Lista 7 [AD-UFPE-2019]

Antonio Fernandes

28 de maio de 2019

Conte?do

Apresenta??o

Este documento apresenta as respostas da lista de exerc?cios 7 da disciplina de An?lise de dados.

O link est? dispon?vel no GitHub: https://github.com/alvesat/AD_7

A lista envolve a execu??o de diversos modelos de regress?o linear multivariada envolvendo um banco de dados espec?ficos.

Quest?o 1

O primeiro passo para responder a quest?o ? abrir o banco de dados contendo os dados necess?rios. ? importante verificar que os dados est?o no formato *.dta*, que ? o formato relacionado ao *software* Stata. Devido a isso, teremos que abrir o pacote `haven` e executar o comando para abrir o banco.

```
library(haven)
```

```
## Warning: package 'haven' was built under R version 3.5.3
```

```
fair <- read_dta("~/Dados/Listas/AD_7/DATA_L7/fair.dta")
```

a) An?lise descritiva de todas as vari?veis da base de dados

Vamos verificar o nome das vari?veis presentes no banco

```r
names(fair)
```

```
## [1] "YEAR"      "VOTE"      "PARTY"     "PERSON"    "DURATION"  "WAR"
## [7] "GROWTH"    "INFLATION" "GOODNEWS"
```

Podemos observar que o banco apresenta 9 vari?veis: *Year, Vote, Party, Person, Duration, War, Growth, Inflation* e *Good news*.

Um outro passo envolvendo o processo de descri??o das vari?veis ? identificar a estrutura de cada vari?vel:

```r
str(fair)
```

```
## Classes 'tbl_df', 'tbl' and 'data.frame':    32 obs. of  9 variables:
##  $ YEAR     : num  1880 1884 1888 1892 1896 ...
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ VOTE     : num  50.2 49.8 50.4 48.3 47.8 ...
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ PARTY    : num  -1 -1 1 -1 1 -1 -1 -1 -1 1 ...
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ PERSON   : num  0 0 1 1 0 1 0 0 1 1 ...
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ DURATION : num  1.75 2 0 0 0 0 1 1.25 1.5 0 ...
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ WAR      : num  0 0 0 0 0 0 0 0 0 0 ...
##   ..- attr(*, "format.stata")= chr "%8.0g"
##  $ GROWTH   : num  3.88 1.59 -5.55 2.76 -10.02 ...
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ INFLATION: num  1.974 1.055 0.604 2.274 3.41 ...
##   ..- attr(*, "format.stata")= chr "%9.0g"
##  $ GOODNEWS : num  9 2 3 7 6 7 5 8 8 3 ...
##   ..- attr(*, "format.stata")= chr "%8.0g"
```

? poss?vel identificar que todas as vari?veis presentes no banco s?o n?mericas e que o banco possui 32 observa??es. Agora vamos fazer uma an?lise descritiva de cada vari?vel:

Year

A vari?vel ano apresenta os anos relacionados as outras vari?veis. Por meio do comando `fivenum`, podemos obter um sum?rio da vari?vel (valores min?mos, 1? quartil, mediana, 3? quartil e valor m?ximo). No caso dessa vari?vel, o importante ? verificarmos o come?o da an?lise e o m?ximo.

```r
fivenum(fair$YEAR)
```

```
## [1] 1880 1910 1942 1974 2004
```

Em rela??o a vari?vel Year, percebemos que o primeiro ano do banco ? 1880 e o ano final do banco ? 2004.

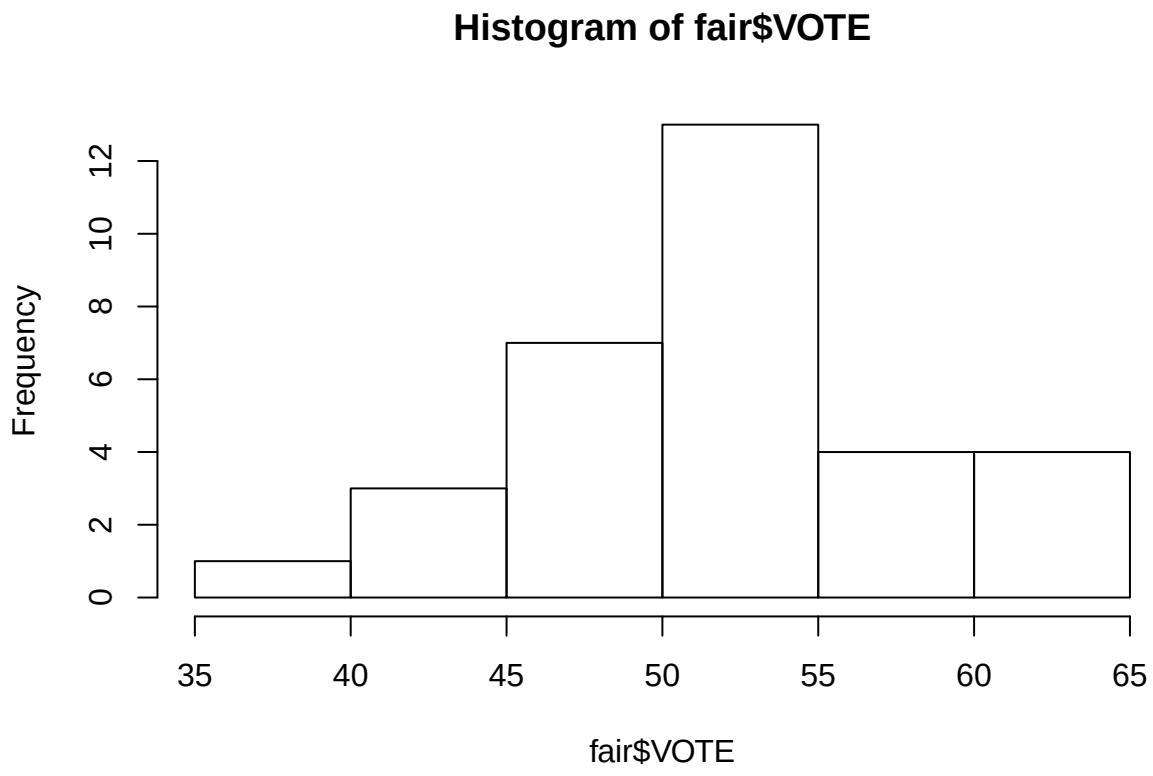
Vote

```
fivenum(fair$VOTE)
```

```
## [1] 36.1190 49.2720 52.0260 56.3815 62.4580
```

Em rela??o a vari?vel vote, a vari?vel representa a porcentagem de votos recebidas pelo partido do incumbente. Observando os valores obtidos do comando fivenum, o valor m?nimo da vari?vel foi de 36.12, o 1? quartil ? de 49.27, a mediana ? de 52.03, o 3? ? de 56.38 e o valor m?ximo ? de 62.46. Podemos analisar a vari?vel vote por meio de um histograma:

```
hist(fair$VOTE)
```



Com o histograma, verificamos que os valores est?o mais concentrados entre 45 e 55, ou seja, no centro da distribui??o.

Party

A vari?vel party ? bin?ria, apresentando como valores -1 e 1.

```
ftable(fair$PARTY)
```

```
##  -1  1
##
##  18 14
```

Com o comando `ftable`, podemos fazer uma tabela de frequ?ncia da vari?vel party e identificar que das 32 observa??es do banco, 18 s?o do valor -1 e 14 s?o do valor 1.

Person

A vari?vel person tamb?m ? bin?ria, apresentando como valores 1 e 0.

```
ftable(fair$PERSON)
```

```
##   0  1
##
##  13 19
```

O resultado mostra que 13 observa??es apresentam o valor 0 e 19 observa??es apresentam o valor 1.

Duration

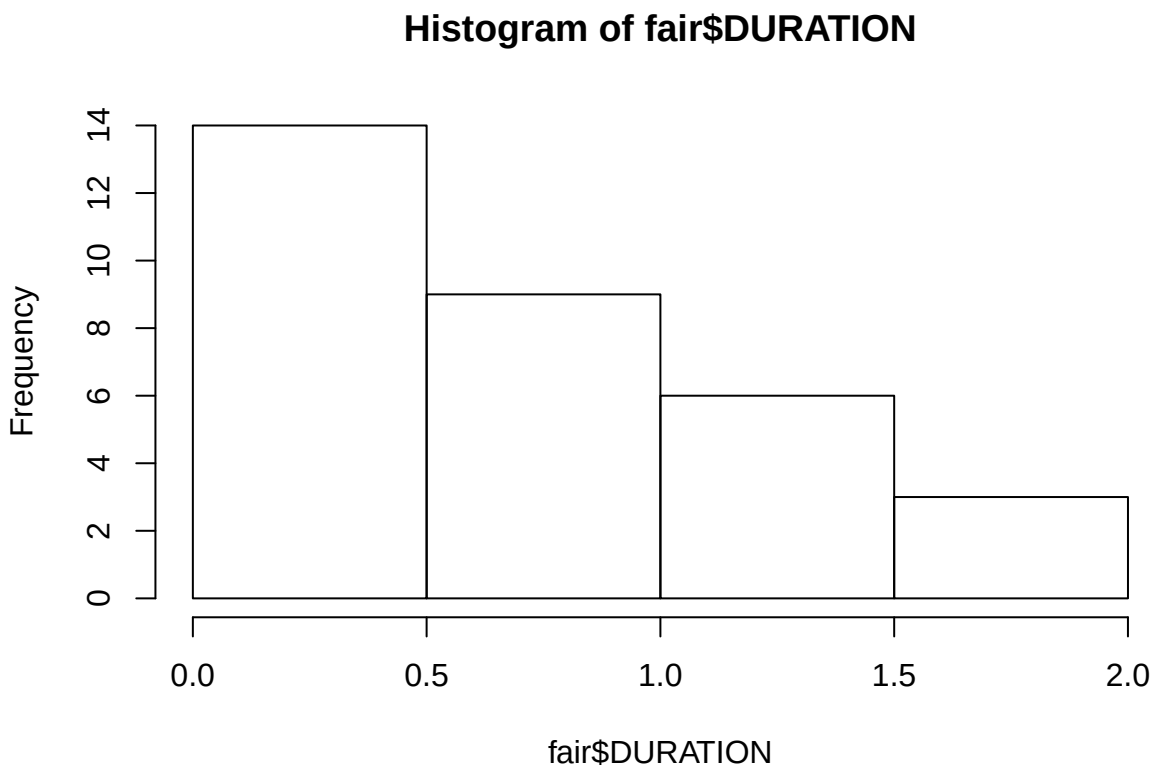
A vari?vel duration apresenta os valores de 2 a 0. Com o comando `fivenum`? poss?vel obter uma an?lise descritiva das vari?vel:

```
fivenum(fair$DURATION)
```

```
## [1] 0.00 0.00 1.00 1.25 2.00
```

O valor m?nimo e do primeiro quartil ? 0, a mediana ? 1, o 3? quartil ? 1.25 e o valor m?ximo da distribui??o ? 2.

```
hist(fair$DURATION)
```



Pelo histograma, vemos que a maior parte das observ??es s?o 0 e 1 (25)

War

A vari?vel war ? uma vari?vel bin?ria que ? melhor observada por meio do comando `ftable`:

```
ftable(fair$WAR)
```

```
##   0  1
##
##  29  3
```

? poss?vel verificar que 29 casos possuem valor 0 e 3 casos apresentam valor 1

Growth

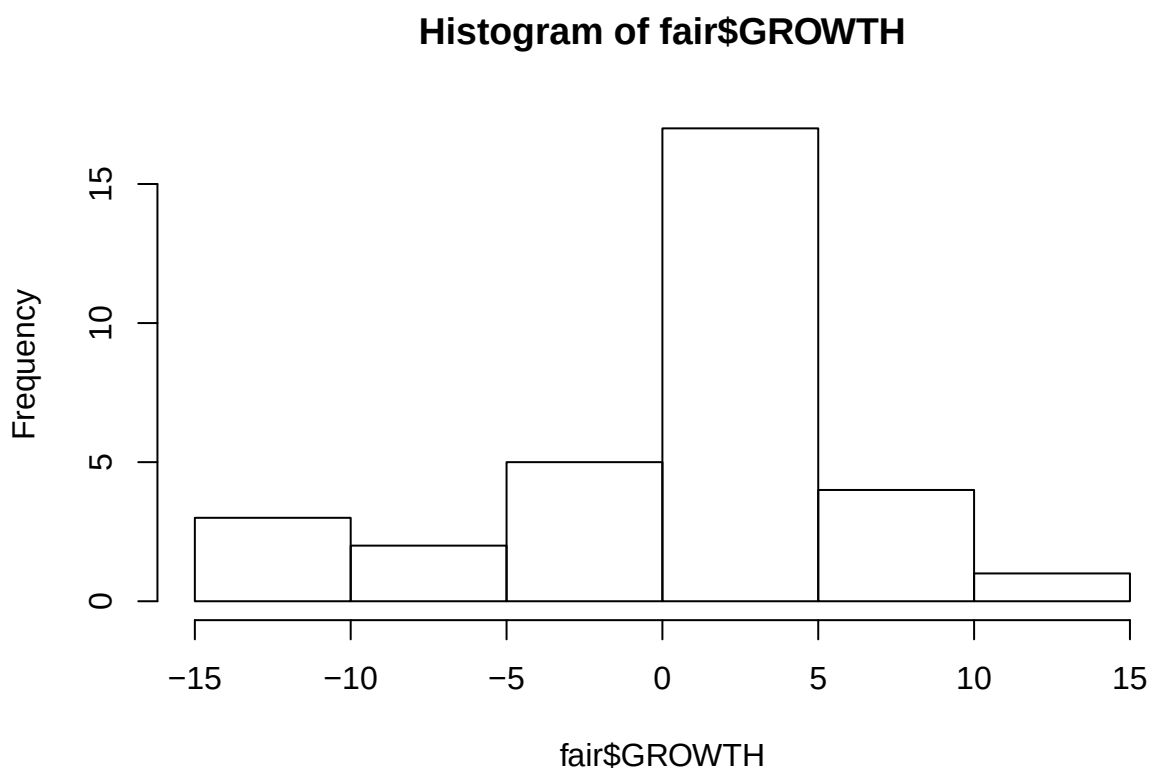
A vari?vel growth cont?m informa??es acerca da varia??o de crescimento do PIB por ano:

```
fivenum(fair$GROWTH)
```

```
## [1] -14.557  -1.923   2.245   4.095  11.677
```

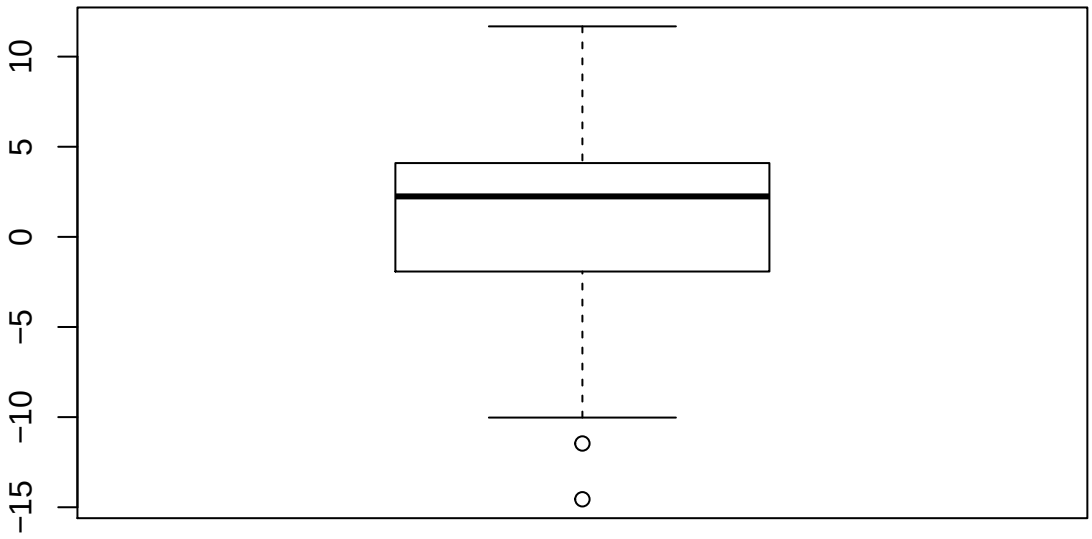
A descri??o por meio do comando fivenum mostra que o valor minimo da distribui??o foi uma contra??o de 14.56 no PIB, enquanto que o 1? quartil apresenta um valor de -1.92, a mediana ? de 2.24, o 3? quartil ? 4.1 e o valor m?ximo da distrbui??o ? de 11.68. Por meio de um histograma ? poss?vel uma melhor visualiza??o da distribui??o.

```
hist(fair$GROWTH)
```



Com o histograma, percebemos que boa parte da distribui??o est? localizada no centro, com casos entre -5 e 5. Com um box plot, podemos identificar a presen?a ou n?o de outliers na distribui??o.

```
boxplot(fair$GROWTH)
```



Por meio do Boxplot, podemos verificar que existem dois valores que s?o outliers: -14.55 e -11.46.

Inflation

Vamos verificar as estat?sticas principais da vari?vel inflation:

```
fivenum(fair$INFLATION)
```

```
## [1] 0.0000 1.3545 2.1590 3.3715 7.9260
```

O valor m?nimo da distribui??o ? 0, o 1? quartil ? 1.35, a mediana ? de 2.16, o 3? quartil ? de 3.38 e o valor m?ximo ? de 7.93.

Goodnews

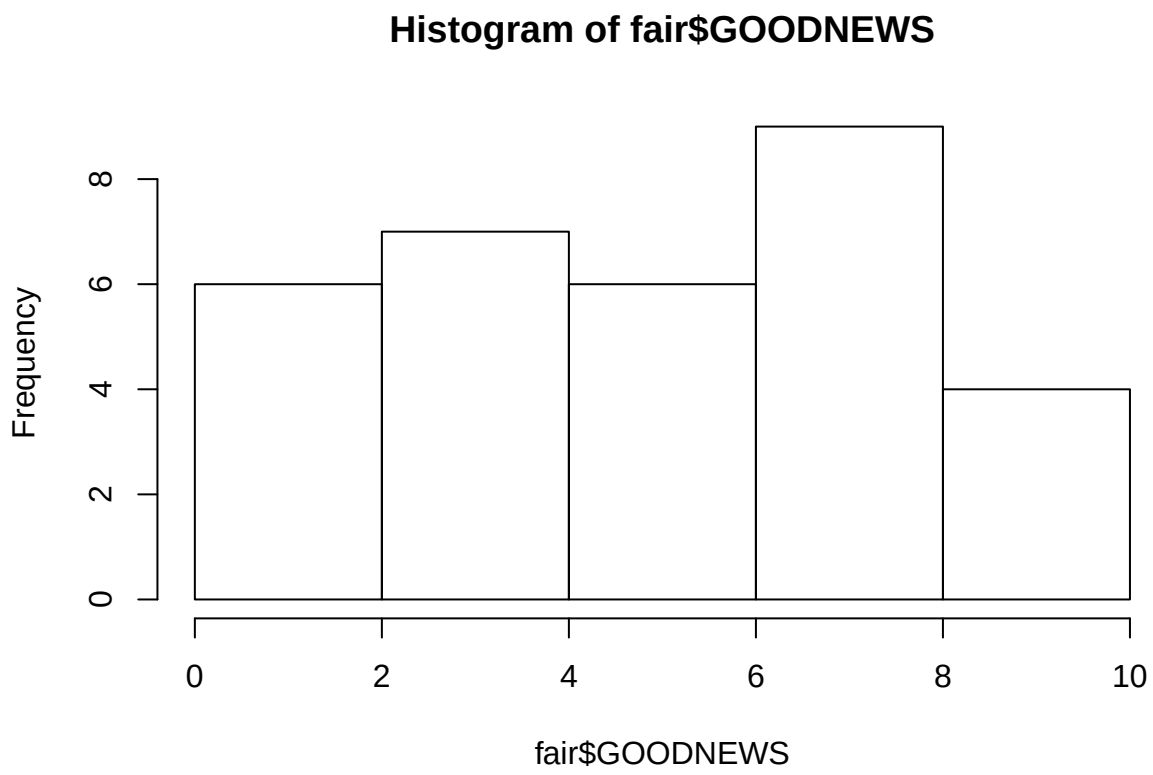
Abaixo vemos os cinco valores da vari?vel goodnews

```
fivenum(fair$GOODNEWS)
```

```
## [1]  0.0  3.5  5.0  7.5 10.0
```

O valor m?nimo da vari?vel ? 0, o 1? quartil ? 3.5, a mediana ? 5, o 3? quartil ? 7.5 e o valor m?ximo ? 10.

```
hist(fair$GOODNEWS)
```



O histograma mostra que a distribui??o dos valores ? bem balanceada.

b) Modelo de regress?o

Para executar o modelo, com a vari?vel Vote como VD e Growth como VI, vamos utilizar o seguinte comando:

```
Linear <- lm(VOTE ~ GROWTH, data = fair)
summary(Linear)
```

```
##
## Call:
## lm(formula = VOTE ~ GROWTH, data = fair)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2487 -3.3330 -0.4282  3.1425  9.7286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.8598     0.8817  58.821  < 2e-16 ***
## GROWTH        0.6536     0.1607   4.068 0.000316 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.955 on 30 degrees of freedom
## Multiple R-squared:  0.3555, Adjusted R-squared:  0.3341
## F-statistic: 16.55 on 1 and 30 DF,  p-value: 0.0003165
```

Resultado da regress?o

Como ? poss?vel observar no resultado, a varia??o de uma unidade na VI (Growth) leva a um aumento de 0.65 na VD (Votes). Nesse caso, o modelo de regress?o utilizado ? bivariado, dado a presen?a de apenas duas vari?veis no modelo.

Em rela??o aos res?duos do modelo, o valor m?nimo ? de -8.25, o 1? quartil ? de -3.33, a mediana ? de -0.423, o 3? quartil ? de 3.14 e o valor m?ximo ? de 9.72.

Quando a VI assume o valor de 0, espera-se que o valor da VD seja de 51.86 (Valor do intercepto). J? em rela??o a VI, al?m do valor do coeficiente, o erro padr?o foi de 0.16, o teste-f que testa a hip?tese nula de que n?o h? rela??o entre as vari?veis apresentou um p-valor menor que 0.001. Ou seja, rejeita-se a hip?tese nula de que n?o h? rela??o entre as vari?veis.

Capacidade explicativa do modelo

No que se refere a capacidade explicativa do modelo, o R? ajustado foi de 0.33, significando que o modelo consegue explicar 33% da vari?ncia da vari?vel dependente. O erro padr?o do res?duo foi de 4.955.

Ajuste do modelo

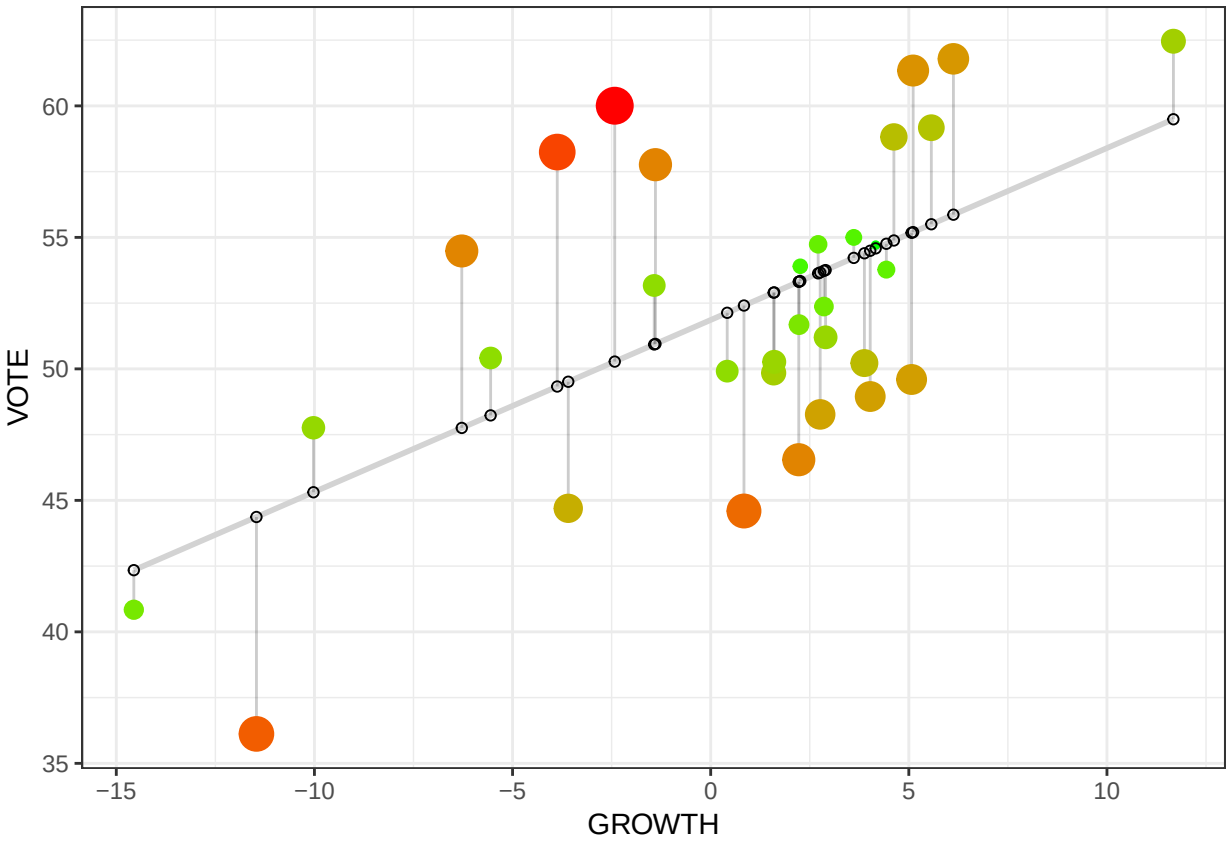
J? no tocante ao ajuste, podemos analisar alguns gr?ficos e verificar a adequabilidade do modelo:

```
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.5.3
```

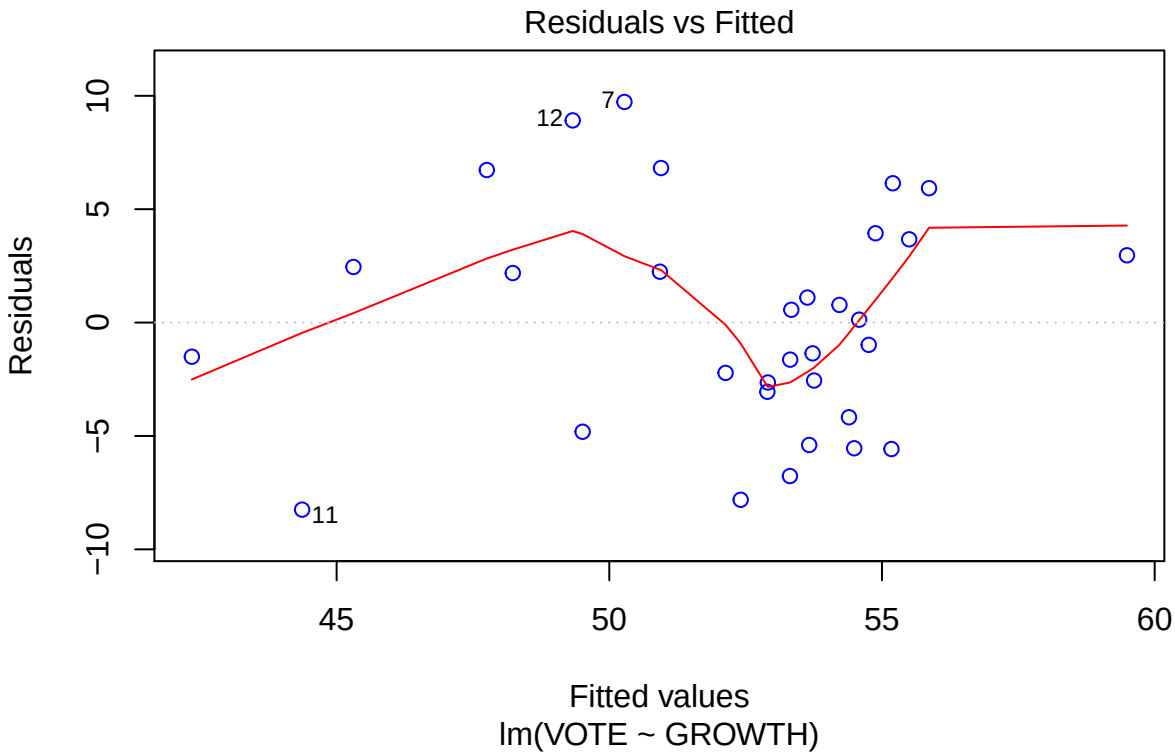
O gr?fico abaixo apresenta o tamanho do res?duo por meio to tamanho do ponto e cor (quanto mais vermelho mais longe e quanto mais verde menor o res?duo). O tamanho do res?duo ? a distancia entre o ponto e a linha de regress?o.

```
fair$predicted <- predict(Linear)
fair$residuals <- residuals(Linear)
ggplot(fair, aes(x = GROWTH, y = VOTE)) +
  geom_smooth(method = "lm", se = FALSE, color = "lightgrey") +
  geom_segment(aes(xend = GROWTH, yend = predicted), alpha = .2) +
  geom_point(aes(color = abs(residuals), size = abs(residuals))) +
  scale_color_continuous(low = "green", high = "red") +
  guides(color = FALSE, size = FALSE) +
  geom_point(aes(y = predicted), shape = 1) +
  theme_bw()
```



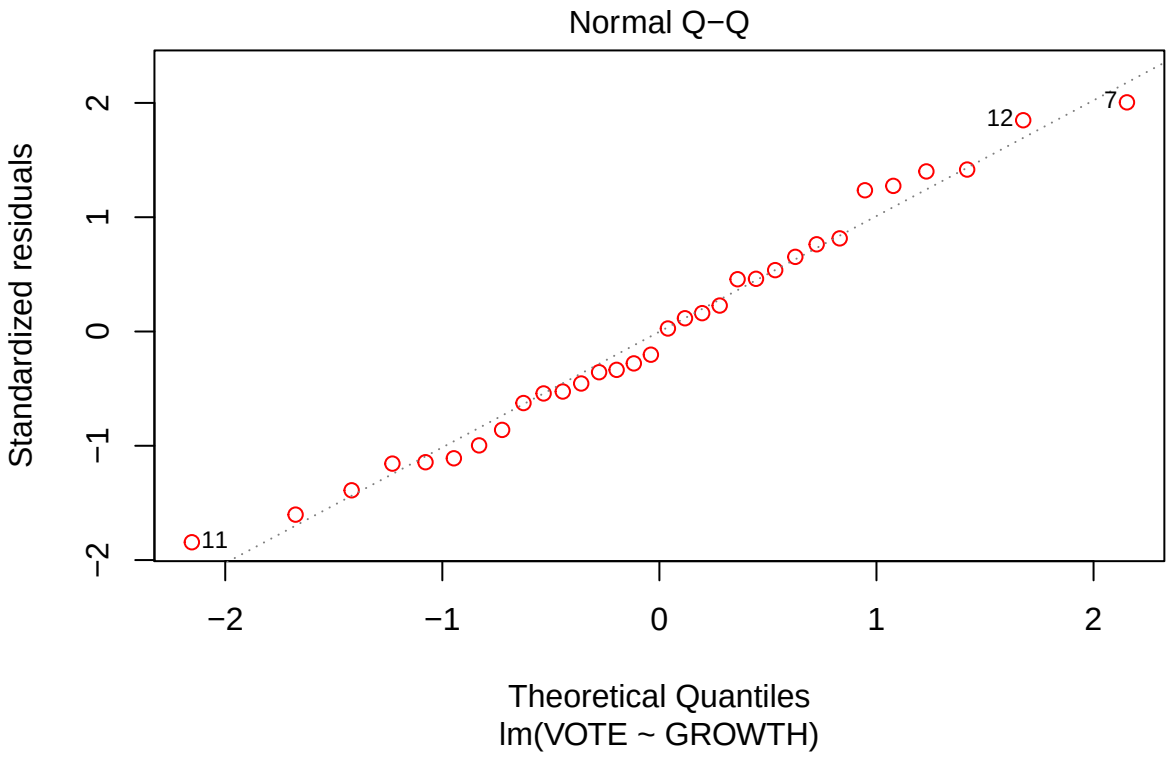
Podemos por meio do gr?fico observar a distribui??o dos res?duos do modelo. Por meio do gr?fico de res?duos, ? poss?vel verificar a adequabilidade dos dados. Dado que o res?duo ? aquela parte do modelo que n?o ? explicada, espera-se que este n?o possua nem um padr?o. Em suma, a distribui??o do res?duo deve ser aleat?ria.

```
plot(Linear, which=1, col=c("blue"))
```



Quando observamos o gr?fico, ? poss?vel verificar uma certa concentra??o de observa??es na parte um pouco a direita do centro do gr?fico, o que pode apresentar uma n?o normalidade dos dados. Por isso ser? executado um gr?fico de quantis (quantile-quantile plot). Caso os res?duos sigam uma linha no gr?fico, ? uma boa indica??o de uma distribui??o normal.

```
plot(Linear, which=2, col=c("red"))
```



Observa-se que os res?duos seguem a linha da regress?o de maneira bastante alinhada.

Com base nessas an?lises, ? poss?vel concluir que o modelo ? adequado.

c) Modelo de regress?o com mais de uma VI

Para executar o modelo de regress?o com VOTES como VD com Growth e Goodnews como VI, executamos o seguinte comando:

```
Linear_2 <- lm(VOTE ~ GROWTH + GOODNEWS, data = fair)
summary(Linear)
```

```
##
## Call:
## lm(formula = VOTE ~ GROWTH, data = fair)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.2487 -3.3330 -0.4282  3.1425  9.7286
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  51.8598     0.8817  58.821  < 2e-16 ***
## GROWTH        0.6536     0.1607   4.068 0.000316 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.955 on 30 degrees of freedom
## Multiple R-squared:  0.3555, Adjusted R-squared:  0.3341
## F-statistic: 16.55 on 1 and 30 DF,  p-value: 0.0003165
```

Resultados do modelo

A vari?vel adicionada ao modelo ? a de GoodNews, relacionada a quantidade de not?cias boas na ?rea economica.

Ao an?lisar o modelo, percebemos que ambas as vari?veis independentes apresentam um efeito sobre a vari?vel dependente. Entretanto, a vair?vel Growth apresenta um P-valor menor que 0.001 enquanto que a v?riavel Goodnews s? apresenta um efeito significativo sobre a VD quando consideramos um P-valor menor que 0.05. Mesmo assim, ambas as vari?veis est?o abaixo do patamar de 0.05, permitindo rejeitar a hip?tese nula de que n?o h? rela??o entre as vari?veis. Observando o coeficiente da vari?vel Gwroth, podemos interpretar que, o aumento de 1% na varia??o do PIB leva a um aumento de 0.57 na VD (mantendo todo o resto constante). J? em rela??o a GOODNEWS, o aumento em 25% de not?cias economicas positivas, leva ao aumento de 0.71 na VD (mantendo todo o resto constante).

Os valores do res?duo mostram um res?duo m?nimo de -8.3125, com o 1? quartil igual a -3.9191, Mediana de 0.4876, 3? quartil de 3.05 e valor m?ximo de 9.68.

Capacidade explicativa do modelo

No que se refere a capacidade explicativa, o modelo apresenta um R? ajustado de 0.43, significando uma explica??o de 43% de varia??o na VD. O erro padr?o do res?duo foi de 4.596.

RMSE

Para analisar o RMSE (Raiz quadrada do erro m?dio) ser? necess?rio primeiro abrir o pacote `sjstats`

```
library("sjstats")
```

```
## Warning: package 'sjstats' was built under R version 3.5.3
```

Assim ? poss?vel verificar o valor da RMSE:

```
rmse(Linear)
```

```
## [1] 4.797286
```

O RMSE ? calculado para mensurar a diferen?a entre os valores preditos pelo modelo e os valores observados. Tamb?m ? conhecido como erro padr?o do modelo e ? calculado elevando ao quadrado cada erro do modelo, somando-os, dividindo pelo n?mero de casos e obtendo a raiz quadrada. Nesse caso, o valor obtido foi de 4.37.O RMSE tem a mesma unidade que a VD e quanto menor, melhor.

Comparando resultados dos modelos

Para poder comparar os resultados desse modelo com o anterior, ? necess?rio padronizar os coeficientes de ambos os modelos. Para isso ser? utilizado o pacote `lm.beta`:

```
library('lm.beta')
```

```
## Warning: package 'lm.beta' was built under R version 3.5.2
```

```
lm.beta(Linear)
```

```
##
## Call:
## lm(formula = VOTE ~ GROWTH, data = fair)
##
## Standardized Coefficients::
## (Intercept)      GROWTH
##   0.0000000   0.5962706
```

```
lm.beta(Linear_2)
```

```
##
## Call:
## lm(formula = VOTE ~ GROWTH + GOODNEWS, data = fair)
##
## Standardized Coefficients::
## (Intercept)      GROWTH    GOODNEWS
##   0.0000000   0.5227285   0.3373268
```

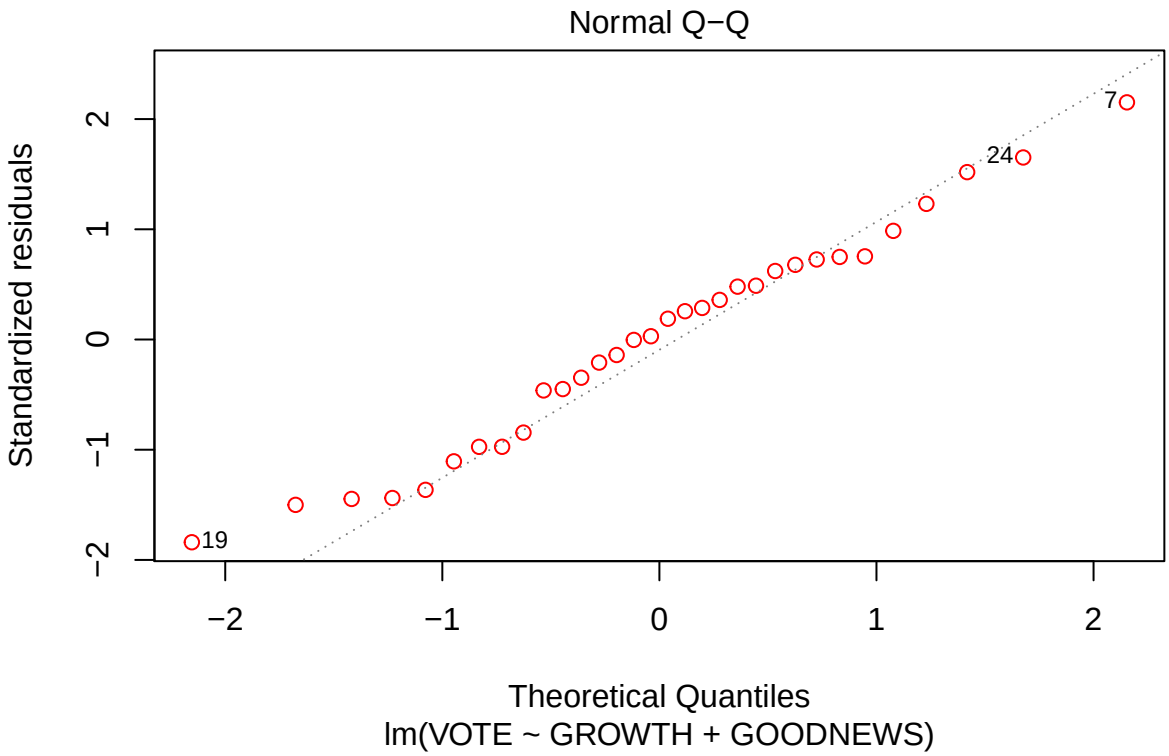
Por meio dos coeficientes padronizados, ? poss?vel comparar os modelos. No primeiro modelo executado, verifica-se que para o aumento de 1 desvio padr?o da VI Growth, espera-se o aumento de .60 desvio padr?o na VD.

J? no segundo modelo, o aumento de 1 desvio padr?o da VI Growth leva a um aumento de 0.52 desvio padr?o na VD. Ou seja, houve uma redu??o no efeito da VI Growth do modelo 1 para o modelo 2.

Ajuste do modelo

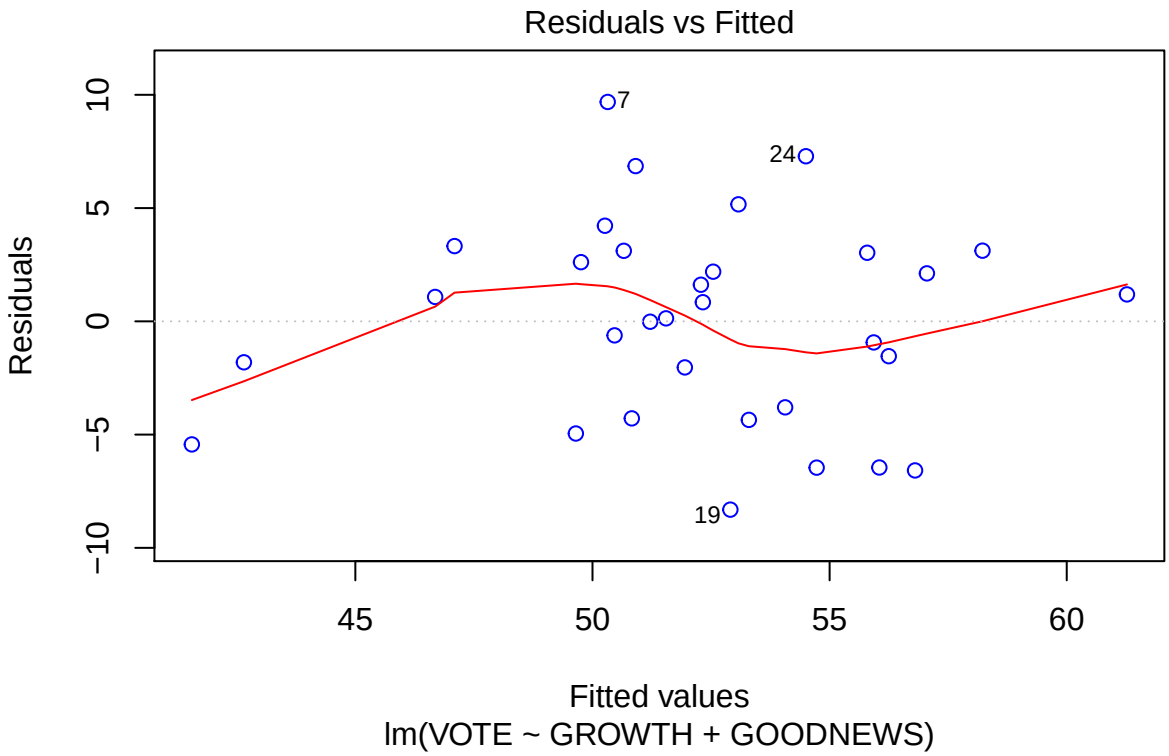
Para verificar o ajuste do modelo, ser? analisado os residuos graficamente.

```
plot(Linear_2, which=2, col=c("red"))
```



O gr?fico Q-Q mostra um ajuste adequado do modelo dado a tend?ncia dos pontos em rela??o a linha de regress?o.

```
plot(Linear_2, which=1, col=c("blue"))
```



J? o gr?fico dos res?duos apresenta valores tanto acima quanto abaixo da linha, representando tamb?m um bom ajuste do modelo.

Por fim, o teste de Shapiro permite verificar a normalidade da distribui??o:

```
res <- residuals(Linear_2)

shapiro.test(res)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  res
## W = 0.97866, p-value = 0.7598
```

Os resultados do teste apontam um p-valor de 0.76, levando a n?o rejeitar a hip?tese nula de que os dados testados n?o est?o normalmente distribu?dos.