Regressão Linear Múltipla no **R**

MAE-0217: ESTATÍSTICA DESCRITIVA – MAIO DE 2017

PROFESSORA: MÁRCIA D'ELIA BRANCO

MONITORA PAE: SIMONE HARNIK

Regressão Linear Simples

- O modelo de regressão linear simples é utilizado para descrever a relação entre duas variáveis: uma explicativa e outra resposta
- Cada observação pode ser expressa como:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad (i = 1,\ldots,n)$$

Em que:

y_i: valor observado da variável Y para a observação i

x_i: valor observado da variável X para a observação i

β_0: parâmetro referente ao intercepto da reta ajustada

β_1: parâmetro referente ao coeficiente angular da reta ajustada

ϵ_i: erro aleatório associado à observação i

Regressão Linear Múltipla

- Quando há mais de uma variável explicativa, sejam elas contínuas ou discreta, podemos modelar a resposta média por uma função dessas variáveis:

$$y_i = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi} + \epsilon_i \qquad (i = 1,\ldots,n)$$

Em que:

y_i: valor observado da variável Y para a observação i

x_{ji}: valor observado da variável X_j para a observação i, com j = 0, ..., p

β_j: parâmetro a ser estimado no modelo

ϵ_i: erro aleatório associado à observação i

Regressão Linear Múltipla

- **Suposições**
  - A relação entre a variável dependente e as independentes é linear
  - Os erros são independentes
  - Os erros são identicamente distribuídos segundo uma N(0,1)
  - A variância dos erros permanece constante (homocedasticidade)

Como checar?!?!?!?!

Análise dos resíduos!

Regressão no **R**

- A função lm pode ser usada para ajustar modelos de regressão linear
- Há vários gráficos já implementados para verificar as suposições do modelo
- A função summary também tem saída específica para a função lm

Exemplo 1 – Preço de casas

Utilizaremos o conjunto de dados chamado “Housing”

Vamos explicar o preço das casas (price) por meio de:

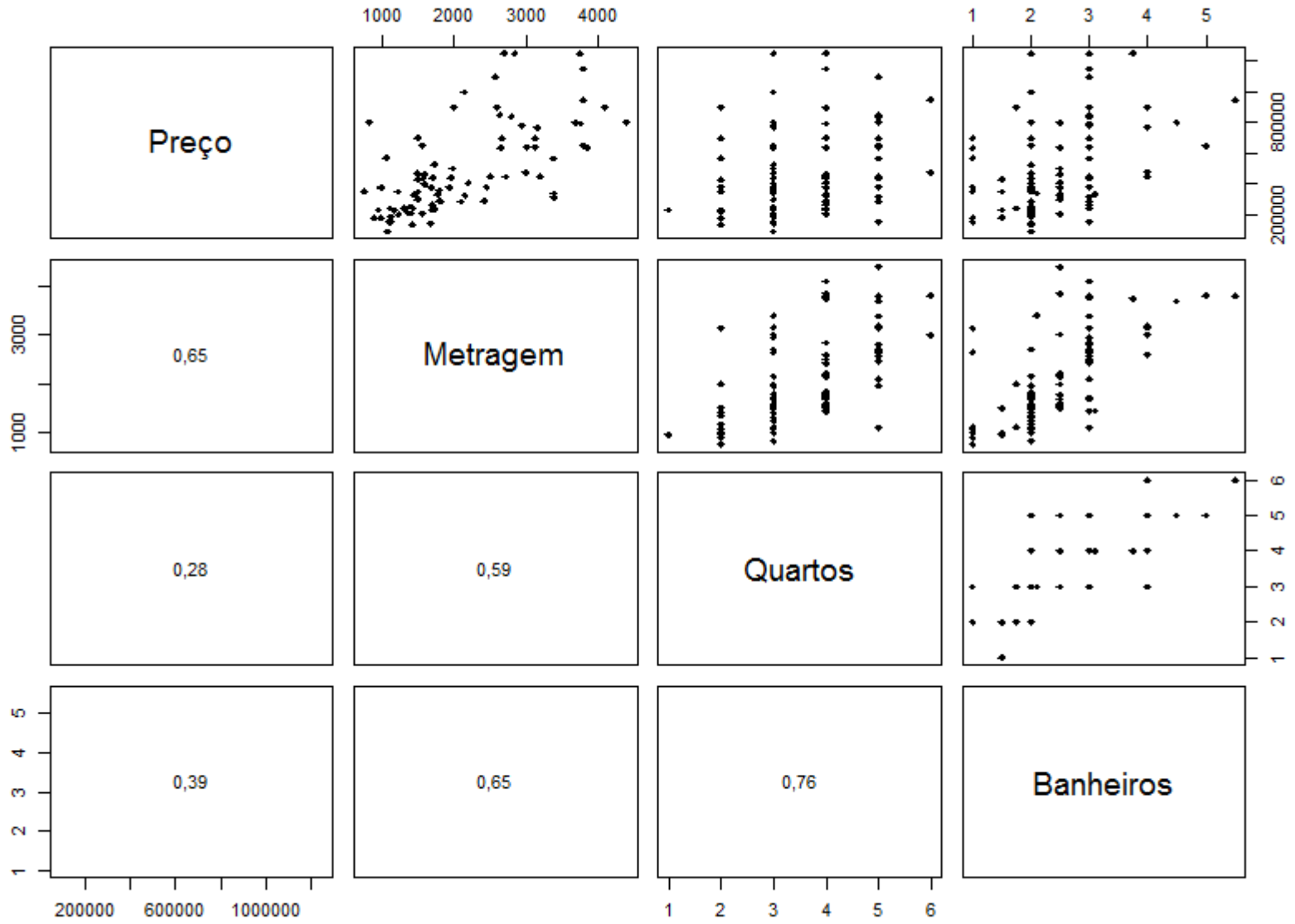
- Tamanho em pés^2 (sqft)
- Cidade (city)
- Número de quartos (bedrooms)
- Número de banheiros (baths)

Exemplo 1 – Leitura dos dados

```
options(OutDec = ",") #Saída com decimal em ","

#Leitura dos dados
house <- read.table("http://www.rossmanchance.com/iscam2/data/housing.txt",
                    header = T, sep = "\t")
attach(house)
names(house)
```

1º passo – Estatísticas descritivas

É importante verificar a relação entre as variáveis antes da modelagem



Ajuste de modelo de regressão

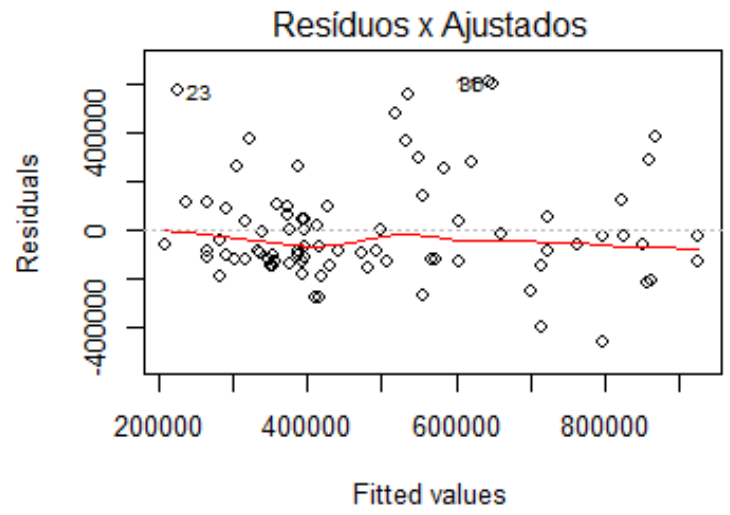
```
fit1<-lm(price~sqft+bedrooms+baths)
summary(fit1)
par(mfrow=c(2,2))
plot(fit1, caption=c("Resíduos x Ajustados", "QQ-Plot Normal",
                     "Locação e Escala", "", "Resíduos e alavancagem"))
```

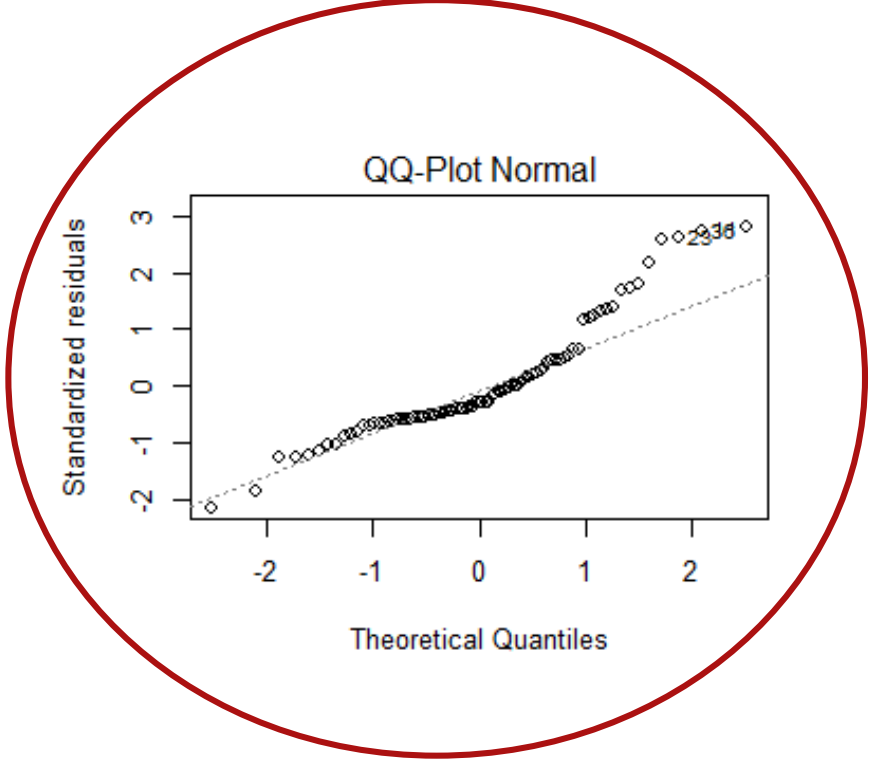
Saída do R

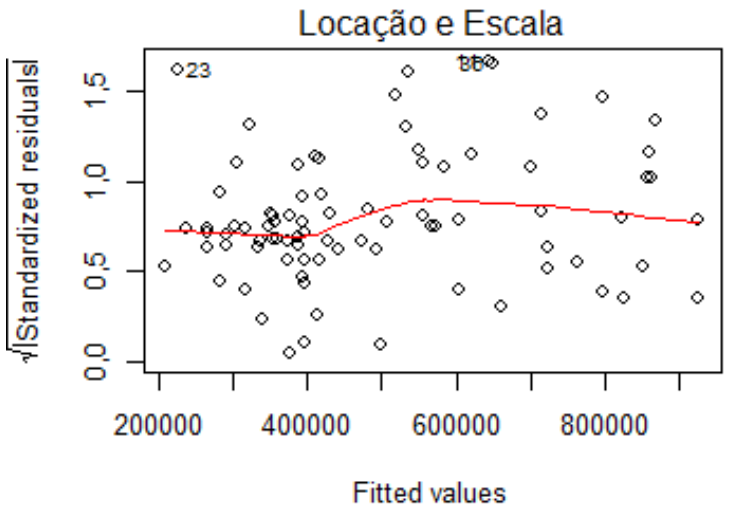
Lembre-se: esta é só a saída do *R*. Em uma consultoria, é necessário transformá-la em uma tabela compreensível ao leitor leigo

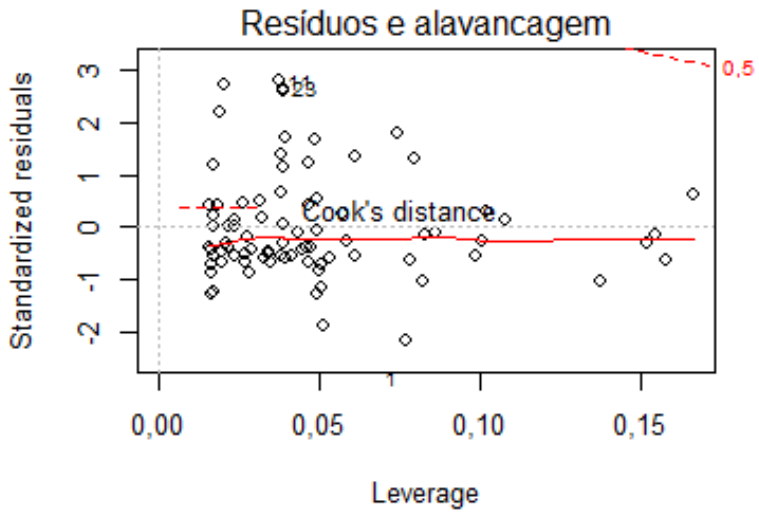
```
> summary(fit1)

Call:
lm(formula = price ~ sqft + bedrooms + baths)

Residuals:
    Min      1Q  Median      3Q     Max
-458105 -127923  -67537   92551  607742

Coefficients:
             Estimate Std. Error t value     Pr(>|t|)
(Intercept) 146817,28   85751,71   1,712       0,0908 .
sqft           221,38      35,19   6,291 0,0000000164 ***
bedrooms    -52862,71   35755,95  -1,478       0,1433
baths        27600,71   46382,98   0,595       0,5535
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 221700 on 79 degrees of freedom
Multiple R-squared:  0,4374,    Adjusted R-squared:  0,416
F-statistic: 20,47 on 3 and 79 DF,  p-value: 0,0000000006526
```

Gráficos de resíduos









Ajuste com resposta transformada (log)

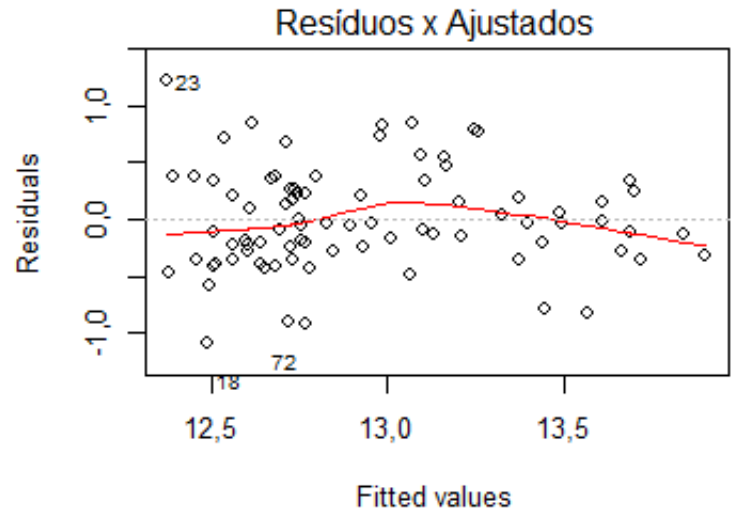
```
fit2<-lm(log(price)~sqft+bedrooms+baths)
summary(fit2)
par(mfrow=c(2,2))
plot(fit1, caption=c("Resíduos x Ajustados", "QQ-Plot Normal",
                     "Locação e Escala", "", "Resíduos e alavancagem"))
```

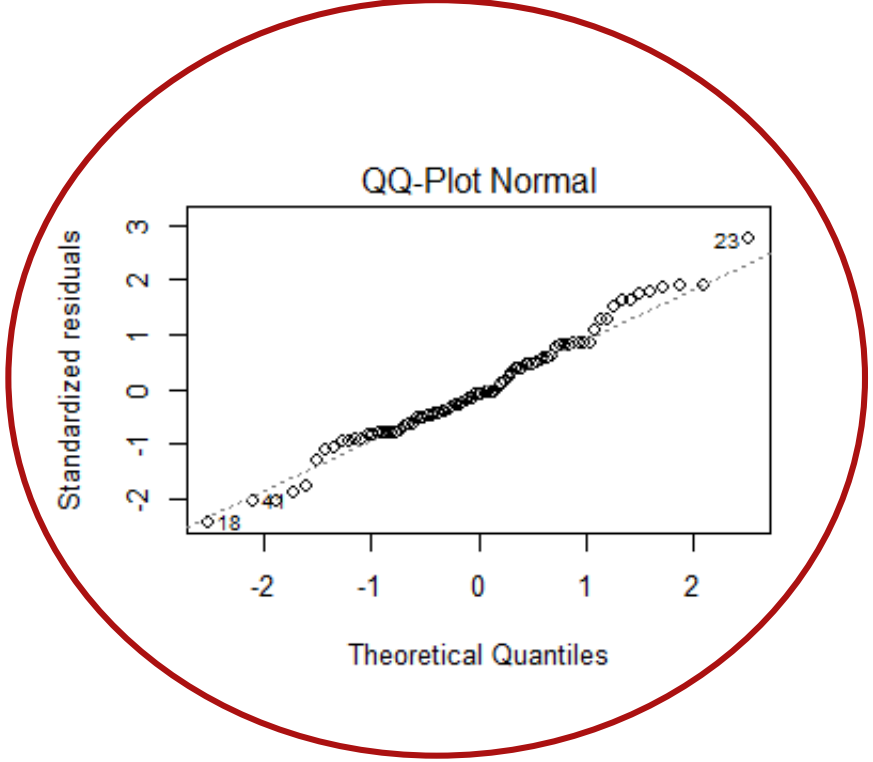
Saída do R

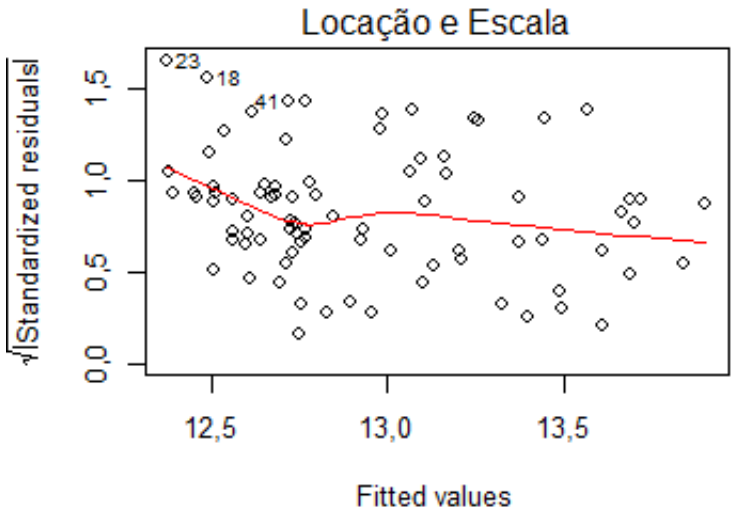
Lembre-se: esta é só a saída do *R*. Em uma consultoria, é necessário transformá-la em uma tabela compreensível ao leitor leigo

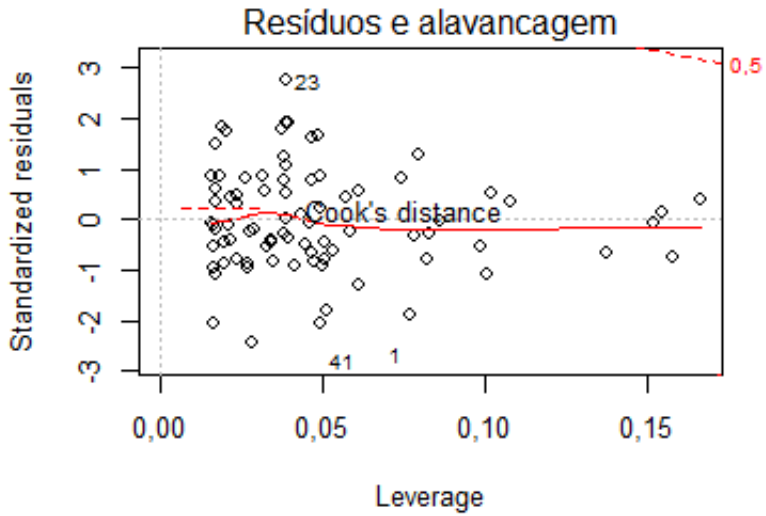
```
> summary(fit2)

Call:
lm(formula = log(price) ~ sqft + bedrooms + baths)

Residuals:
     Min       1Q   Median       3Q      Max
-1,08355 -0,28314 -0,04766  0,26424  1,21742

Coefficients:
             Estimate  Std. Error t value      Pr(>|t|)
(Intercept) 12,16259203  0,17511975  69,453       < 2e-16 ***
sqft         0,00046338  0,00007187   6,448 0,00000000832 ***
bedrooms    -0,06826716  0,07301979  -0,935         0,353
baths        0,01680607  0,09472202   0,177         0,860
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 0,4527 on 79 degrees of freedom
Multiple R-squared:  0,4486,    Adjusted R-squared:  0,4277
F-statistic: 21,43 on 3 and 79 DF,  p-value: 0,0000000002982
```

Gráficos de resíduos









Seleção de variáveis

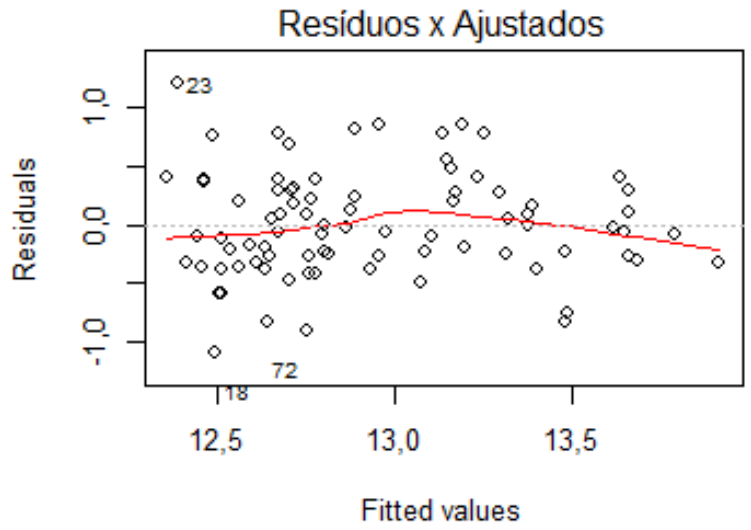
RETIRAMOS BANHEIROS

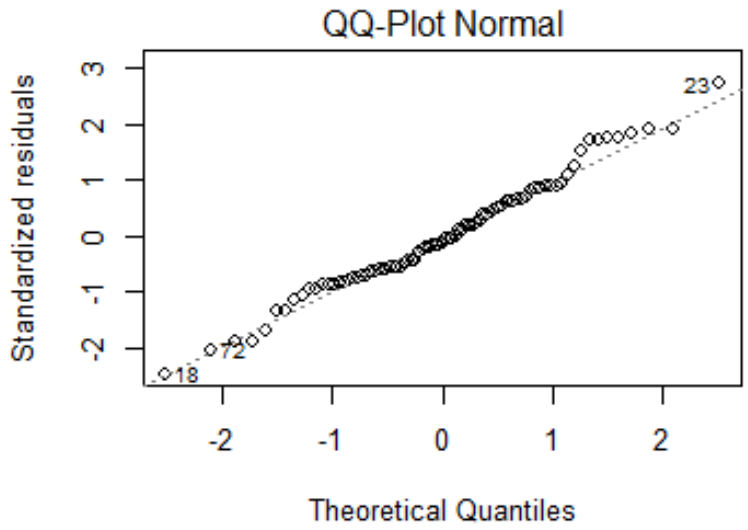
```
> fit3<-lm(log(price)~sqft+bedrooms)
> summary(fit3)

Call:
lm(formula = log(price) ~ sqft + bedrooms)

Residuals:
    Min      1Q  Median      3Q     Max
-1,0814 -0,2776 -0,0530  0,2680  1,2208

Coefficients:
              Estimate  Std. Error t value       Pr(>|t|)
(Intercept) 12,16494285  0,17355757  70,092        < 2e-16 ***
sqft         0,00046825  0,00006603   7,092 0,000000000473 ***
bedrooms    -0,06029335  0,05720160  -1,054          0,295
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 0,4499 on 80 degrees of freedom
Multiple R-squared:  0,4484,    Adjusted R-squared:  0,4346
F-statistic: 32,52 on 2 and 80 DF,  p-value: 0,00000000004623
```

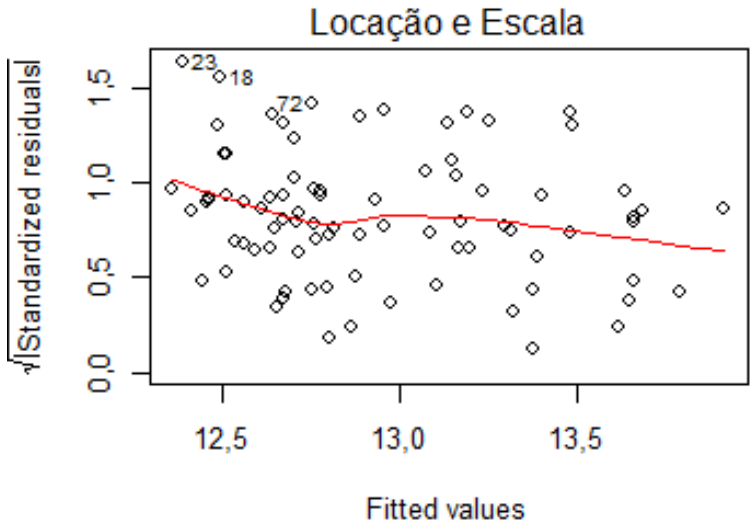
RETIRAMOS QUARTOS

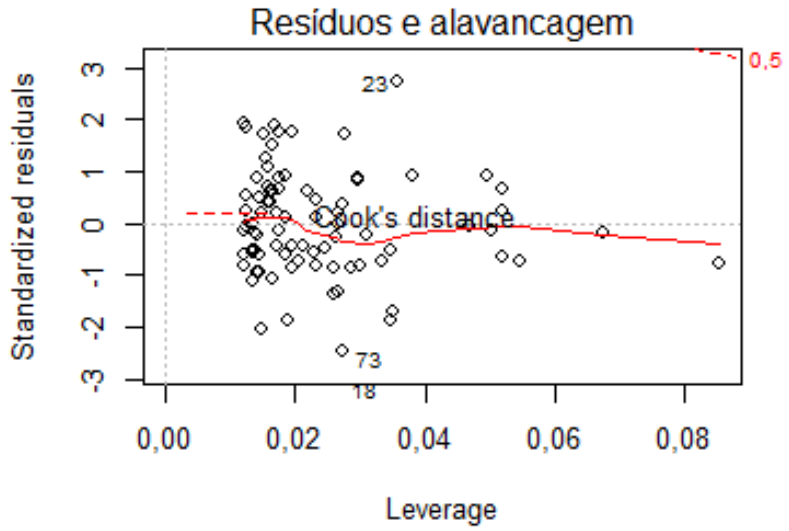
```
> fit4<-lm(log(price)~sqft)
> summary(fit4)

Call:
lm(formula = log(price) ~ sqft)

Residuals:
     Min       1Q   Median       3Q      Max
-1,08988 -0,29591 -0,05899  0,28717  1,20206

Coefficients:
              Estimate  Std. Error t value         Pr(>|t|)
(Intercept) 12,03645213  0,12362471   97,36          < 2e-16 ***
sqft         0,00042740  0,00005349    7,99 0,00000000000787 ***
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 0,4502 on 81 degrees of freedom
Multiple R-squared:  0,4407,    Adjusted R-squared:  0,4338
F-statistic: 63,83 on 1 and 81 DF,  p-value: 0,000000000007874
```

Gráficos de resíduos



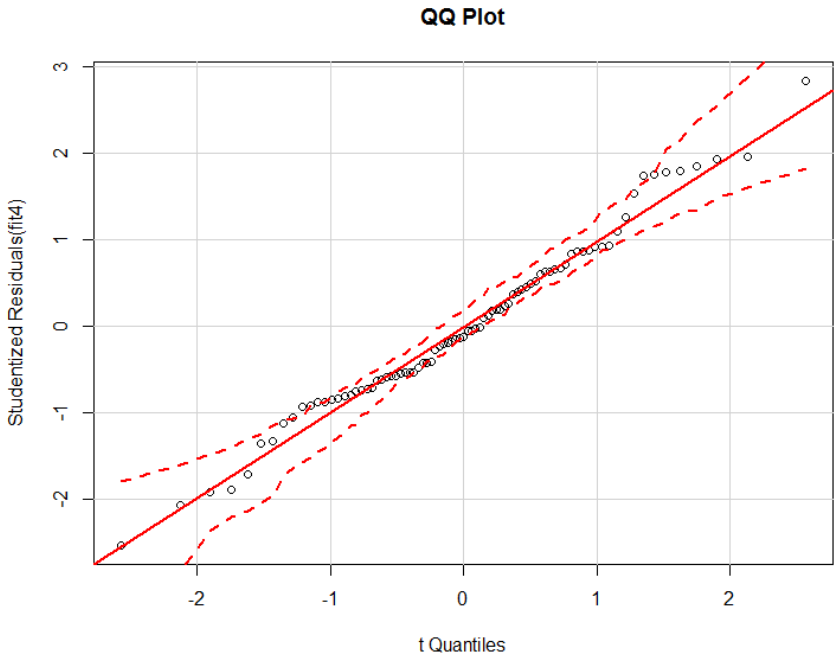




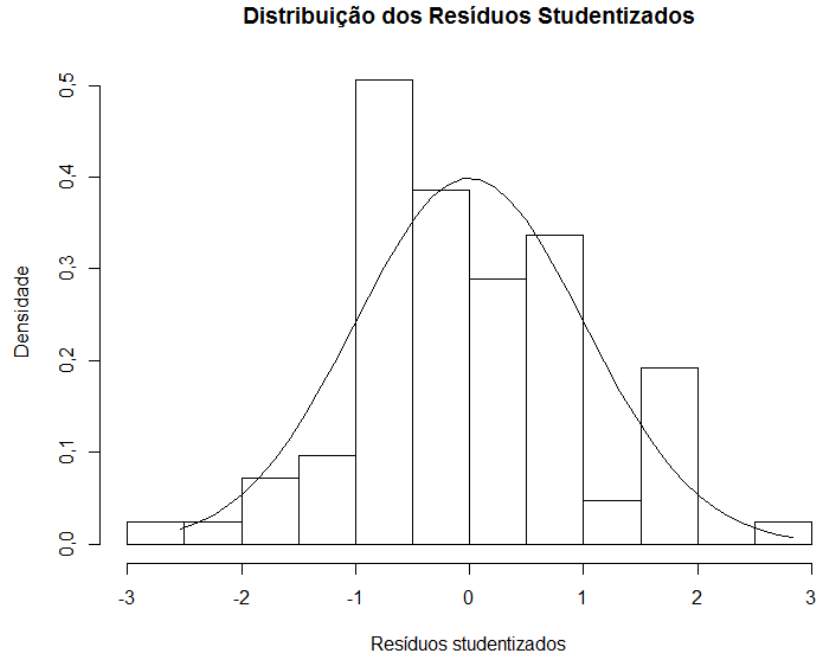


Detalhamento: normalidade

```r
library(MASS)
library(car)
qqPlot(fit4, main="QQ Plot")
```




```r
sresid <- studres(fit4)
hist(sresid, freq=FALSE, ylab='Densidade', xlab = "Resíduos studentizados",
     main="Distribuição dos Resíduos Studentizados")
xfit<-seq(min(sresid),max(sresid),length=40)
yfit<-dnorm(xfit)
lines(xfit, yfit)
```



Exemplo 2 – Hanseníase

- Dados reais de um estudo com pacientes com hanseníase acompanhados pela dermatologia do Hospital das Clínicas
- Objetivo: verificar se há relação entre a temperatura de sensação de frio com o fato de sentir dor neuropática ou não e com a quantidade de fibras nervosas (verificada em microscópio após biópsia)
- Cada unidade amostral é uma mão, braço ou perna do paciente (pois essa doença acomete os lados do corpo independentemente)

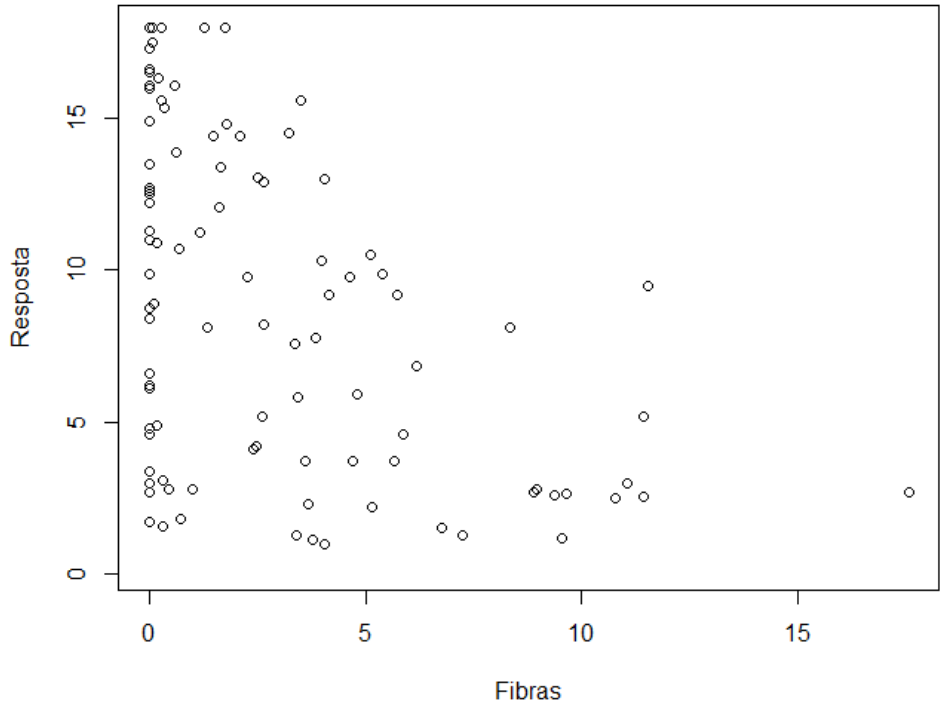
Exemplo 2 – Hanseníase

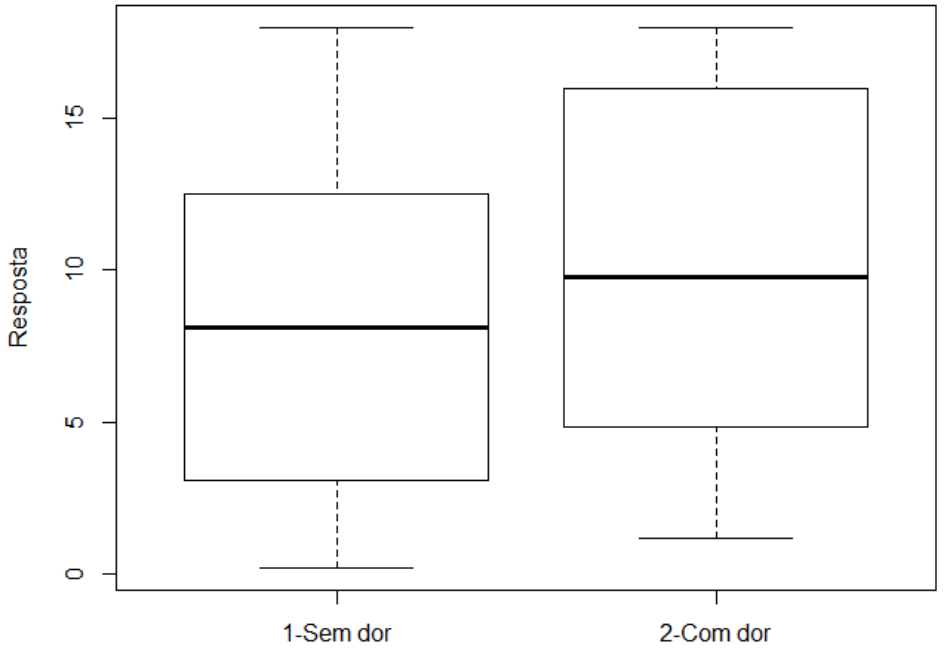
- A resposta é a variação de temperatura até o paciente declarar a sensação de calor
- Todos os pacientes foram orientados a declarar a sensação no momento em que sentissem
- É usado um equipamento que vai aumentando a temperatura
- A temperatura inicial é de 32°C
- Assim, um paciente com registro de 10°C na resposta só sentiu calor a 42°C
- Detalhe: por questões médicas e éticas, é necessário parar o aquecimento a 50°C

Medidas-resumo

Variável	Mínimo	Q1	Mediana	Q3	Máximo	Média	DP	N
Resposta	0,20	3,44	8,58	13,31	18,00	8,83	5,52	118
Fibras	0,00	0,00	1,42	4,08	17,55	2,82	3,59	100

Análise descritiva





Ajuste de modelo de regressão

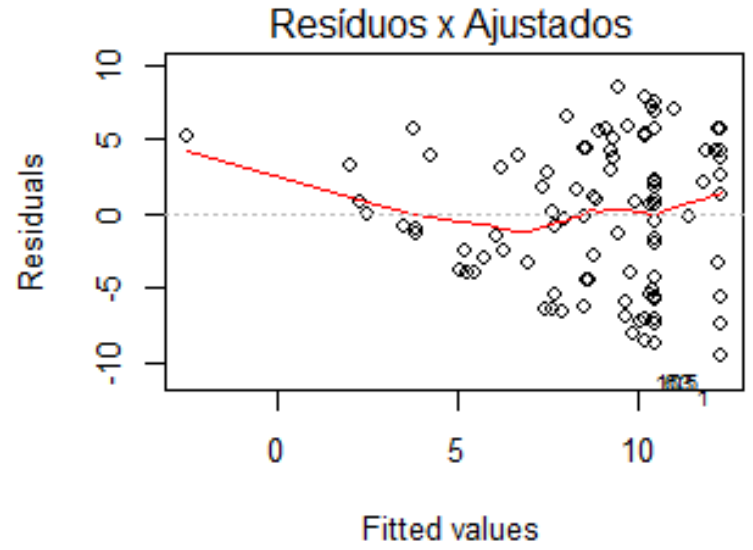
```
fit1<-lm(RESP~FIBRAS_PGP+GRUPO2)
summary(fit1)
par(mfrow=c(2,2))
plot(fit1, caption=c("Resíduos x Ajustados", "QQ-Plot Normal",
                     "Locação e Escala", "", "Resíduos e alavancagem"))
```

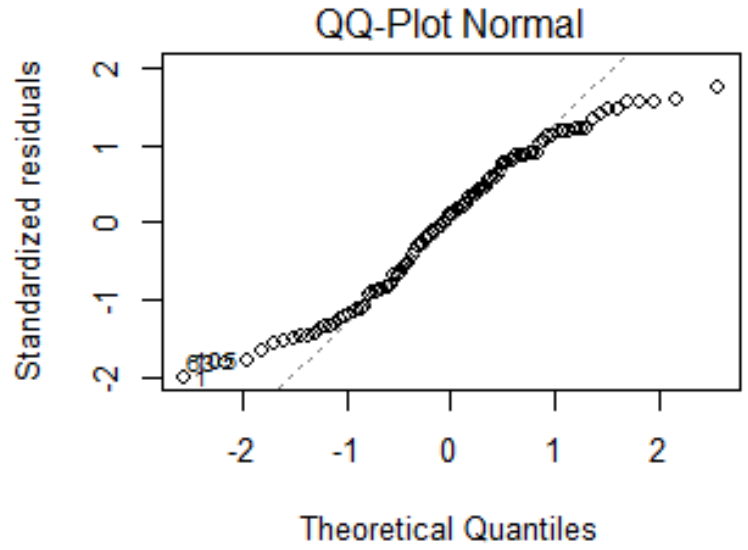
Saída do R

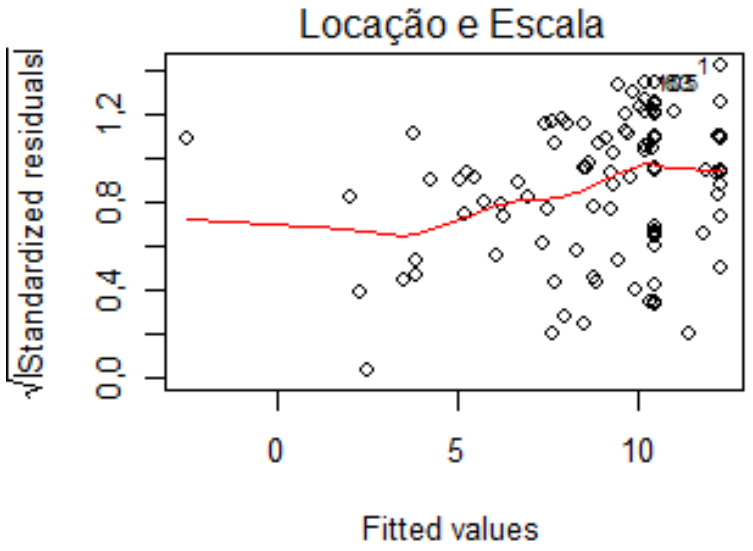
Lembre-se: esta é só a saída do *R*. Em uma consultoria, é necessário transformá-la em uma tabela compreensível ao leitor leigo

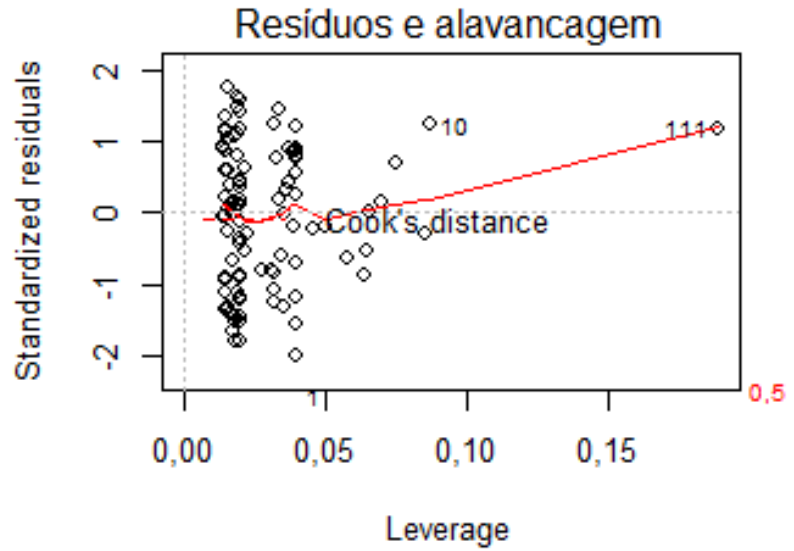
```
> summary(fit1)

Call:
lm(formula = RESP ~ FIBRAS_PGP + GRUPO2)

Residuals:
    Min      1Q  Median      3Q     Max
-9,6146 -4,1178  0,5698  4,1865  8,5078

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      10,4451     0,6900  15,137  < 2e-16 ***
FIBRAS_PGP       -0,7379     0,1366  -5,400 4,74e-07 ***
GRUPO22-Com dor   1,8695     1,0566   1,769     0,08 .
---
Signif. codes:  0 ‘***’ 0,001 ‘**’ 0,01 ‘*’ 0,05 ‘.’ 0,1 ‘ ’ 1

Residual standard error: 4,875 on 97 degrees of freedom
  (18 observations deleted due to missingness)
Multiple R-squared:  0,2428,    Adjusted R-squared:  0,2272
F-statistic: 15,55 on 2 and 97 DF,  p-value: 1,383e-06
```

Gráficos de resíduos





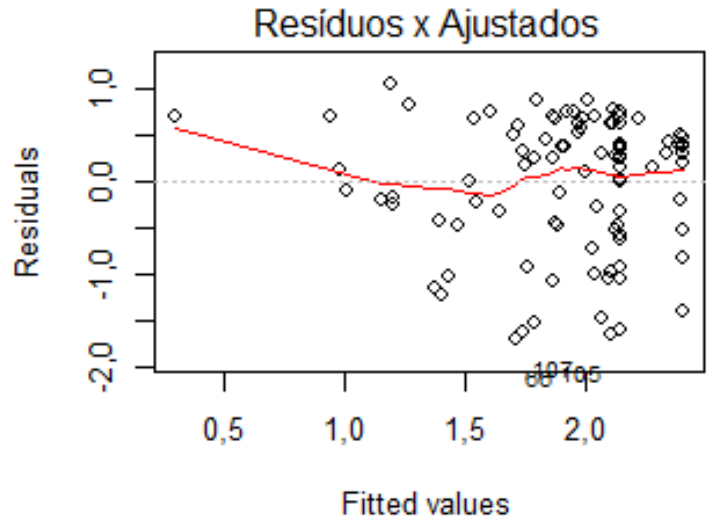


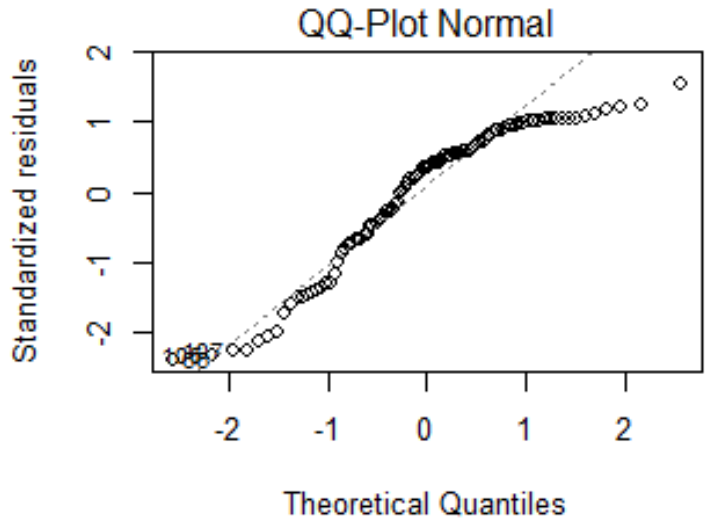


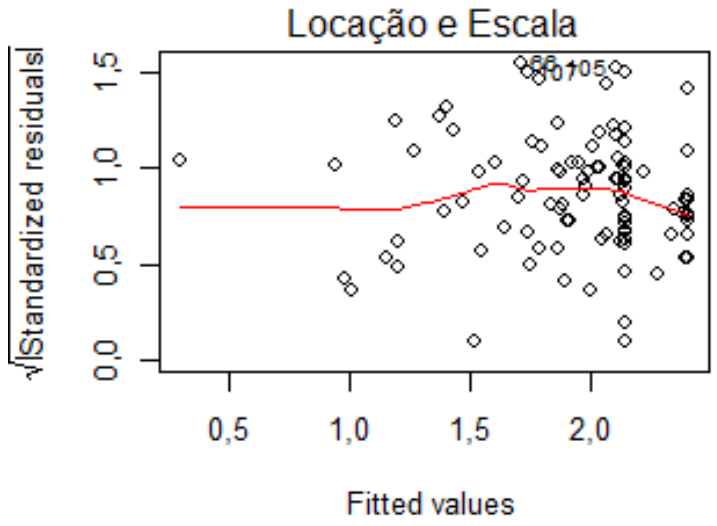
Transformação na resposta

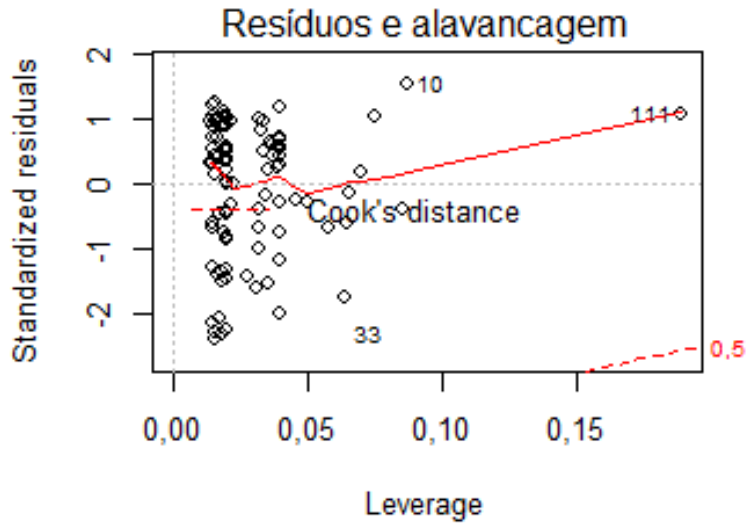
```
fit2<-lm(log(RESP)~FIBRAS_PGP+GRUPO2)
summary(fit2)
par(mfrow=c(2,2))
plot(fit2, caption=c("Resíduos x Ajustados", "QQ-Plot Normal",
                     "Locação e Escala", "", "Resíduos e alavancagem"))
```

Gráficos de resíduos





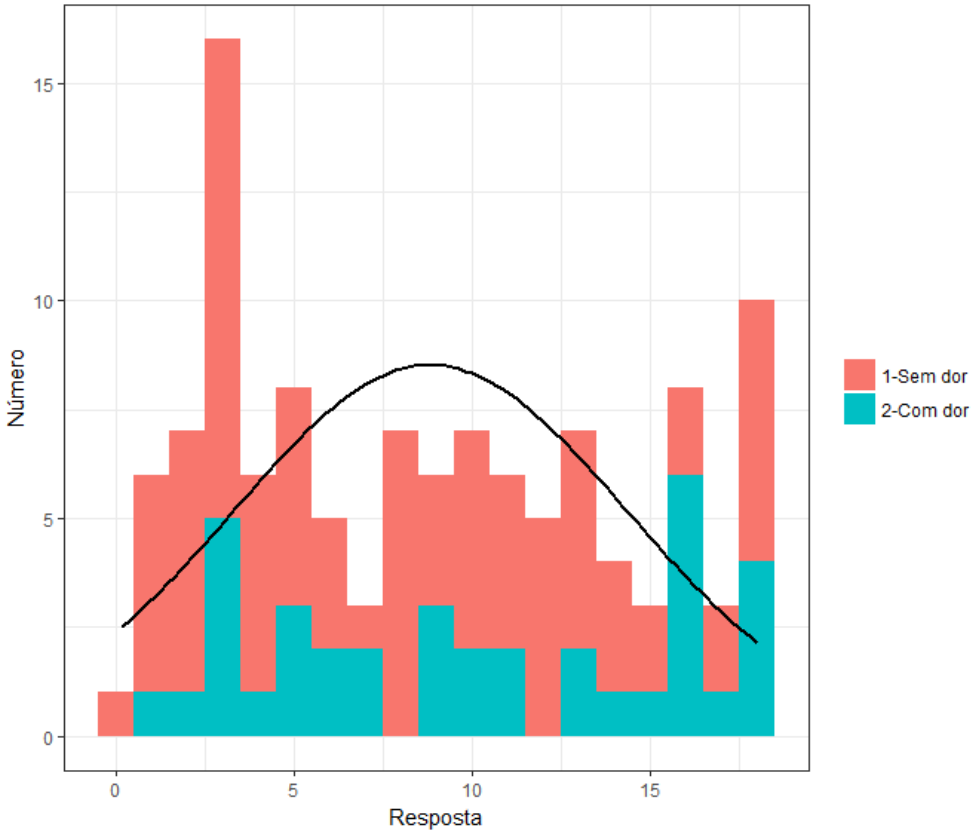




E agora?

- Tentar outras transformações?
- Tentar outros tipos de ajuste?
- Voltar para as estatísticas descritivas!

Característica dos dados



- A resposta não é bem comportada como alguma distribuição conhecida
- Não é discreta, pois assume valores não inteiros no intervalo de 0 a 18
- Mas...
- Há concentração de dados em 18! Indício de dados censurados
- Temos que procurar um modelo para dados censurados!

Referências

HARDIN, J. **Example R code / analysis for housing data.** Disponível em: http://pages.pomona.edu/~jsh04747/courses/math58/Final_examp.html. Acessado em 9 de maio de 2017.

QUICK R. Quick R: Regression Diagnostics. Disponível em: http://www.statmethods.net/stats/rdiagnostics.html. Acessado em 9 de maio de 2017.