053512460
depth    -0.23063770 -0.073515097
mag       1.00000000  0.851182422
stations  0.85118242  1.000000000
```

ANÁLISE: É possível notar que `stations` possui alta correlação com a variável `mag`. Nenhuma das variáveis explicativas apresentam correlação entre si, portando não apresentam problemas de multicolinearidade.

Testando possíveis modelos:

modelo 2
mod2 <- lm(mag ~., data=df)
summary(mod2)

```
Call:
lm(formula = mag ~ ., data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-0.62156 -0.13401 -0.00419  0.12857  0.79298

Coefficients:
             Estimate Std. Error t value
(Intercept)  5.731e+00  1.878e-01  30.514
lat         -7.690e-03  1.308e-03  -5.879
long        -9.452e-03  1.096e-03  -8.627
depth       -2.726e-04  2.878e-05  -9.473
stations     1.531e-02  2.795e-04  54.777
            Pr(>|t|)
(Intercept)  < 2e-16 ***
lat         5.63e-09 ***
long         < 2e-16 ***
depth        < 2e-16 ***
stations     < 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
```

modelo 3
mod3 <- step(mod2, direction = "backward")
summary(mod3)

modelo 4
mod4 <- lm(mag ~ lat + long, data = df)
summary(mod4)

modelo 5
mod5 <- lm(mag ~ depth + stations, data = df)
summary(mod5)

modelo 6
mod6 <- step(mod2, direction = "both")
summary(mod6)

modelo 7 (Propondo transformações nas variáveis)
```
mod7_log <- lm(log(mag) ~ log(.),data = df)
mod_mag_step <- step(mod7_log, direction = "both")
summary(mod_mag_step)
```

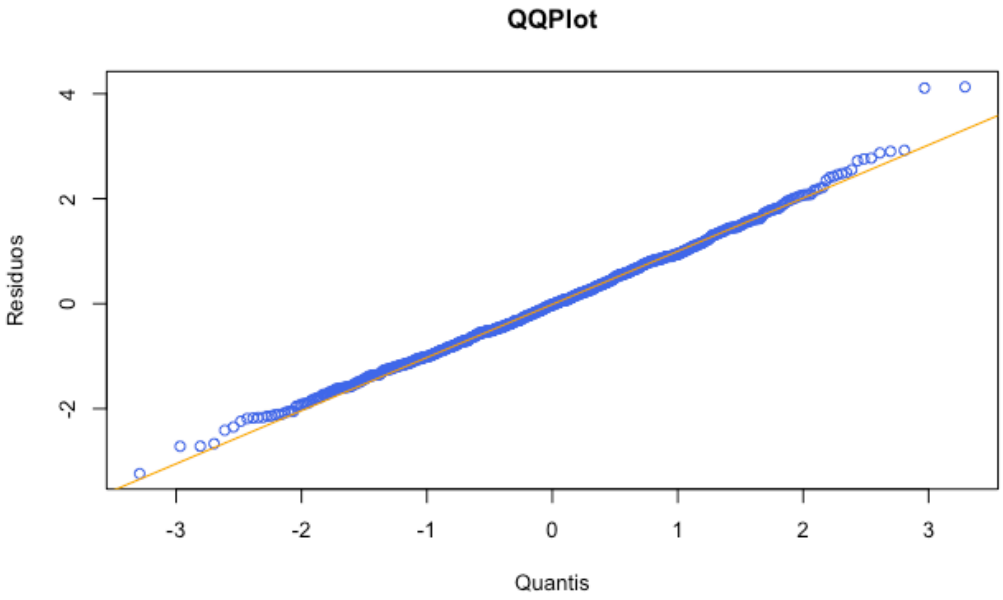
Otimização do modelo 2 – Que apresentou os melhores resultados.
```
mod2 <- lm(mag ~., data=df)
mod_mag_step <- step(mod2, direction = "both")
summary(mod_mag_step)
```

```
Call:
lm(formula = mag ~ lat + long + depth + stations, data = df)

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
```

Análise de Resíduos com modelo 2 que teve melhor R2.
```
anares <- rstandard(mod_mag_step)
par(mfrow=c(2,2))
```

ANÁLISE: O R2 do mod2 se apresentou como o melhor modelo de regressão linear. Seu p-value confirma que o modelo é melhor do que o modelo nulo.

Teste de Normalidade - Gráfico
```
qqnorm(anares, ylab="Residuos", xlab="Quantis", main="QQPlot", col="royalblue")
qqline(anares, col="orange")
```



Teste formais de Normalidade

Hipóteses:
- H_0: Normalidade
- H_1: Não Normalidade

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

```
        Anderson-Darling normality test

data:  anares
A = 0.50785, p-value = 0.1992
```

ANÁLISE: O gráfico apresenta uma tendência linear com resíduos próximos a reta. Se nota a presença de possíveis outliers. Ao realizar o teste formal de normalidade encontra-se um p-valor bem maior que 0.05, portanto não rejeitamos a hipótese nula de que os dados são distribuídos normalmente.

Teste de Homocedasticidade

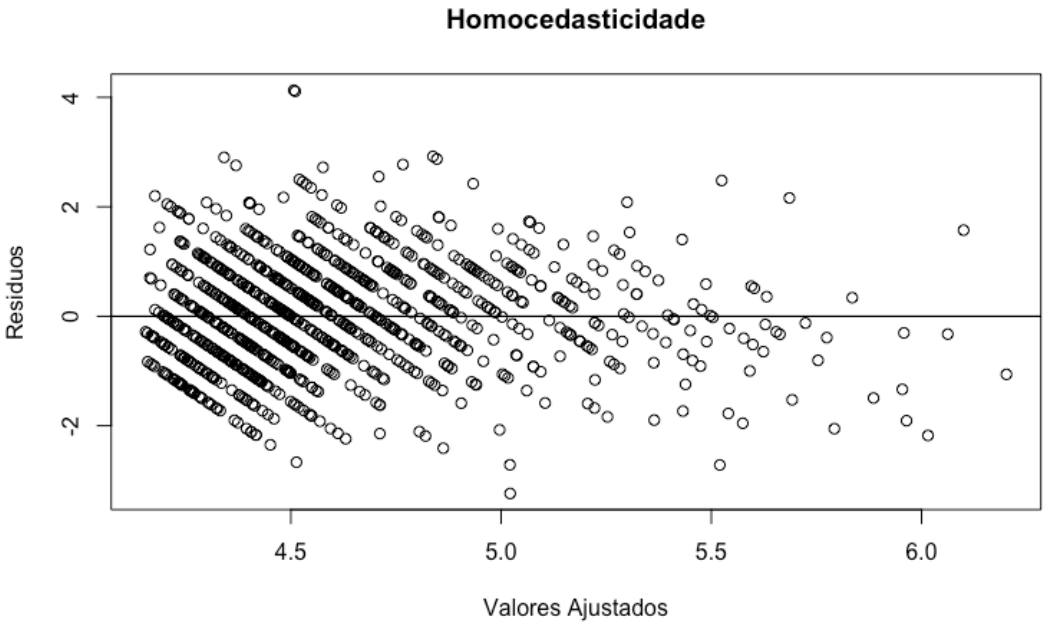
Hipóteses (de forma resumida):

- H_0: Homocedasticidade
- H_1: Heteroscedasticidade

Teste de Homocedasticidade - Gráfico

```
fit=fitted.values(mod2)   ### Valores ajustados da variável resposta pelo modelo gerado
plot(fit, anares, ylab="Residuos", xlab="Valores Ajustados",
main="Homocedasticidade")
```

abline(0,0)



Teste formal de Homocedasticidade
library(lmtest)
bptest(mod2)

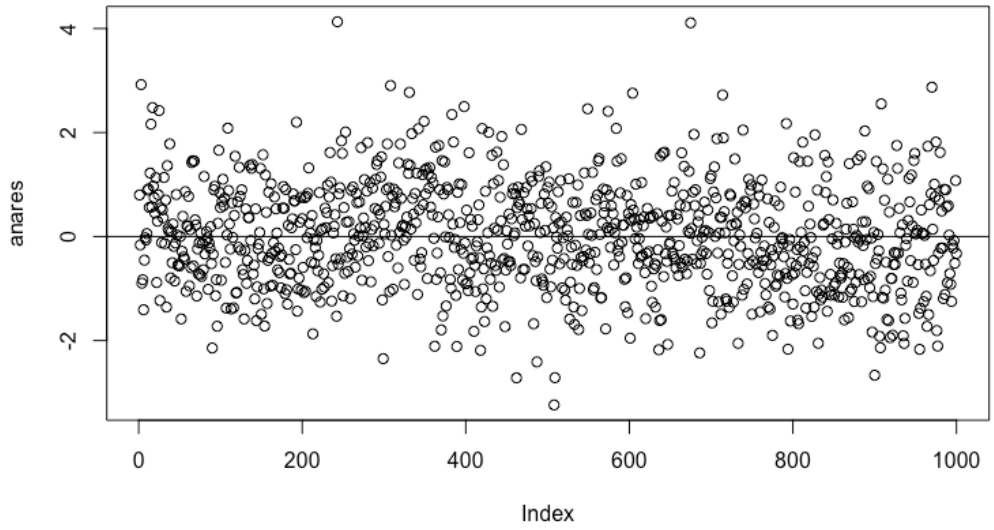
```
        studentized Breusch-Pagan test

data:  mod2
BP = 6.7019, df = 4, p-value = 0.1525
```

ANÁLISE: Não se rejeita a hipótese de que os dados sejam homocedásticos.

Teste de autocorrelação
Hipóteses:

- H_0: inicialmente sem autocorrelação
- H_1: inicialmente existe correlação

Teste de Autocorrelação - Gráfico
plot(anares)
abline(0,0)



Teste formal de Autocorrelação
dwtest(mod2)

```
        Durbin-Watson test

data:  mod2
DW = 1.9414, p-value = 0.1751
alternative hypothesis: true autocorrelation is greater than 0
```

ANÁLISE: Não existem evidências de que há autocorrelação nos dados.

RESPOSTAS – BASE QUAKES (MAG):

a) Como mencionado anteriormente, de todos os modelos testados o que mais se ajustou foi o mod2, que se apresentou como o melhor modelo de regressão linear múltipla. Seu p-value confirma que o modelo é melhor do que o nulo e ele apresenta bom ajuste aos dados, e com ele, todas as suposições foram satisfeitas.

b) O modelo atendeu todos os pressupostos.

c) Previsão e valores de MAG segundo o modelo linear. O modelo parece bem ajustado dados a proximidade entre a visão real e a previsão:

predict(mod2, interval="predict")

	fit	lwr	upr
1	4.646071	4.267300	5.024841
2	4.231062	3.852026	4.610099
3	4.837914	4.458852	5.216977
4	4.272523	3.893527	4.651519
5	4.159755	3.780660	4.538850
6	4.270922	3.892006	4.649838

O intervalo de confiança não se apresenta muito distante do valor real:

predict(mod2, interval="confidence")

	fit	lwr	upr
1	4.646071	4.626807	4.665334
2	4.231062	4.207129	4.254995
3	4.837914	4.813573	4.862256
4	4.272523	4.249241	4.295804
5	4.159755	4.134915	4.184594
6	4.270922	4.248986	4.292858

TREES (VOLUME)

A descrição do dataset:

Este conjunto de dados fornece medidas do diâmetro, altura e volume da madeira em 31 cerejeiras pretas abatidas. Observe que o diâmetro (em polegadas) está erroneamente rotulado como circunferência nos dados. É medido a 4 pés 6 polegadas acima do solo.

Atributos:

- Volume: Volume de madeira em pés cúbicos (nossa variável resposta);
- Girth: Diâmetro da árvore em polegadas (em vez de em perímetros);
- Height: Altura em pés.

Visualização das 6 primeiras linhas dos dados desse dataset:

help("trees")
df <- trees
head(df)

	Girth	Height	Volume
1	8.3	70	10.3
2	8.6	65	10.3
3	8.8	63	10.2
4	10.5	72	16.4
5	10.7	81	18.8
6	10.8	83	19.7

Checando as correlações:

cor(df)

```
           Girth    Height    Volume
Girth  1.0000000 0.5192801 0.9671194
Height 0.5192801 1.0000000 0.5982497
Volume 0.9671194 0.5982497 1.0000000
```

ANÁLISE: É possível dizer que a variável “Girth” possui forte correlação com a variável resposta “Volume” e que também se correlaciona c om “Height”. Nenhuma das variáveis explicativas apresentam considerável correlação entre si, portanto não apresentam problemas de multicolinearidade.

Testando possíveis modelos:

```
mod <- lm(Volume ~., data=df)
summary(mod)
```

```
mod <- lm(Volume ~ Girth, data=df)
summary(mod)
```

```
mod <- lm(Volume ~ Height, data=df)
summary(mod)
```

```
mod <- lm(Volume ~ sqrt(Girth) + sqrt(Height), data=df)
summary(mod)
```

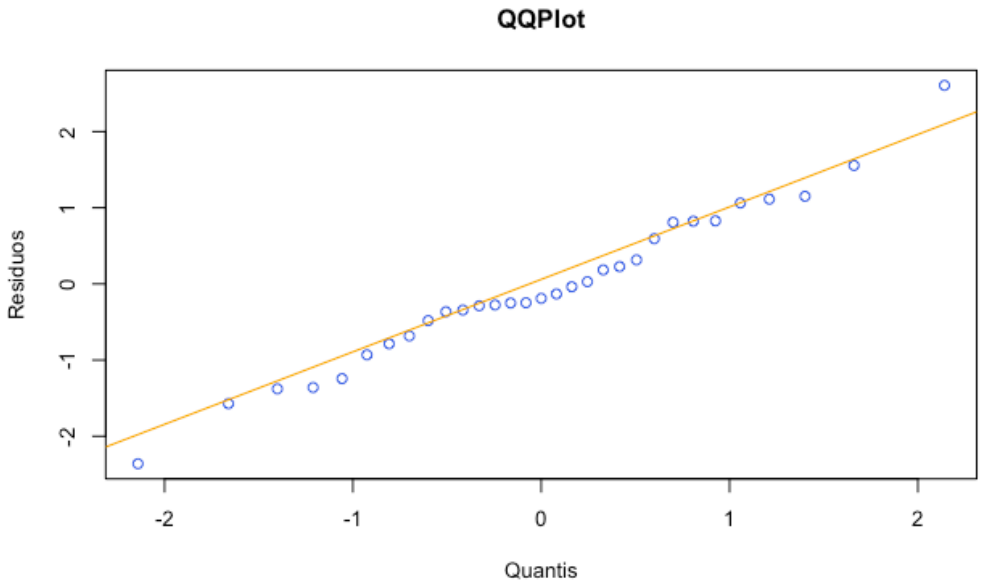
mod <-lm(Volume ~. -1, data=df) **## melhor modelo**
summary(mod)

```
Call:
lm(formula = Volume ~ . - 1, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-14.159  -3.490  -1.107   4.144  14.620

Coefficients:
       Estimate Std. Error t value Pr(>|t|)
Girth   5.04401    0.41187  12.247 5.52e-13 ***
Height -0.47732    0.07347  -6.497 4.12e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.162 on 29 degrees of freedom
Multiple R-squared:  0.9697,    Adjusted R-squared:  0.9676
F-statistic: 463.9 on 2 and 29 DF,  p-value: < 2.2e-16
```

```
mod=glm(log(Volume) ~ ., family = Gamma, data=df)
summary(mod)
```

```
mod=glm(Volume ~., family = Gamma, data=df)
summary(mod)
```

```
mod=glm(sqrt(Volume) ~ ., family = inverse.gaussian, data=df)
summary(mod)

mod=lm(Volume ~ log(.), data=df)
summary(mod)

mod <- glm.nb(Volume^0.5 ~ ., data=df)
summary(mod)

## Análise de Resíduos com modelo que teve melhor R2.
anares <- rstandard(mod)
par(mfrow=c(2,2))

### Teste de Normalidade – Gráfico
qqnorm(anares, ylab="Residuos", xlab="Quantis", main="QQPlot", col="royalblue")
qqline(anares, col="orange")
```



Teste formais de Normalidade

Hipóteses:
- H_0: Normalidade
- H_1: Não Normalidade

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

```
        Anderson-Darling normality test

data:  anares
A = 0.26786, p-value = 0.6612
```

ANÁLISE: O gráfico apresenta uma tendência linear com resíduos próximos a reta. Não se nota a presença de possíveis outliers. Ao realizar o teste formal de normalidade encontra-se um p-valor bem maior que 0.05, portanto não rejeitamos a hipótese nula de que os dados são distribuídos normalmente.

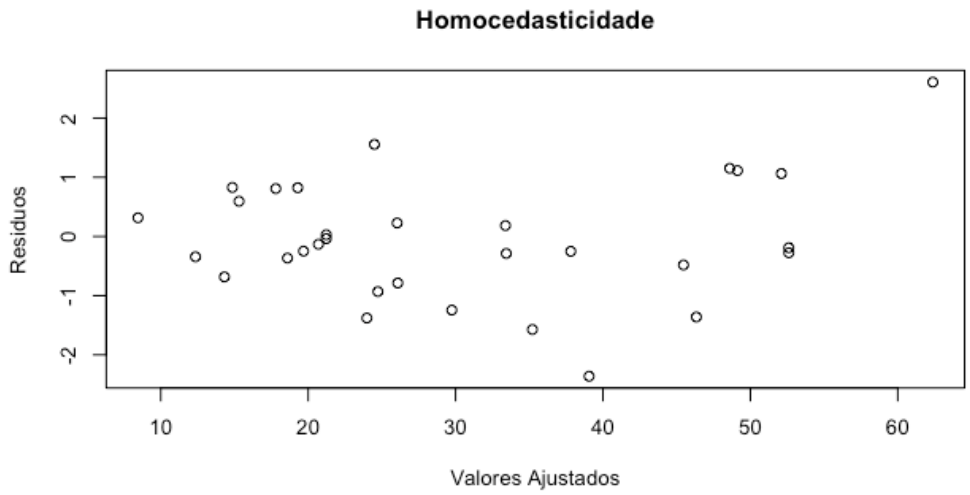
Teste de Homocedasticidade

Hipóteses (de forma resumida):

- H_0: Homocedasticidade
- H_1: Heteroscedasticidade

Teste de Homocedasticidade – Gráfico

```
fit=fitted.values(mod)    ### Valores ajustados da variável resposta pelo modelo gerado
plot(fit, anares, ylab="Residuos", xlab="ValoresAjustados",
main="Homocedasticidade")
```



Teste formal de Homocedasticidade

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

```
        studentized Breusch-Pagan test

data:  mod
BP = 6.195, df = 1, p-value = 0.01281
```

ANÁLISE: Não se rejeita a hipótese de que os dados sejam homocedásticos.

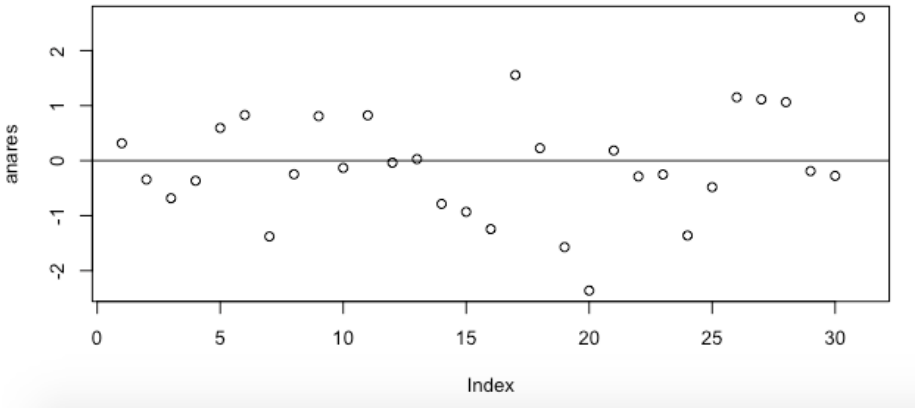
Teste de autocorrelação

Hipóteses:

- H_0: inicialmente sem autocorrelação
- H_1: inicialmente existe correlação

Teste de Autocorrelação – Gráfico

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



Teste formal de Autocorrelação

```
dwtest(mod)
```

```
        Durbin-Watson test

data:  mod
DW = 1.4957, p-value = 0.05211
alternative hypothesis: true autocorrelation is greater than 0
```

ANÁLISE: Não existem evidências de que há autocorrelação nos dados.

RESPOSTAS – TREES (Volume):

a)Como mencionado, de todos os modelos testados o que mais se ajustou foi o mod assinalado no qual foi retirado o intercepto. Seu p-value confirma que o modelo é melhor do que o nulo e ele apresenta excelente ajuste aos dados, R2 de 96%. Com ele, todas as suposições foram satisfeitas.

b) O modelo atendeu todos os pressupostos.

c) Previsão e valores de “Volume” segundo o modelo linear. O modelo parece bem ajustado dados a proximidade entre a visão real e a previsão:

```
predict(mod, interval="predict")
```

```
         fit        lwr      upr
1   8.452922 -4.7432964 21.64914
2  12.352721 -0.6070641 25.31251
3  14.316161  1.4403511 27.19197
4  18.595102  5.6927323 31.49747
5  15.308031  2.1512905 28.46477
6  14.857793  1.6418574 28.07373
```

O intervalo de confiança não se apresenta muito distante do valor real:
predict(mod2, interval="confidence")

```
         fit        lwr      upr
1   8.452922  4.536633 12.36921
2  12.352721  9.327302 15.37814
3  14.316161 11.673508 16.95881
4  18.595102 15.825937 21.36427
5  15.308031 11.526901 19.08916
6  14.857793 10.875572 18.84001
```

COLLEGE (APPS)

A descrição do dataset:

Esse dataset possui 777 linhas e 19 colunas.

Atributos:

- **Apps**: Number of applications received (Nossa variável resposta)
- **Private**: Public/private indicator
- **Accept**: Number of applicants accepted
- **Enroll**: Number of new students enrolled
- **Top10perc**: New students from top 10 % of high school class
- **Top25perc**: New students from top 25 % of high school class
- **F.Undergrad**: Number of full-time undergraduates
- **P.Undergrad**: Number of part-time undergraduates
- **Outstate**: Out-of-state tuition
- **Room.Board**: Room and board costs
- **Books**: Estimated book costs
- **Personal**: Estimated personal spending
- **PhD**: Percent of faculty with Ph.D.’s
- **Terminal**: Percent of faculty with terminal degree
- **S.F.Ratio**: Student/faculty ratio
- **perc.alumni**: Percent of alumni who donate Expend
- **Instructional**: expenditure per student
- **Grad.Rate**: Graduation rate

Análise inicial no dataset:

```
dados <- read.csv2(file.choose(), head=T)
df <- dados
head(df)
```

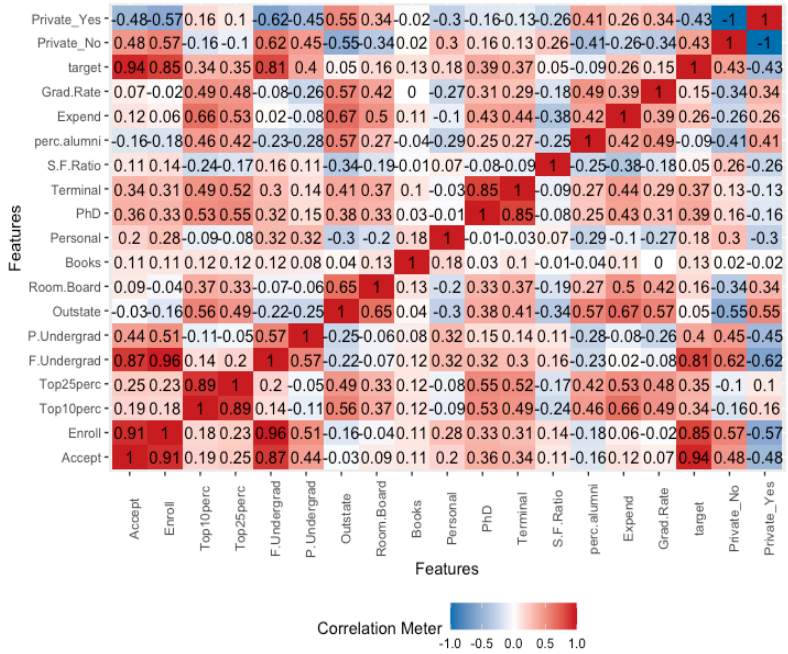
```
                            X Private Apps Accept Enroll Top10perc Top25perc
1 Abilene Christian University     Yes 1660   1232    721        23        52
2           Adelphi University     Yes 2186   1924    512        16        29
3               Adrian College     Yes 1428   1097    336        22        50
4          Agnes Scott College     Yes  417    349    137        60        89
5    Alaska Pacific University     Yes  193    146     55        16        44
6            Albertson College     Yes  587    479    158        38        62
  F.Undergrad P.Undergrad Outstate Room.Board Books Personal PhD Terminal
1        2885         537     7440       3300   450     2200  70       78
2        2683        1227    12280       6450   750     1500  29       30
3        1036          99    11250       3750   400     1165  53       66
4         510          63    12960       5450   450      875  92       97
5         249         869     7560       4120   800     1500  76       72
6         678          41    13500       3335   500      675  67       73
  S.F.Ratio perc.alumni Expend Grad.Rate
1      18.1          12   7041        60
2      12.2          16  10527        56
3      12.9          30   8735        54
4       7.7          37  19016        59
5      11.9           2  10922        15
6       9.4          11   9727        55
```

```
names(dados)
```

```
 [1] "X"           "Private"     "Apps"        "Accept"      "Enroll"
 [6] "Top10perc"   "Top25perc"   "F.Undergrad" "P.Undergrad" "Outstate"
[11] "Room.Board"  "Books"       "Personal"    "PhD"         "Terminal"
[16] "S.F.Ratio"   "perc.alumni" "Expend"      "Grad.Rate"
```

Checando as correlações:

```
DataExplorer::plot_correlation(df)
```



ANÁLISE: Existem diversas variáveis correlacionadas.

Testando possíveis modelos:

```
mod <- lm(Apps ~., data=df)
summary(mod)
```

```
mod <- lm(Apps ~ Private + Accept + Enroll + Top10perc + Outstate + Expend,
data=df)
summary(mod)
```

```
mod <- lm(Apps ~ Accept + Enroll + Top10perc + Outstate + Expend, data=df)
summary(mod)
```

```
mod <- lm(Apps ~ Private + sqrt(Accept) + sqrt(Enroll) + Top10perc + Outstate +
Expend, data=df)
summary(mod)
```

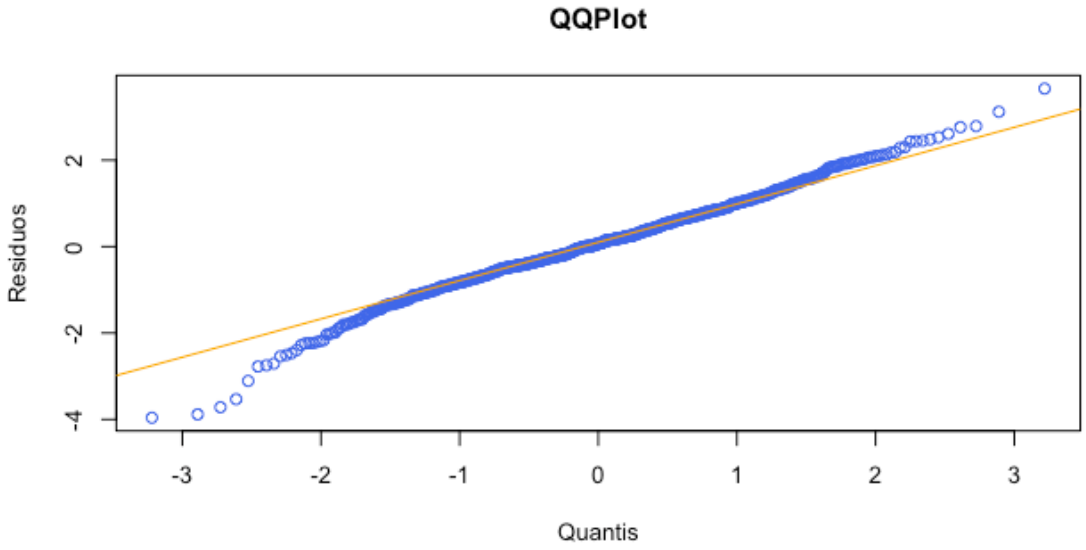
```
mod <-lm(log(Apps) ~. -1, data=df) ### Melhor Modelo
mod_step <- step(mod, direction="backward")
summary(mod_step)
```

```
Call:
lm(formula = log(Apps) ~ X + Private + Accept + Enroll + Top10perc +
    Top25perc + P.Undergrad + Outstate + Room.Board + Books +
    Personal + PhD + Terminal + S.F.Ratio + perc.alumni + Grad.Rate -
    1, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.0466 -0.3820  0.0538  0.5364  2.8333

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
X            5.012e-04  1.283e-04   3.906 0.000102 ***
Private      6.132e-01  8.180e-02   7.497 1.82e-13 ***
Accept       1.030e-04  3.021e-05   3.411 0.000682 ***
Enroll       4.699e-04  8.504e-05   5.525 4.53e-08 ***
Top10perc   -2.026e-02  3.645e-03  -5.560 3.73e-08 ***
Top25perc    1.489e-02  3.206e-03   4.643 4.04e-06 ***
P.Undergrad  3.721e-05  2.321e-05   1.603 0.109374
Outstate    -3.356e-05  1.296e-05  -2.589 0.009817 **
Room.Board   1.366e-04  3.599e-05   3.796 0.000159 ***
Books        1.132e-03  1.734e-04   6.528 1.22e-10 ***
Personal     2.603e-04  4.547e-05   5.724 1.50e-08 ***
PhD          1.478e-02  3.431e-03   4.307 1.87e-05 ***
Terminal     2.204e-02  3.651e-03   6.037 2.45e-09 ***
S.F.Ratio    2.998e-03  6.414e-04   4.675 3.48e-06 ***
perc.alumni -7.387e-03  3.049e-03  -2.423 0.015646 *
Grad.Rate    1.808e-02  2.131e-03   8.483  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7772 on 761 degrees of freedom
Multiple R-squared:  0.9895,    Adjusted R-squared:  0.9893
F-statistic:  4479 on 16 and 761 DF,  p-value: < 2.2e-16
```

```
mod=glm(log(Apps) ~ ., family = Gamma, data=df)
summary(mod)

mod=glm(Apps ~., family = Gamma, data=df)
summary(mod)

mod=glm(sqrt(Apps) ~ ., family = inverse.gaussian, data=df)
summary(mod)

mod=lm(Apps ~ log(.), data=df)
summary(mod)

mod <- glm.nb(Volume^0.5 ~ ., data=df)
summary(mod)

mod <- glm(Apps~., family = poisson(link = "log"), data = df)
summary(mod)
```

Análise de Resíduos com modelo que teve melhor R2 (acima citado).
anares <- rstandard(mod)
par(mfrow=c(2,2))

Teste de Normalidade – Gráfico
qqnorm(anares, ylab="Residuos", xlab="Quantis", main="QQPlot", col="royalblue")
qqline(anares, col="orange")



Teste formais de Normalidade

Hipóteses:
- H_0: Normalidade
- H_1: Não Normalidade

library(nortest)
ad.test(anares)

```
        Anderson-Darling normality test

data:  anares
A = 1.6176, p-value = 0.0003702
```

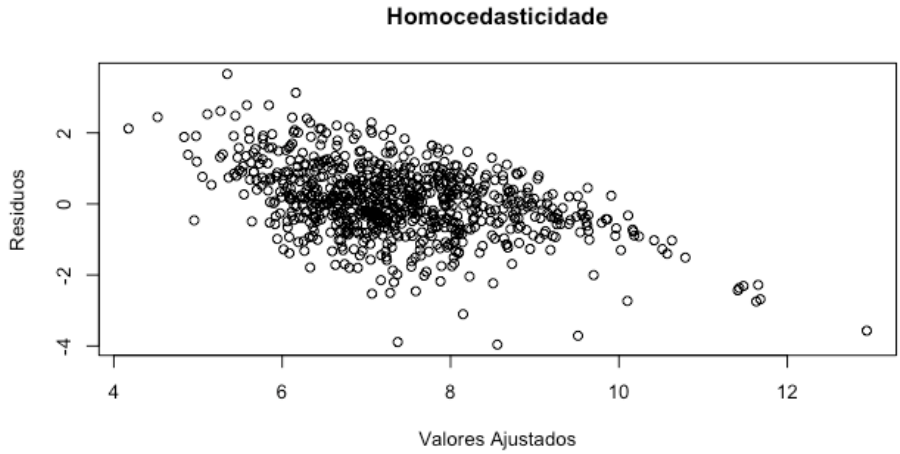
ANÁLISE: O gráfico apresenta uma tendência linear com resíduos próximos a reta. Se nota a presença de possíveis outliers. Ao realizar o teste formal de normalidade encontra-se um p-valor bem menor que 0.05, portanto temos indicação de que os resíduos não seguem distribuição normal.

Teste de Homocedasticidade

Hipóteses (de forma resumida):

- H_0: Homocedasticidade
- H_1: Heteroscedasticidade

Teste de Homocedasticidade – Gráfico
fit=fitted.values(mod) ### Valores ajustados da variável resposta pelo modelo gerado
plot(fit, anares, ylab="Residuos", xlab="ValoresAjustados",
main="Homocedasticidade")



Teste formal de Homocedasticidade

library(lmtest)
bptest(mod)

```
        studentized Breusch-Pagan test

data:  mod
BP = 79.469, df = 17, p-value = 4.764e-10
```

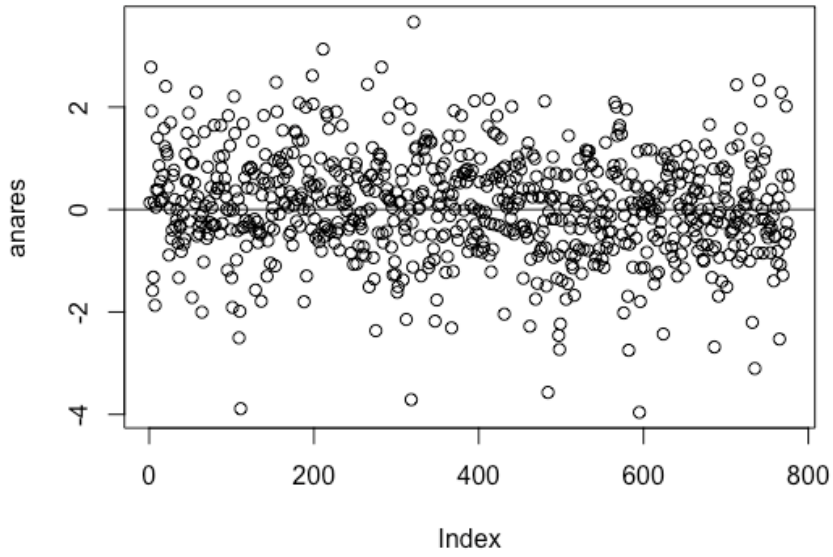
ANÁLISE: Rejeita a hipótese de que os dados sejam homocedásticos.

Teste de autocorrelação
Hipóteses:

- H_0: inicialmente sem autocorrelação
- H_1: inicialmente existe correlação

Teste de Autocorrelação – Gráfico

plot(anares)
abline(0,0)



Teste formal de Autocorrelação
dwtest(mod)

```
        Durbin-Watson test

data:  mod
DW = 1.7475, p-value = 0.0001711
alternative hypothesis: true autocorrelation is greater than 0
```

ANÁLISE: Rejeita H_0 já que existem evidências de autocorrelação nos dados.

RESPOSTAS – COLLEGE (Apps):

a) O modelo parece não estar completamente ajustado aos dados.

b) Nem todas as suposições foram satisfeitas.

c) Previsão e valores de Apps segundo o modelo linear. O modelo parece não estar muito bem ajustado dados a proximidade entre a visão real e a previsão:
predict(mod, interval="predict")

	fit	lwr	upr
1	7.311147	5.772678	8.849617
2	5.568083	4.013036	7.123129
3	5.776138	4.242980	7.309296
4	7.247679	5.701629	8.793730
5	6.273179	4.720908	7.825450
6	6.260115	4.710466	7.809765

O intervalo de confiança se apresenta distante do valor real:
predict(mod, interval="confidence")

```
        fit       lwr       upr
1  7.311147  7.124994  7.497301
2  5.568083  5.274938  5.861228
3  5.776138  5.640719  5.911556
4  7.247679  7.006772  7.488587
5  6.273179  5.995131  6.551227
6  6.260115  5.997099  6.523132
```

CREDIT (BALANCE)

A descrição do dataset:

Esse dataset possui 400 linhas e 11 colunas.

Atributos:

Descrição de cada atributo:

- **balance**: average credit card debt for a number of individuals age (in years) NOSSA VARIÁVEL RESPOSTA)
- **cards**: number of credit cards
- **education**: years of education
- **income**: in thousands of dollars
- **limit**: credit limit
- **rating**: credit rating
- **gender**: gender
- **student**: student status
- **status**: marital status
- **ethnicity**: Caucasian, African American or Asian.

Análise inicial no dataset:

```
dados <- read.csv2(file.choose(), head=T)
df <- dados
head(df)
```

```
   Income Limit Rating Cards Age Education Gender Student Married
1  14.891  3606    283     2  34        11   Male      No     Yes
2 106.025  6645    483     3  82        15 Female     Yes     Yes
3 104.593  7075    514     4  71        11   Male      No      No
4 148.924  9504    681     3  36        11 Female      No      No
5  55.882  4897    357     2  68        16   Male      No     Yes
6   80.18  8047    569     4  77        10   Male      No      No
  Ethnicity Balance
1 Caucasian     333
2     Asian     903
3     Asian     580
4     Asian     964
5 Caucasian     331
6 Caucasian    1151
```

names(dados)

```
 [1] "Income"    "Limit"     "Rating"    "Cards"     "Age"
 [6] "Education" "Gender"    "Student"   "Married"   "Ethnicity"
[11] "Balance"
```

Checando as correlações:
cor(df)

	Income	Limit	Rating	Cards
Income	1.000000000	0.143989177	0.134799839	-0.082734361
Limit	0.143989177	1.000000000	0.996879737	0.010231333
Rating	0.134799839	0.996879737	1.000000000	0.053239030
Cards	-0.082734361	0.010231333	0.053239030	1.000000000
Age	-0.001035415	0.100887922	0.103164996	0.042948288
Education	-0.057000466	-0.023548534	-0.030135627	-0.051084217
Gender	-0.061952771	-0.009396678	-0.008884590	0.022658021
Student	-0.051554574	-0.006015094	-0.002027646	-0.026164127
Married	-0.009824417	0.031154829	0.036750773	-0.009695060
Ethnicity	0.059349148	-0.020837489	-0.020287751	-0.003866747
Balance	1.000000000	0.143989177	0.134799839	-0.082734361

	Age	Education	Gender	Student
Income	-0.001035415	-0.057000466	-0.061952771	-0.051554574
Limit	0.100887922	-0.023548534	-0.009396678	-0.006015094
Rating	0.103164996	-0.030135627	-0.008884590	-0.002027646
Cards	0.042948288	-0.051084217	0.022658021	-0.026164127
Age	1.000000000	0.003619285	-0.004015496	-0.029844426
Education	0.003619285	1.000000000	0.005049071	0.072085400
Gender	-0.004015496	0.005049071	1.000000000	-0.055033718
Student	-0.029844426	0.072085400	-0.055033718	1.000000000
Married	-0.073135503	0.048910587	-0.012451711	-0.076973701
Ethnicity	-0.032451326	-0.030054892	-0.001513996	-0.030261377
Balance	-0.001035415	-0.057000466	-0.061952771	-0.051554574

	Married	Ethnicity	Balance
Income	-0.009824417	0.059349148	1.000000000
Limit	0.031154829	-0.020837489	0.143989177
Rating	0.036750773	-0.020287751	0.134799839
Cards	-0.009695060	-0.003866747	-0.082734361
Age	-0.073135503	-0.032451326	-0.001035415
Education	0.048910587	-0.030054892	-0.057000466
Gender	-0.012451711	-0.001513996	-0.061952771
Student	-0.076973701	-0.030261377	-0.051554574
Married	1.000000000	0.060562584	-0.009824417
Ethnicity	0.060562584	1.000000000	0.059349148
Balance	-0.009824417	0.059349148	1.000000000

ANÁLISE: Se nota que existem poucas variáveis correlacionadas.

Testando possíveis modelos:

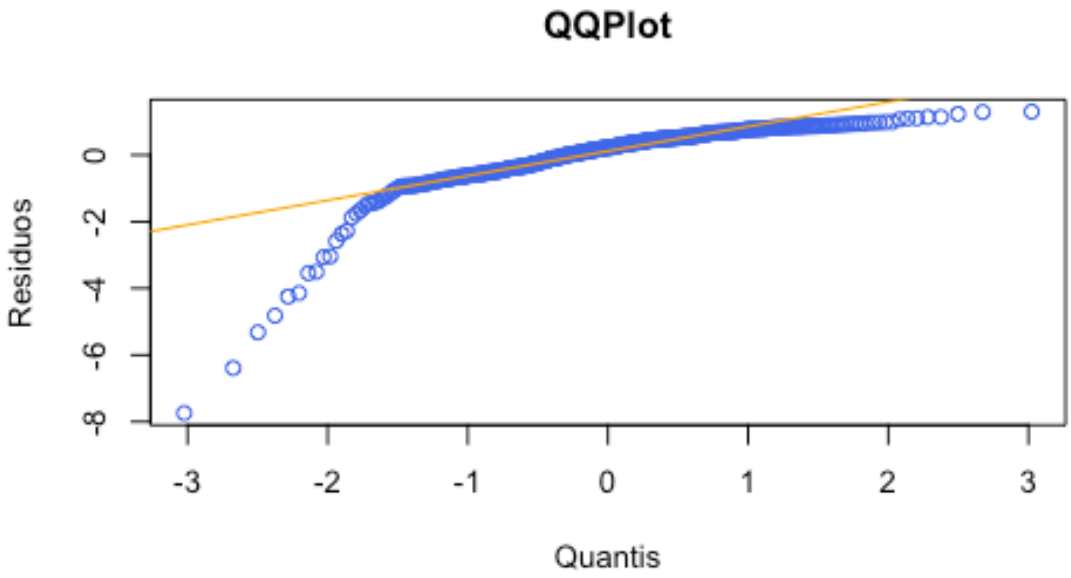
mod <- lm(Balance ~., data=df)
summary(mod)

```
mod <- lm(Balance ~.-1, data=df)
summary(mod)

mod <- lm(Balance ~ Rating-1, data=df)
summary(mod)

mod <- lm(log(Balance) ~ log(Rating)-1, data=df)
summary(mod)

mod <- lm(Balance ~ sqrt(Rating), data=df)
summary(mod)

mod <-lm(log(Balance) ~.-1, data=df) ### melhor modelo
mod_step <- step(mod, direction="both")
summary(mod_step)
```

```
Call:
lm(formula = log(Balance) ~ Income + Limit + Rating + Age + Education +
    Gender + Student + Married + Ethnicity - 1, data = df)

Residuals:
    Min      1Q  Median      3Q     Max
-3.9794 -0.2910  0.0701  0.4917  1.2585

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
Income     0.0083772  0.0002664  31.446  < 2e-16 ***
Limit     -0.0007595  0.0001636  -4.641 4.73e-06 ***
Rating     0.0113511  0.0024281   4.675 4.05e-06 ***
Age        0.0078207  0.0016930   4.619 5.23e-06 ***
Education  0.0578044  0.0089259   6.476 2.83e-10 ***
Gender     0.2598495  0.0580989   4.473 1.01e-05 ***
Student    0.5596919  0.0950541   5.888 8.42e-09 ***
Married    0.2186851  0.0604905   3.615 0.000339 ***
Ethnicity  0.1258699  0.0359939   3.497 0.000525 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.6178 on 391 degrees of freedom
Multiple R-squared:  0.9856,    Adjusted R-squared:  0.9853
F-statistic:  2976 on 9 and 391 DF,  p-value: < 2.2e-16
```

```
mod=glm(log(Balance) ~ Rating + Income, family = Gamma, data=df)
summary(mod)

mod=glm(sqrt(Balance) ~ Rating + Income, family = inverse.gaussian, data=df)
summary(mod)

mod=lm(Balance ~ log(.), data=df)
summary(mod)

mod <- glm.nb(Balance^0.5 ~ ., data=df)
```

```
summary(mod)

mod <- glm(Balance~., family = poisson(link = "log"), data = df)
summary(mod)

## Análise de Resíduos com modelo que teve melhor R2 (acima citado).
anares <- rstandard(mod)
par(mfrow=c(2,2))

### Teste de Normalidade – Gráfico
qqnorm(anares, ylab="Residuos", xlab="Quantis", main="QQPlot", col="royalblue")
qqline(anares, col="orange")
```




```
### Teste formais de Normalidade

Hipóteses:
- H_0: Normalidade
- H_1: Não Normalidade

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

```
        Anderson-Darling normality test

data:  anares
A = 19.668, p-value < 2.2e-16
```

ANÁLISE: Se nota que a maioria dos atributos não foram suficientes para explicar a variável "`Balance`" porém como nenhuma variável apresentou evidência estatística para não rejeitar a hipótese nula de normalidade no teste de Aderson-Darling os resultados são inconclusivos.

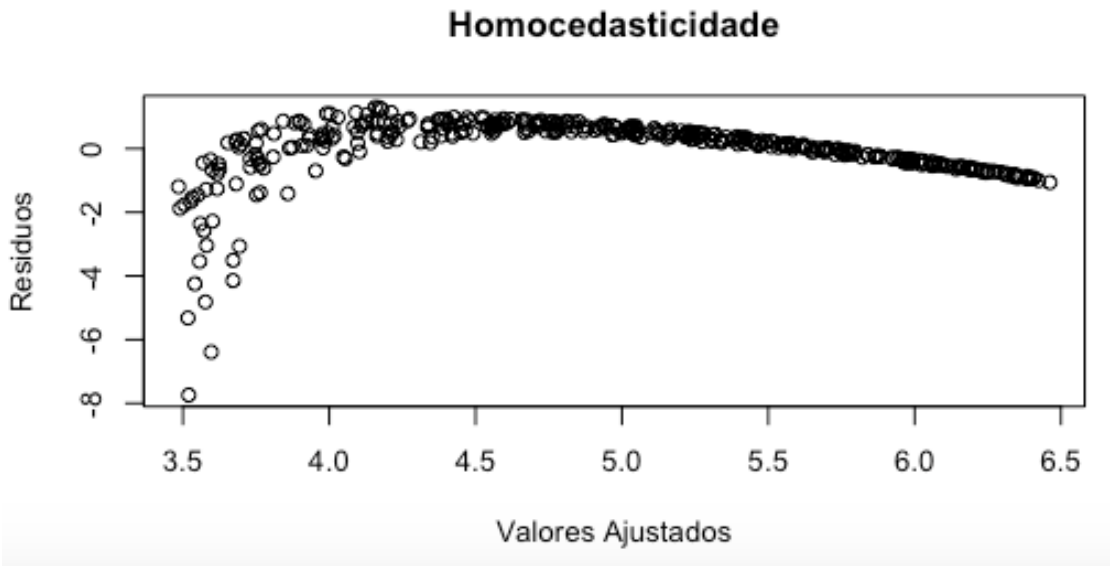
Teste de Homocedasticidade

Hipóteses (de forma resumida):

- H_0: Homocedasticidade
- H_1: Heteroscedasticidade

Teste de Homocedasticidade – Gráfico

```
fit=fitted.values(mod)    ### Valores ajustados da variável resposta pelo modelo gerado
plot(fit, anares, ylab="Residuos", xlab="ValoresAjustados",
main="Homocedasticidade")
```



Teste formal de Homocedasticidade

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

```
        studentized Breusch-Pagan test

data:  mod
BP = 23.931, df = 9, p-value = 0.004412
```

ANÁLISE: Rejeitada a hipótese de que os dados sejam homocedásticos.

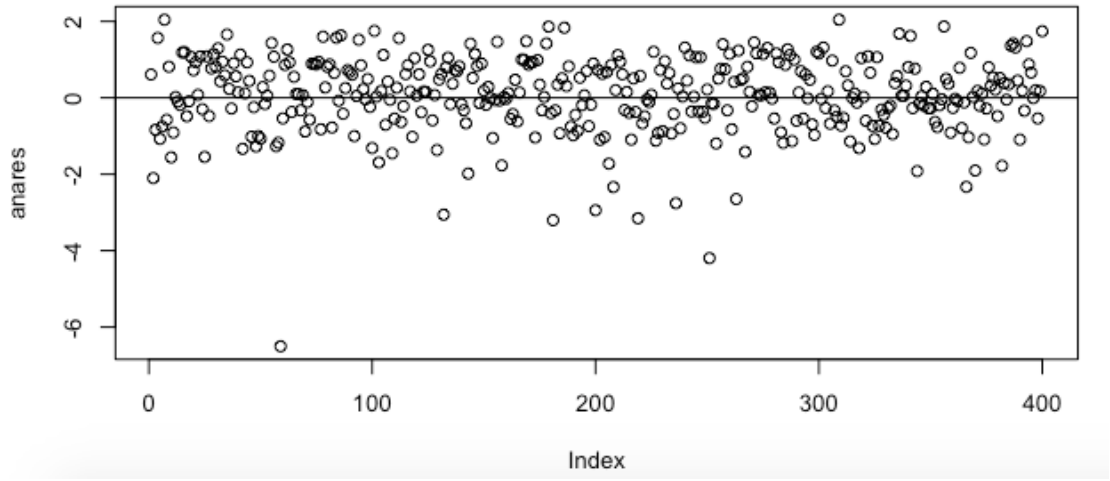
Teste de autocorrelação

Hipóteses:

- H_0: inicialmente sem autocorrelação
- H_1: inicialmente existe correlação

Teste de Autocorrelação – Gráfico

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




```
### Teste formal de Autocorrelação
dwtest(mod)
```

```
        Durbin-Watson test

data:  mod
DW = 1.954, p-value = 0.3197
alternative hypothesis: true autocorrelation is greater than 0
```

ANÁLISE: Não rejeita H_0 já que não existem evidências de autocorrelação nos dados. Suposição satisfeita.

RESPOSTAS – CREDIT (Balance):

a) O modelo parece estar ajustado aos dados como modelo explicativo com R2 alto, mas apenas um pressuposto foi atendido, o que invalida interpretações de intervalos de confiança das estimativas do modelo

b) Nem todas as suposições foram satisfeitas.

c) Previsão e valores de “Balance” segundo o modelo linear.
O modelo parece não estar muito bem ajustado dados a proximidade entre a visão real e a previsão:

```
predict(mod, interval="predict")
```

```
        fit      lwr      upr
1  3.801694 2.575271 5.028117
2  4.164541 2.926708 5.402375
3  3.362590 2.134380 4.590800
4  3.311882 2.080402 4.543363
5  6.451461 5.229127 7.673794
6  6.437499 5.204416 7.670582
```

O intervalo de confiança se apresenta distante do valor real:
predict(mod, interval="confidence")

```
        fit      lwr      upr
1  3.801694 3.632517 3.970871
2  4.164541 3.926342 4.402740
3  3.362590 3.180912 3.544268
4  3.311882 3.109272 3.514493
5  6.451461 6.315052 6.587869
6  6.437499 6.225366 6.649633
```

ADVERTISING (SALES)

A descrição do dataset:

Esse dataset possui 200 linhas e 4 colunas.

Atributos:

- **Sales** in millions USD **(Variável resposta);**
- **TV** thousands of dollars in tv advertising;
- **radio** thousands of dollars in radio advertising;
- **newspaper** thousands of dollars in newspaper advertising.

Análise inicial no dataset:

dados <- read.csv2(file.choose(), head=T)
df <- dados
head(df)

```
     TV radio newspaper sales
1 230.1  37.8      69.2  22.1
2  44.5  39.3      45.1  10.4
3  17.2  45.9      69.3   9.3
4 151.5  41.3      58.5  18.5
5 180.8  10.8      58.4  12.9
6   8.7  48.9      75.0   7.2
```

names(dados)

```
[1] "TV"        "radio"     "newspaper" "sales"
```

Checando as correlações:

```
cor(df)
```

	TV	radio	newspaper	sales
TV	1.00000000	0.05480866	0.05664787	0.7822244
radio	0.05480866	1.00000000	0.35410375	0.5762226
newspaper	0.05664787	0.35410375	1.00000000	0.2282990
sales	0.78222442	0.57622257	0.22829903	1.0000000

ANÁLISE: Nenhuma das variáveis explicativas apresentam correlação forte entre si (sem problema de multicolinearidade).

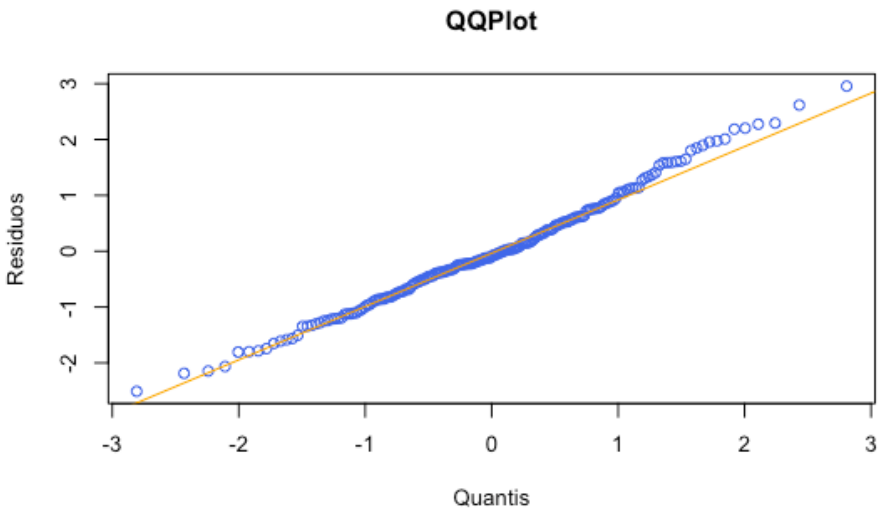
Testando possíveis modelos:

```
mod <- lm(sales ~., data=df)
summary(mod)

mod <- lm(sales ~.-1, data=df)
summary(mod)

mod <- lm(log(sales) ~ log(.), data=df)
summary(mod)

mod <- lm(log(sales) ~ ., data=df)
summary(mod)

mod <- lm(sales ~ sqrt(.), data=df)
summary(mod)

mod <- lm(sales ~.-1, data=df)
mod_step <- step(mod, direction="both")
summary(mod_step)

mod <-lm(sqrt(sales) ~ sqrt(.)-1, data=df)
summary(mod)

mod <-lm(sales + tv + radio, data=df)
mod_step <- step(mod, direction="both")
summary(mod_step)

mod=glm(sales ~ ., family = Gamma, data=df)
summary(mod)

mod <- glm(sales ~ ., family = binomial(link = "logit"),
   data = df)
summary(mod)

mod=glm(sqrt(sales) ~ ., family = inverse.gaussian, data=df)
summary(mod)
```

```
mod <- glm.nb(sales^0.5 ~ ., data=df)
summary(mod)

### MELHOR MODELO
mod <- glm(sales^2 ~ TV + radio, family = gaussian, data=df)
mod_step <- step(mod, direction="both")
summary(mod_step)
```

```
Call:
glm(formula = sales^2 ~ TV + radio, family = gaussian, data = df)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-150.040   -41.080    -5.875    36.813   177.224

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -112.08467   10.66220  -10.51   <2e-16 ***
TV             1.30461    0.05034   25.92   <2e-16 ***
radio          6.18844    0.29109   21.26   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for gaussian family taken to be 3705.746)

    Null deviance: 5130946  on 199  degrees of freedom
Residual deviance:  730032  on 197  degrees of freedom
AIC: 2216.1

Number of Fisher Scoring iterations: 2
```

```
## Análise de Resíduos com o melhor modelo encontrado.
anares <- rstandard(mod)
par(mfrow=c(2,2))

### Teste de Normalidade – Gráfico
qqnorm(anares, ylab="Residuos", xlab="Quantis", main="QQPlot", col="royalblue")
qqline(anares, col="orange")
```



Teste formais de Normalidade

Hipóteses:
- H_0: Normalidade
- H_1: Não Normalidade

library(nortest)
ad.test(anares)

```
        Anderson-Darling normality test

data:  anares
A = 0.58116, p-value = 0.129
```

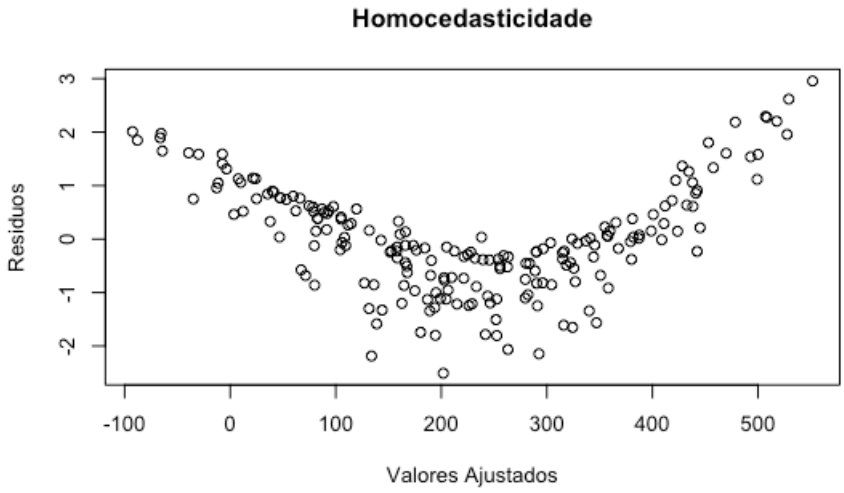
ANÁLISE: O gráfico apresenta uma tendência linear com resíduos próximos a reta. Se nota a presença de possíveis outliers. Ao realizar o teste formal de normalidade encontra-se um p-valor bem maior que 0.05, portanto não rejeitamos a hipótese nula de que os dados são distribuídos normalmente.

Teste de Homocedasticidade

Hipóteses (de forma resumida):

- H_0: Homocedasticidade
- H_1: Heteroscedasticidade

Teste de Homocedasticidade – Gráfico
fit=fitted.values(mod) ### Valores ajustados da variável resposta pelo modelo gerado
plot(fit, anares, ylab="Residuos", xlab="ValoresAjustados",
main="Homocedasticidade")



Teste formal de Homocedasticidade

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

```
        studentized Breusch-Pagan test

data:  mod
BP = 5.2091, df = 2, p-value = 0.07394
```

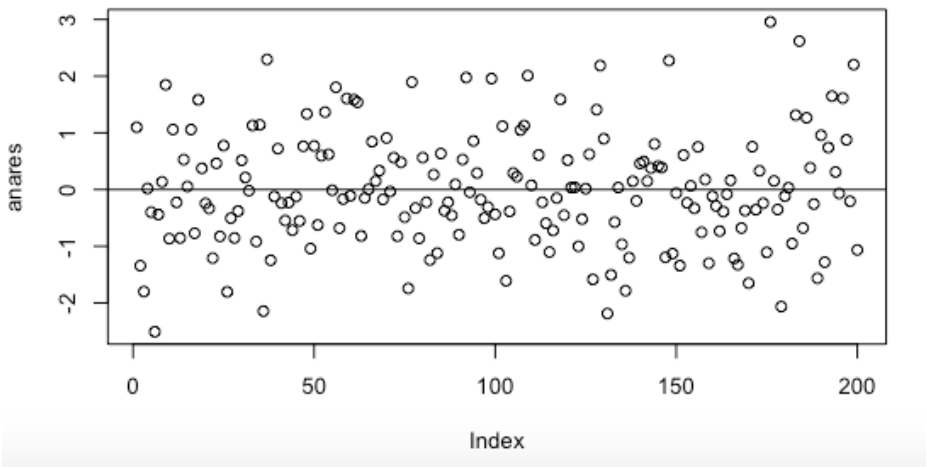
ANÁLISE: Não rejeita a hipótese de que os dados sejam homocedásticos, mesmo com o gráfico tendo indicado certo padrão.

Teste de autocorrelação
Hipóteses:

- H_0: inicialmente sem autocorrelação
- H_1: inicialmente existe correlação

Teste de Autocorrelação – Gráfico

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



Teste formal de Autocorrelação
dwtest(mod2)

```
        Durbin-Watson test

data:  mod
DW = 2.2015, p-value = 0.924
alternative hypothesis: true autocorrelation is greater than 0
```

ANÁLISE: Não existem evidências de que há autocorrelação nos dados. Suposição satisteita!

RESPOSTAS – ADVERTISING (sales):

a) O modelo parece estar bem ajustado aos dados, e todas as suposições foram satisfeitas, o que valida interpretações de intervalos de confiança das estimativas desse modelo.

b) O modelo atendeu a todos os pressupostos.

c) Previsão e valores de “sales” segundo o modelo linear.

O modelo parece bem ajustado dados a proximidade entre a visão real e a previsão:

predict(mod2, interval="predict")

	fit	lwr	upr
1	21.9413973	17.91062636	25.972168
2	11.8856909	7.85220574	15.919176
3	12.2908295	8.21672838	16.364931
4	18.3112160	14.28568939	22.336743
5	13.1077034	9.06457723	17.150830
6	12.5961339	8.50598899	16.686279

O intervalo de confiança não se apresenta muito distante do valor real:

predict(mod2, interval="confidence")

	fit	lwr	upr
1	21.9413973	21.4166987	22.4660959
2	11.8856909	11.3405330	12.4308488
3	12.2908295	11.4993154	13.0823437
4	18.3112160	17.8284544	18.7939776
5	13.1077034	12.4952792	13.7201275
6	12.5961339	11.7258007	13.4664672