lab/exercicios_cap3

Felipe Albuquerque

18 de junho de 2019

Exercícios

1.

A hipótese nula indica que o orçamento para Televisão, Jornais e Rádio não impactam as vendas. Em notação: $H_0^1 : \beta_1 = 0$, $H_0^2 : \beta_2 = 0$, $H_0^3 : \beta_3 = 0$. Para TV(β_1) e Rádio(β_2), podemos rejeitar a hipótese nula. Para jornais(β_3), não podemos.

2.

O KNN (K-nearest-neighbors) é um método de classificação de dados categóricos. Nele, se identifica os valores mais próximos de x_0 e depois estima a probabilidade de x_0 ser daquela categoria. O modelo de regressão KNN, da mesma forma, identifica os valores mais próximos de x_0, e depois estima $f(x_0)$ como a média das respostas (na training data) entre esses "vizinhos".

3.

A função fica $Y = 50 + 20gpa + 0.07iq + 35gender + 0.01gpa * iq - 10gpa * gender$

a

b

```
iq = 110
gpa = 4
gender = 1

Y = 50 + 20*gpa + 0.07*iq + 35*gender + 0.01*110*4 - 10*gpa*gender
Y
```

```
## [1] 137.1
```

c

Falso. O parâmetro da interação só mede a intensidade da interação. Para verificar se há um efeito significativo da interação, devemoz conduzir um teste de hipótese e olhar o p-valor.

4

a

A primeira vista, esperamos que a regressão linear tenha um RSS menor, já que a relação entre X e Y é linear.

b

c

Aqui acontece o oposto. A regressão polinomial é mais flexível que a linear. Como a relação entre X e Y é não-linear, esperamos portanto um menor RSS com a polinomial.

d

5

$$\hat{y}_i = x_i \hat{\beta}$$

e:

$$\hat{\beta} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} x_{i'}^2}$$

Então:

$$\hat{y}_i = x_i \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i'=1}^{n} x_{i'}^2}$$

logo:

$$\hat{y}_i = \sum_{i=1}^{n} (\frac{x_i y_i * x_i}{\sum_{i'=1}^{n} x_{i'}^2})$$

$$\hat{y}_i = \sum_{i=1}^{n} (\frac{x_i * x_i}{\sum_{i'=1}^{n} x_{i'}^2} y_i)$$

$$a_{i'} = \frac{x_i * x_i}{\sum_{i'=1}^{n} x_{i'}^2}$$

6

7

8

a

```r
auto = Auto
lm1 = lm(mpg ~ horsepower, data = auto)
summary(lm1)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower, data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 39.935861   0.717499   55.66 <0.0000000000000002 ***
## horsepower  -0.157845   0.006446  -24.49 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.906 on 390 degrees of freedom
## Multiple R-squared:  0.6059, Adjusted R-squared:  0.6049
## F-statistic: 599.7 on 1 and 390 DF,  p-value: < 0.00000000000000022
```

i Há uma associação clara entre horsepower e mpg.

ii o p valor é menor que $2e^{-16}$, o que indica uma associação forte.

iii associação negativa, já que o coeficiente é de -0.157.

$iiii$ Para horsepower - 98, valor de mpg =

```r
hpr98 <- data.frame(horsepower=98)
predict(lm1, hpr98)
```

```
##        1
## 24.46708
```

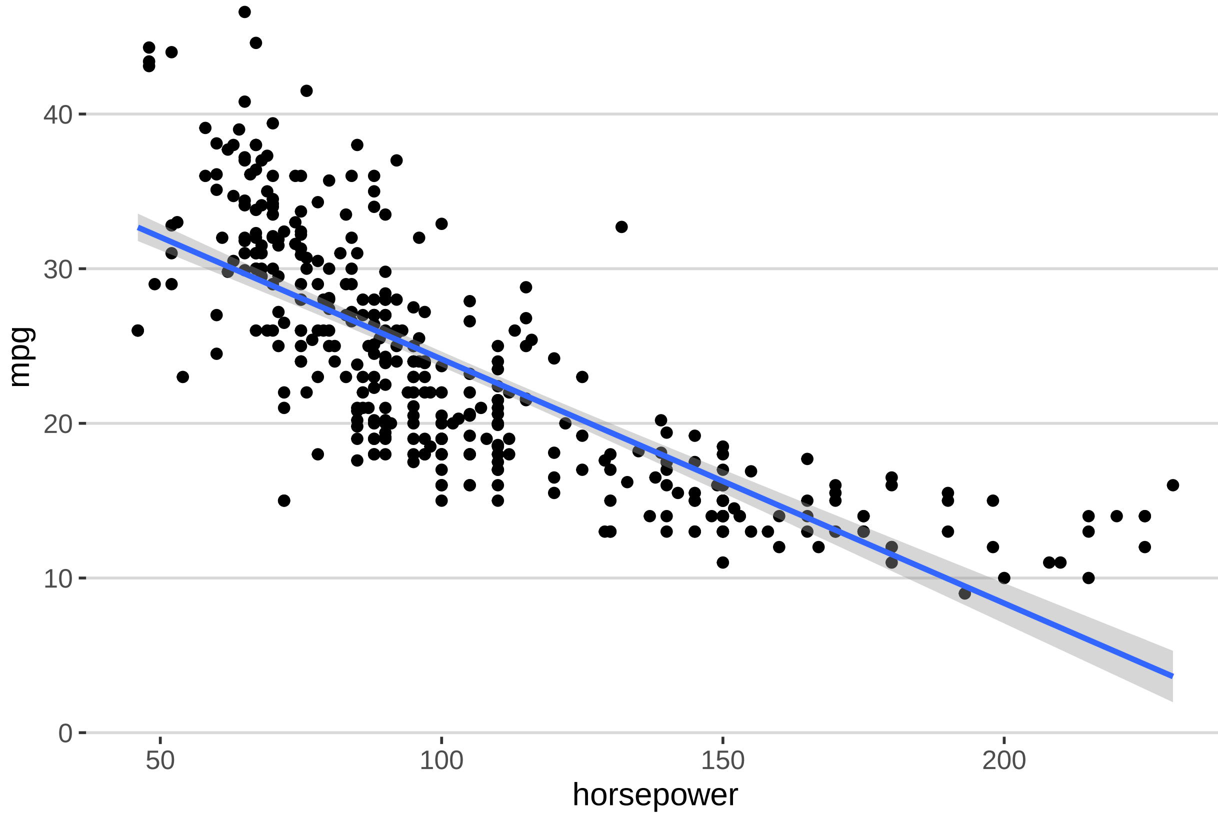
Intervalo de confiança:

```r
predict(lm1, hpr98, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 24.46708 23.97308 24.96108
```

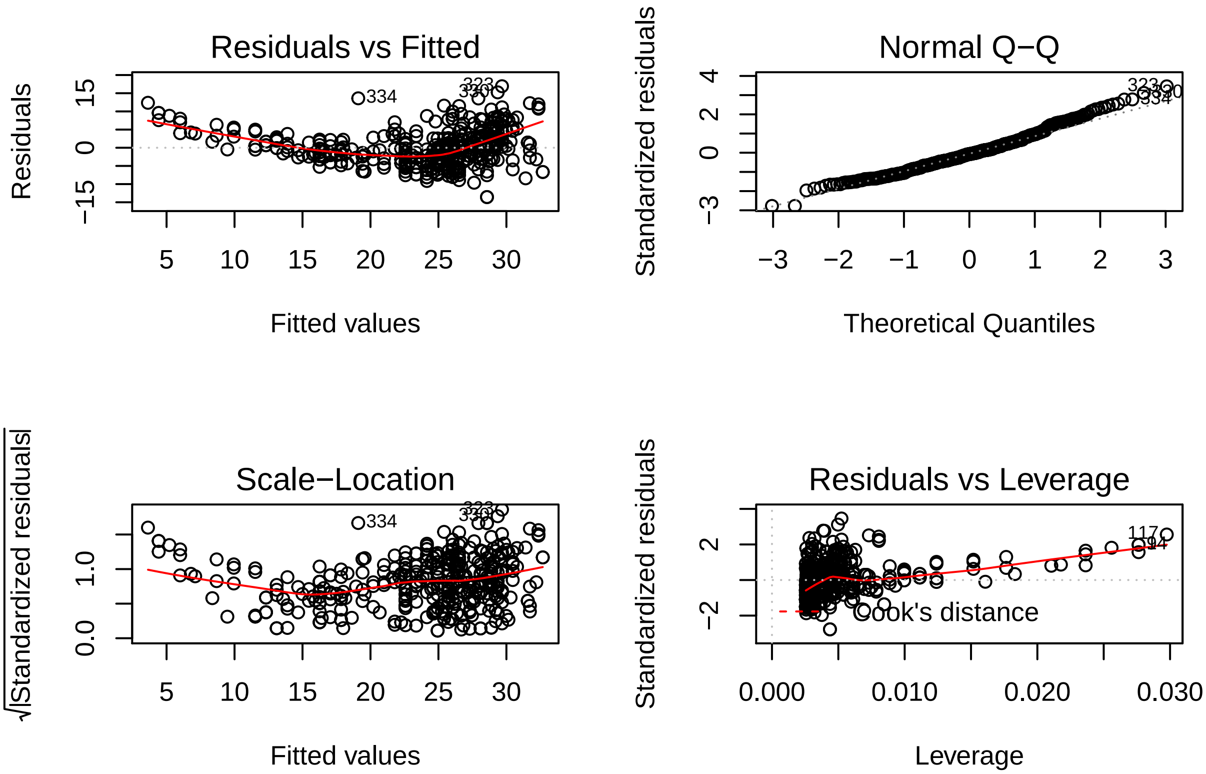
b

```r
auto %>%
  ggplot(aes(x = horsepower, y = mpg)) + geom_point() +
  geom_smooth(method = "lm") +
  theme_hc()
```



c

```
par(mfrow=c(2,2))
plot(lm1)
```

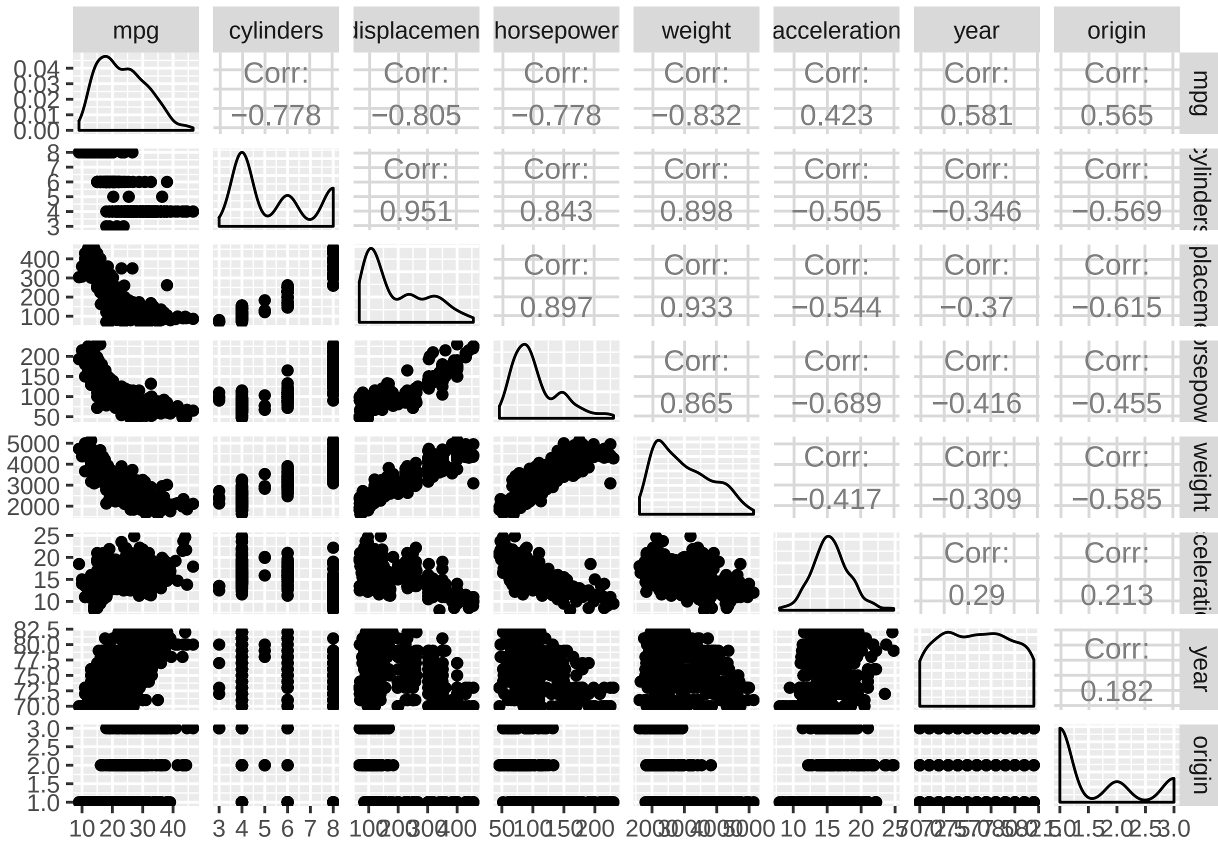


Há uma não-linearidade dos resíduos.

9

a

```r
auto %>%
  dplyr::select(-name) %>%
  GGally::ggpairs()
```



b

```r
auto %>%
  dplyr::select(1:8) %>%
  cor()
```

```
##                     mpg  cylinders displacement horsepower     weight
## mpg           1.0000000 -0.7776175   -0.8051269 -0.7784268 -0.8322442
## cylinders    -0.7776175  1.0000000    0.9508233  0.8429834  0.8975273
## displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
## horsepower   -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
## weight       -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
## acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
## year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
## origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
##              acceleration       year     origin
## mpg             0.4233285  0.5805410  0.5652088
## cylinders      -0.5046834 -0.3456474 -0.5689316
## displacement   -0.5438005 -0.3698552 -0.6145351
## horsepower     -0.6891955 -0.4163615 -0.4551715
## weight         -0.4168392 -0.3091199 -0.5850054
## acceleration    1.0000000  0.2903161  0.2127458
## year            0.2903161  1.0000000  0.1815277
## origin          0.2127458  0.1815277  1.0000000
```

c

```r
lm2 <- lm(mpg ~ . -name, data = auto)
summary(lm2)
```

```
##
## Call:
## lm(formula = mpg ~ . - name, data = auto)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5903 -2.1565 -0.1169  1.8690 13.0604
##
## Coefficients:
##                Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  -17.218435   4.644294  -3.707              0.00024 ***
## cylinders     -0.493376   0.323282  -1.526              0.12780
## displacement   0.019896   0.007515   2.647              0.00844 **
## horsepower    -0.016951   0.013787  -1.230              0.21963
## weight        -0.006474   0.000652  -9.929 < 0.0000000000000002 ***
## acceleration   0.080576   0.098845   0.815              0.41548
## year           0.750773   0.050973  14.729 < 0.0000000000000002 ***
## origin         1.426141   0.278136   5.127          0.000000467 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.328 on 384 degrees of freedom
## Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
## F-statistic: 252.4 on 7 and 384 DF,  p-value: < 0.00000000000000022
```

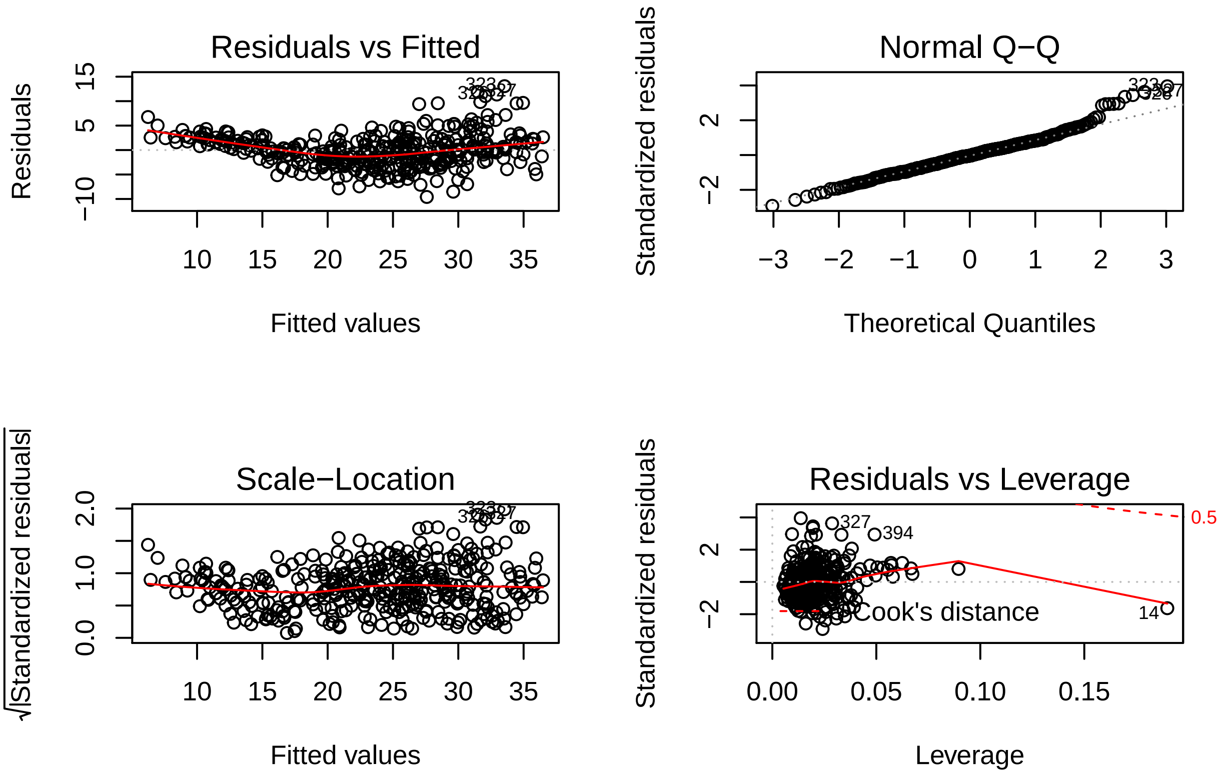
i Existe

ii displacement (0.019896) ; weight (-0.006474); year (0.750773) e origin (1.426141)

iii versões mais recentes de carros conseguem consumir menos combustível

d

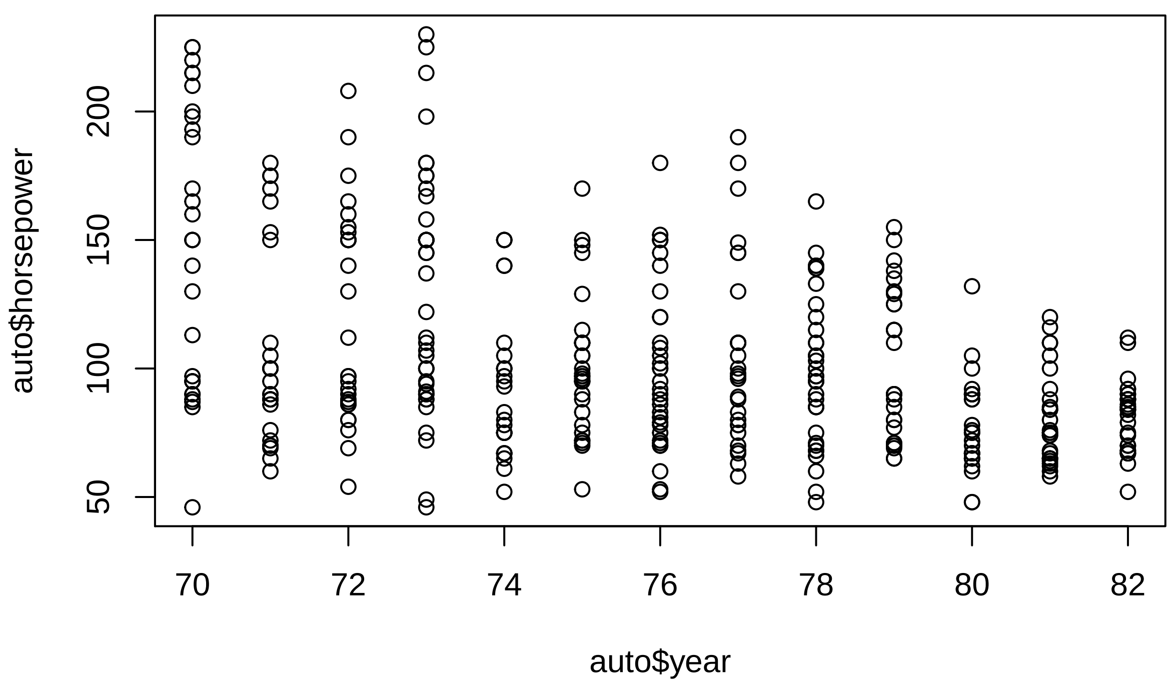
```r
par(mfrow=c(2,2))
plot(lm2)
```



Os erros estão mais lineares, o que aponta para uma homocedasticidade desejada.

e

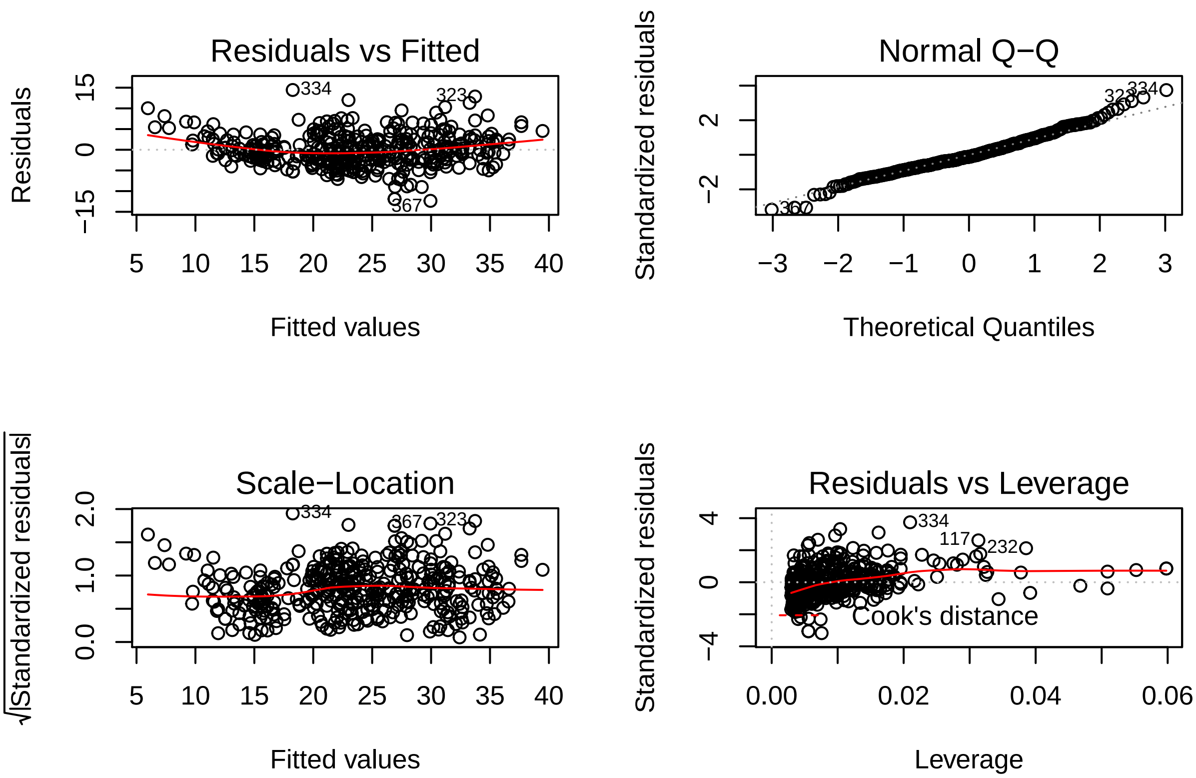
```
plot(auto$year, auto$horsepower)
```



```r
lm3 <- lm(mpg ~ horsepower + year + horsepower*year, data = auto)
summary(lm3)
```

```
##
## Call:
## lm(formula = mpg ~ horsepower + year + horsepower * year, data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.3492  -2.4509  -0.4557   2.4056  14.4437
##
## Coefficients:
##                    Estimate  Std. Error t value            Pr(>|t|)
## (Intercept)     -126.608853   12.117256 -10.449 <0.0000000000000002 ***
## horsepower         1.045674    0.115374   9.063 <0.0000000000000002 ***
## year               2.191976    0.161350  13.585 <0.0000000000000002 ***
## horsepower:year   -0.015959    0.001562 -10.217 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.901 on 388 degrees of freedom
## Multiple R-squared:  0.7522, Adjusted R-squared:  0.7503
## F-statistic: 392.5 on 3 and 388 DF,  p-value: < 0.00000000000000022
```

```r
par(mfrow=c(2,2))
plot(lm3)
```

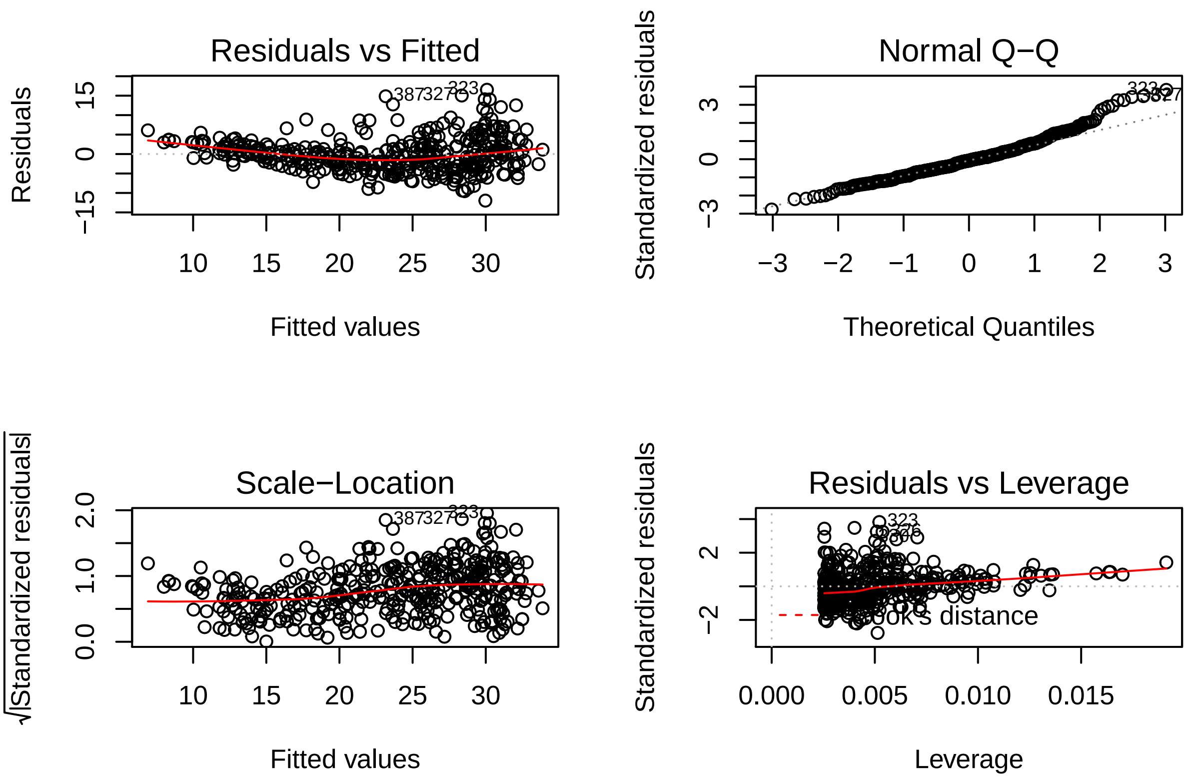


f

```r
lm4 <- lm(mpg ~ weight, data = auto)
summary(lm4)
```

```
##
## Call:
## lm(formula = mpg ~ weight, data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.9736  -2.7556  -0.3358   2.1379  16.5194
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 46.216524   0.798673   57.87 <0.0000000000000002 ***
## weight      -0.007647   0.000258  -29.64 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.333 on 390 degrees of freedom
## Multiple R-squared:  0.6926, Adjusted R-squared:  0.6918
## F-statistic: 878.8 on 1 and 390 DF,  p-value: < 0.00000000000000022
```

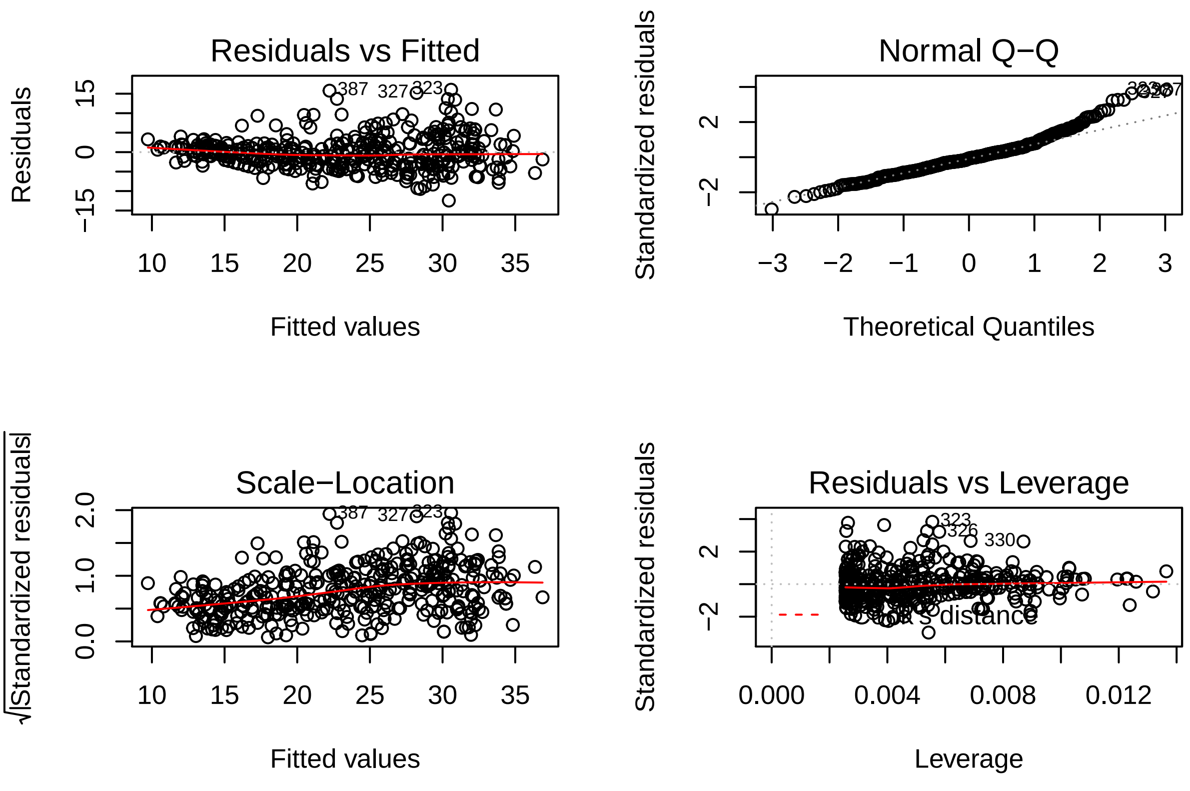
```r
par(mfrow = c(2,2))
plot(lm4)
```




```r
lm5 <- lm(mpg ~ log(weight), data = auto)
summary(lm5)
```

```
##
## Call:
## lm(formula = mpg ~ log(weight), data = auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.4315  -2.6752  -0.2888   1.9429  16.0136
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) 209.9433     6.0002   34.99 <0.0000000000000002 ***
## log(weight) -23.4317     0.7534  -31.10 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.189 on 390 degrees of freedom
## Multiple R-squared:  0.7127, Adjusted R-squared:  0.7119
## F-statistic: 967.3 on 1 and 390 DF,  p-value: < 0.00000000000000022
```

```r
par(mfrow = c(2,2))
plot(lm5)
```



Normalizando a variável weight, os erros ficam praticamente lineares.

10

```r
rm(list=ls())
data = Carseats
```

a

```r
lm1 <- lm(Sales ~ Population + Urban + US, data = data)
summary(lm1)
```

```
##
## Call:
## lm(formula = Sales ~ Population + Urban + US, data = data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.3323 -1.9844 -0.0824  1.8783  8.4053
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  6.7262086  0.4009409  16.776 < 0.0000000000000002 ***
## Population   0.0007415  0.0009499   0.781             0.435475
## UrbanYes    -0.1341034  0.3063701  -0.438             0.661830
## USYes        1.0360741  0.2921241   3.547             0.000437 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.787 on 396 degrees of freedom
## Multiple R-squared:  0.03342,    Adjusted R-squared:  0.02609
## F-statistic: 4.563 on 3 and 396 DF,  p-value: 0.003713
```

b

O aumento de uma unidade da **população** aumenta em 0.07 as unidades vendidas (0.0007);

Em **áreas urbanas**, as **vendas** são menores em 22 mil unidades (-0.0219)

As **vendas** aumentam em até 1000 unidades em locais dentro dos **Estados Unidos** (1.0360)

c

$$Sales = 6.7626 \ + 0.0007(Population) \ - 0.1341(Urban) \ + 1.0360(US)$$

d

Pode-se rejeitar a hipótese nula para USYes.

e

```r
lm2 <- lm(Sales ~ US, data = data)
summary(lm2)
```

```
##
## Call:
## lm(formula = Sales ~ US, data = data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.497 -1.929 -0.105  1.836  8.403
##
## Coefficients:
```

```
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)   6.8230     0.2335   29.21 < 0.0000000000000002 ***
## USYes         1.0439     0.2908    3.59             0.000372 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.783 on 398 degrees of freedom
## Multiple R-squared:  0.03136,    Adjusted R-squared:  0.02893
## F-statistic: 12.89 on 1 and 398 DF,  p-value: 0.0003723
```

f

O segundo modelo e o primeiro contam um X^2 muito pequeno, de apenas 0.02. O RSE do segundo modelo é um pouco menor.

g

```
confint(lm2)
```

```
##                 2.5 %   97.5 %
## (Intercept) 6.3638993 7.282157
## USYes       0.4721887 1.615553
```

h

```
outlierTest(lm1)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 377  3.05528          0.0024011      0.96043
```

11

```
rm(list=ls())
set.seed(1)
x <- rnorm(100)
y <- 2*x + rnorm(100)
```

a

```
lm1 <- lm(y ~ x -1)
summary(lm1)
```

```
##
## Call:
## lm(formula = y ~ x - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
```

```
##
## Coefficients:
##   Estimate Std. Error t value            Pr(>|t|)
## x   1.9939     0.1065   18.73 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 0.00000000000000022
```

b

```r
lm2 <- lm(x ~ y -1)
summary(lm2)
```

```
##
## Call:
## lm(formula = x ~ y - 1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
##
## Coefficients:
##   Estimate Std. Error t value            Pr(>|t|)
## y  0.39111    0.02089   18.73 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 0.00000000000000022
```

c

Pelo visto, obtemos os mesmo valores para a estatística-t, e consequentemente, para o p-valor. Em outras palavras, $y = 1.99x + \epsilon$ é igual a $x = 0.39y + \epsilon$.

d

e

f

```r
lm3 <- lm(y ~ x)
summary(lm3)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -1.8768 -0.6138 -0.1395  0.5394  2.3462
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -0.03769    0.09699  -0.389               0.698
## x            1.99894    0.10773  18.556 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9628 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 0.00000000000000022
```

```
##
## Call:
## lm(formula = x ~ y)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90848 -0.28101  0.06274  0.24570  0.85736
##
## Coefficients:
##             Estimate Std. Error t value            Pr(>|t|)
## (Intercept)  0.03880    0.04266    0.91               0.365
## y            0.38942    0.02099   18.56 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4249 on 98 degrees of freedom
## Multiple R-squared:  0.7784, Adjusted R-squared:  0.7762
## F-statistic: 344.3 on 1 and 98 DF,  p-value: < 0.00000000000000022
```

De novo, ambas estatísticas-t se assemelham.

12

a

Se $\hat{\beta} = \frac{\sum_i x_i y_i}{\sum_j x_j^2}$ e $\hat{\beta}' = \frac{\sum_i x_i y_i}{\sum_j y_j^2}$, então os coeficientes são iguais se:

$$\sum_j x_j^2 = \sum_j y_j^2$$

b

c

13

a

```r
rm(list=ls())
```

Table 1:

	Dependent variable:	
	y	x
	(1)	(2)
x	2.000***	
	(0.0001)	
y		0.500***
		(0.00004)
Constant	0.018	−0.009
	(0.017)	(0.009)
Observations	200	200
R^2	1.000	1.000
Adjusted R^2	1.000	1.000
Residual Std. Error (df = 198)	0.121	0.060
F Statistic (df = 1; 198)	182,633,689.000***	182,633,689.000***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Table 2:

	Dependent variable:	
	y	x
	(1)	(2)
x	0.006	
	(0.032)	
y		0.006
		(0.031)
Constant	0.992***	0.992***
	(0.032)	(0.032)
Observations	1,000	1,000
R^2	0.00004	0.00004
Adjusted R^2	−0.001	−0.001
Residual Std. Error (df = 998)	0.104	0.104
F Statistic (df = 1; 998)	0.041	0.041
Note:	*p<0.1; **p<0.05; ***p<0.01	

```r
set.seed(1)
x <- rnorm(1000, 0, 1)
```

b

```r
eps <- rnorm(1000, 0, 0.25)
```

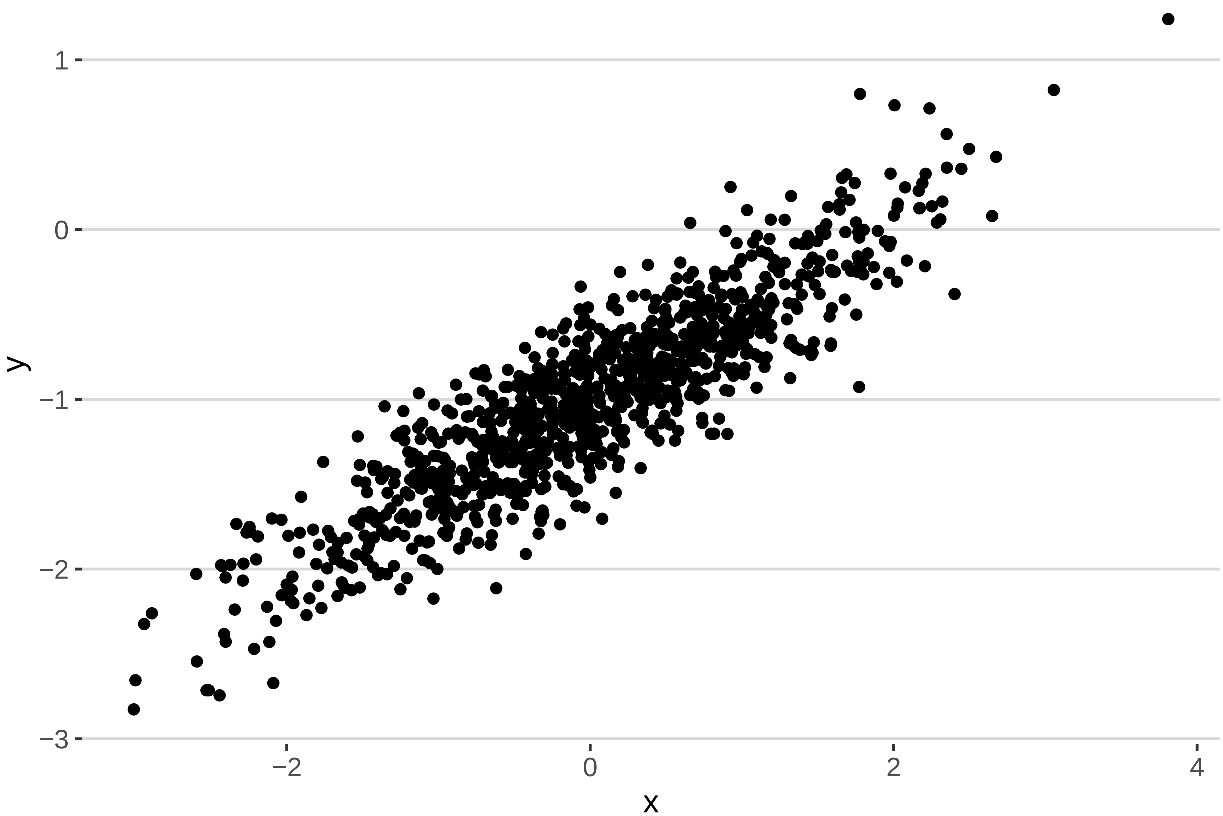
c

```r
y <- -1 + 0.5*x + eps  # eps=epsilon=e
length(y)
```

```
## [1] 1000
```

d

```r
data = as.data.frame(cbind(x, y))
data %>%
ggplot(aes(x = x, y = y)) + geom_point() + theme_hc()
```



Parece haver uma correlação positiva e linear entre as variáveis

e

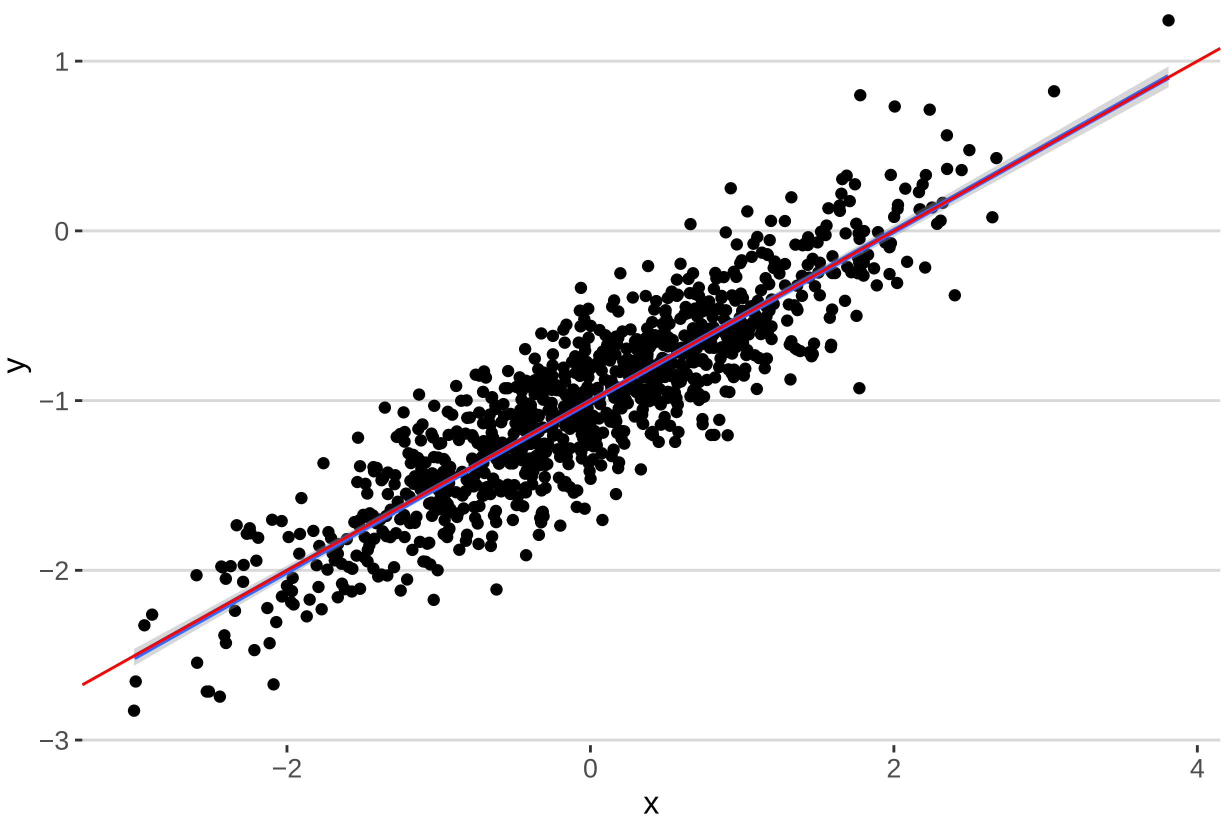
```r
lm1 <- lm(y ~ x)
summary(lm1)
```

```
##
## Call:
## lm(formula = y ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.81211 -0.16799 -0.00344  0.18885  0.91108
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -1.004047   0.008226 -122.05 <0.0000000000000002 ***
## x            0.501608   0.007952   63.08 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2601 on 998 degrees of freedom
## Multiple R-squared:  0.7995, Adjusted R-squared:  0.7993
## F-statistic:  3979 on 1 and 998 DF,  p-value: < 0.00000000000000022
```

Os parâmetros se assemelham.

f

```
data %>%
  ggplot(aes(x = x, y = y)) + geom_point() +
  geom_smooth(method = "lm") + geom_abline(aes(intercept = -1, slope = 0.5), col = "red") +
  theme_hc()
```



g

```r
lm2 <- lm(y ~ x + I(x^2))
summary(lm2)
```

```
##
## Call:
## lm(formula = y ~ x + I(x^2))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82236 -0.16995 -0.00384  0.18910  0.90073
##
## Coefficients:
##              Estimate Std. Error  t value            Pr(>|t|)
## (Intercept) -1.009147   0.010077 -100.147 <0.0000000000000002 ***
## x            0.501814   0.007957   63.069 <0.0000000000000002 ***
## I(x^2)       0.004768   0.005440    0.877               0.381
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2602 on 997 degrees of freedom
## Multiple R-squared:  0.7996, Adjusted R-squared:  0.7992
## F-statistic:  1989 on 2 and 997 DF,  p-value: < 0.00000000000000022
```

Não. Pelo contrário.

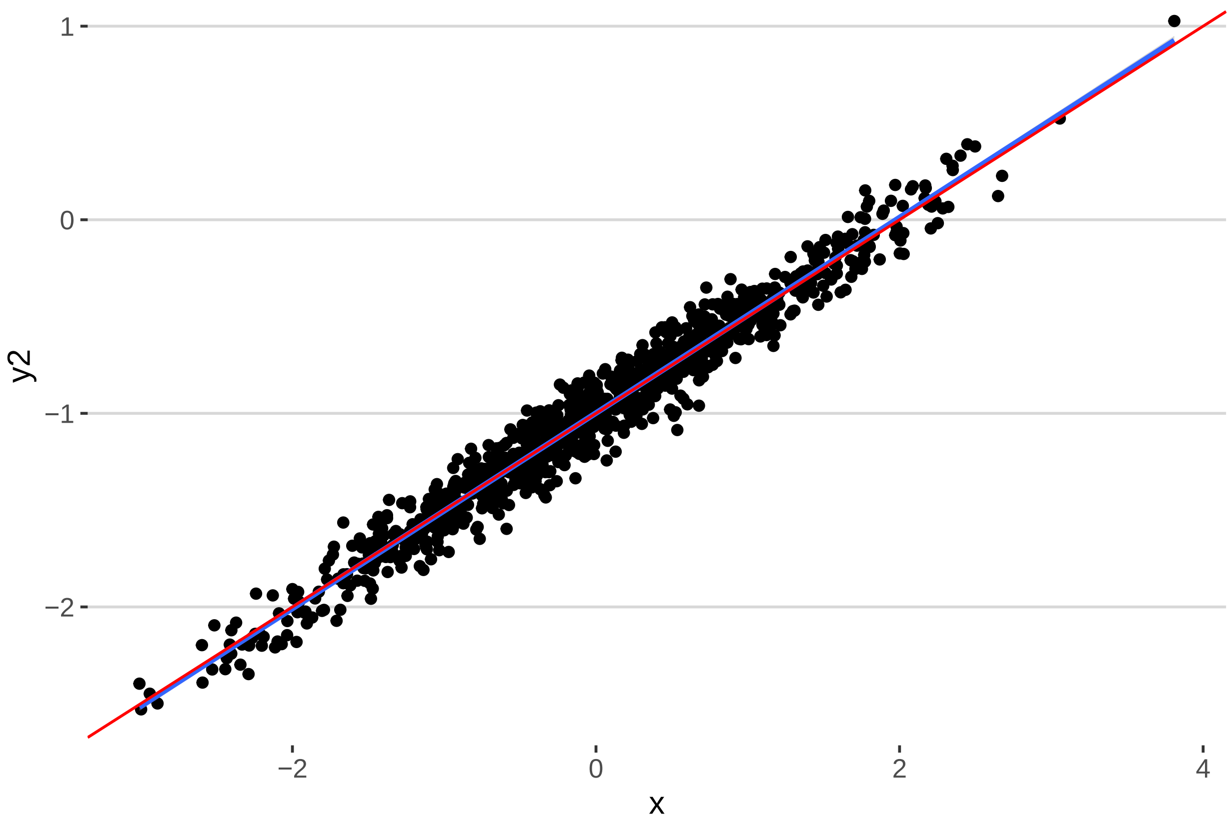
h

```r
eps2 <- rnorm(1000, 0, 0.1)
y2 <- -1 + 0.5*x +eps2
lm3 <- lm(y2 ~ x)
summary(lm3)
```

```
##
## Call:
## lm(formula = y2 ~ x)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.35818 -0.06414 -0.00179  0.06866  0.28105
##
## Coefficients:
##              Estimate Std. Error t value            Pr(>|t|)
## (Intercept) -0.998412   0.003259  -306.4 <0.0000000000000002 ***
## x            0.504919   0.003150   160.3 <0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.103 on 998 degrees of freedom
## Multiple R-squared:  0.9626, Adjusted R-squared:  0.9626
## F-statistic: 2.569e+04 on 1 and 998 DF,  p-value: < 0.00000000000000022
```

```r
data %>%
  ggplot(aes(x = x, y = y2)) + geom_point() +
```

```
  geom_smooth(method = "lm") + geom_abline(aes(intercept = -1, slope = 0.5), col = "red") +
  theme_hc()
```



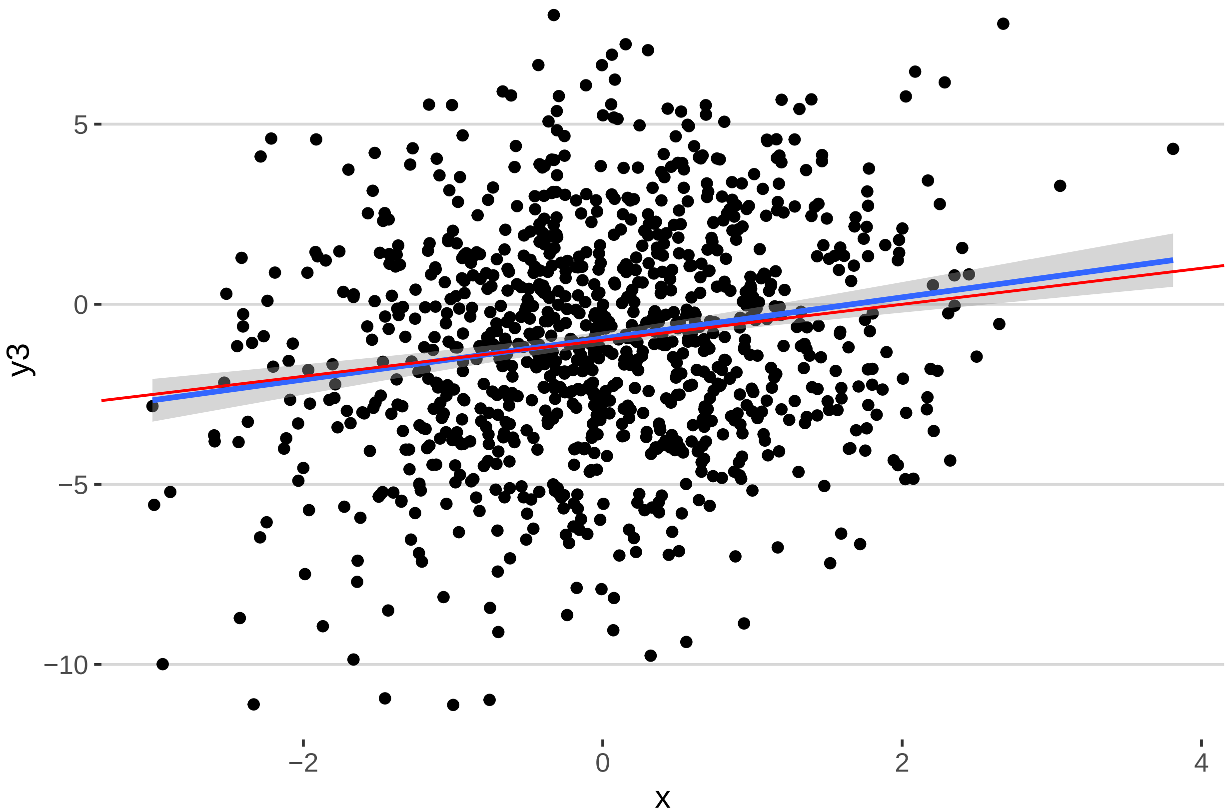
i

```
eps3 <- rnorm(1000, sd=3)  # orig sd was 0.5
y3 <- -1 + 0.5*x + eps3
lm4 <- lm(y3 ~ x)
summary(lm4)
```

```
##
## Call:
## lm(formula = y3 ~ x)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.6046 -2.0975 -0.0328  2.1509  9.1657
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept) -0.94901    0.09856  -9.628 < 0.0000000000000002 ***
## x            0.57067    0.09528   5.989         0.00000000294 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.117 on 998 degrees of freedom
## Multiple R-squared:  0.0347, Adjusted R-squared:  0.03373
```

```
## F-statistic: 35.87 on 1 and 998 DF,  p-value: 0.000000002937
```

```r
data %>%
  ggplot(aes(x = x, y = y3)) + geom_point() +
  geom_smooth(method = "lm") + geom_abline(aes(intercept = -1, slope = 0.5), col = "red") +
  theme_hc()
```



O intervalo de confiança aumenta. O R^2 diminui consideravelmente, devido a grande variabilidade dos dados.

j

```r
confint(lm1)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0201895 -0.9879040
## x            0.4860032  0.5172131
```

```r
confint(lm3)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.0048061 -0.9920175
## x            0.4987378  0.5111003
```

```r
confint(lm4)
```

```
##                  2.5 %     97.5 %
## (Intercept) -1.1424261 -0.7555941
## x            0.3836967  0.7576412
```

Quanto menor a variância, menor o intervalo de confiança.

____- # 14

a

```
rm(list=ls())
set.seed(1)
x1 = runif(100)
x2 = 0.5*x1 + rnorm(100)/10
x3 = 2 + 2*x1 + 0.3*x2 + rnorm(100)
```

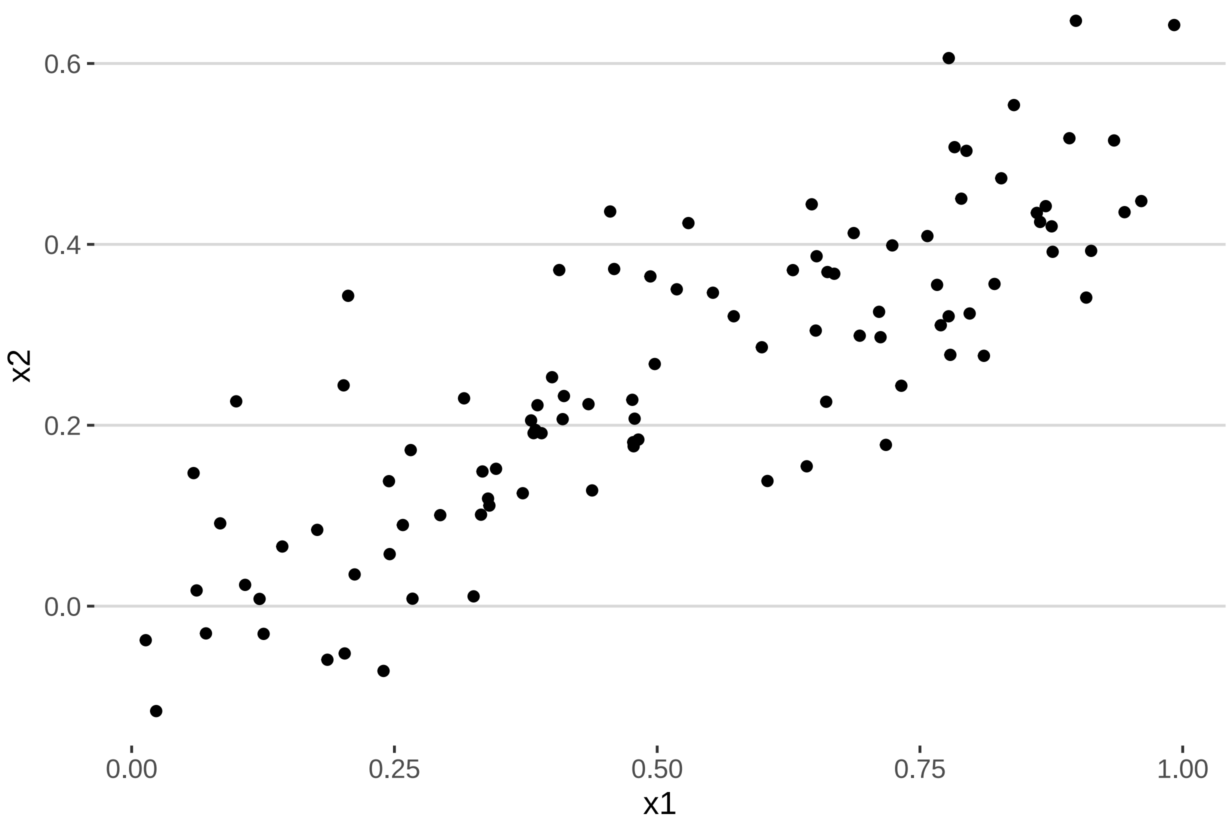
$\beta_0 = 2$, $\beta_1 = 2$, $\beta_2 = 0.3$

b

```
data = as.data.frame(cbind(x1, x2))
cor(x1, x2)
```

```
## [1] 0.8351212
```

```
data %>%
ggplot(aes(x = x1, y = x2)) + geom_point() + theme_hc()
```



c

```
lm1 <- lm(x3 ~ x1 + x2)
summary(lm1)
```

```
##
## Call:
## lm(formula = x3 ~ x1 + x2)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8311 -0.7273 -0.0537  0.6338  2.3359
##
## Coefficients:
##             Estimate Std. Error t value           Pr(>|t|)
## (Intercept)   2.1305     0.2319   9.188 0.00000000000000761 ***
## x1            1.4396     0.7212   1.996              0.0487 *
## x2            1.0097     1.1337   0.891              0.3754
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 97 degrees of freedom
## Multiple R-squared:  0.2088, Adjusted R-squared:  0.1925
## F-statistic:  12.8 on 2 and 97 DF,  p-value: 0.00001164
```

Coeficiente $\beta_1 = 1.43$ e Coefiencie $\beta_2 = 1.0097$. A hipótese nula pode ser rejeitada a 5% no primeiro caso, mas não no segundo.

d

```r
lm2 <- lm(x3 ~ x1)
```

e

Table 3:

	Dependent variable:	
	x3	
	(1)	(2)
x1	1.976^{***}	
	(0.396)	
x2		2.900^{***}
		(0.633)
Constant	2.112^{***}	2.390^{***}
	(0.231)	(0.195)
Observations	100	100
R^2	0.202	0.176
Adjusted R^2	0.194	0.168
Residual Std. Error (df = 98)	1.055	1.072
F Statistic (df = 1; 98)	24.862^{***}	20.980^{***}
Note:	*p<0.1; **p<0.05; ***p<0.01	

Nos dois casos daria para rejeitar a hipótese nula

f

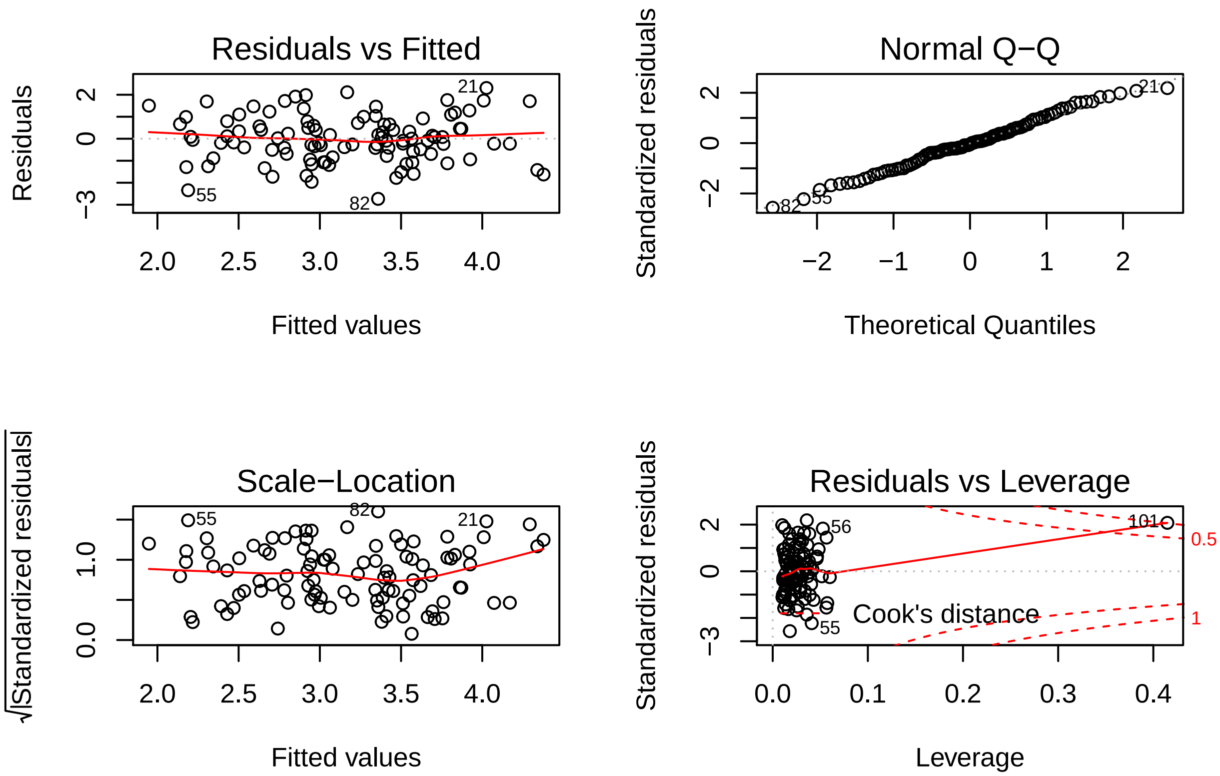
Não. Isso acontece apenas pela multicolinearidade entre as variáveis x1 e x2.

g

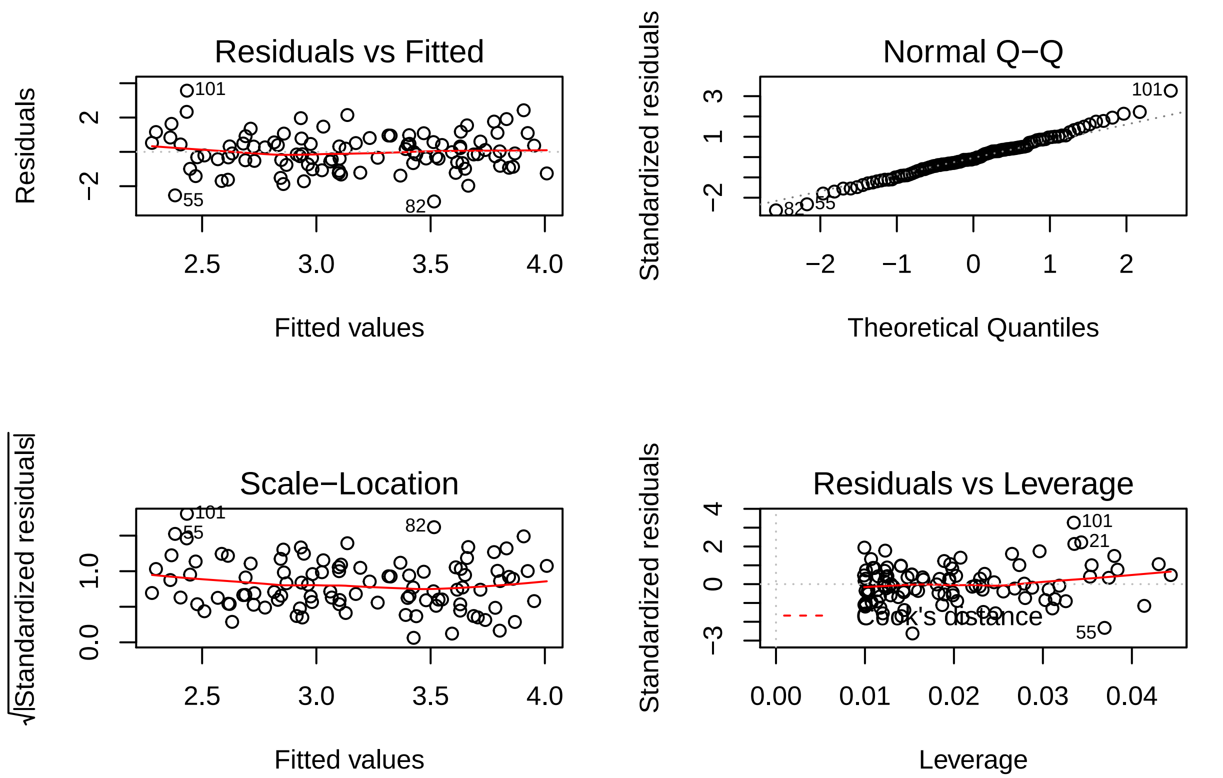
```
x1=c(x1,0.1)
x2=c(x2,0.8)
y=c(x3,6)
```

```
lm4 <- lm(y ~ x1 + x2)
lm5 <- lm(y ~ x1)
lm6 <- lm(y ~ x2)
```

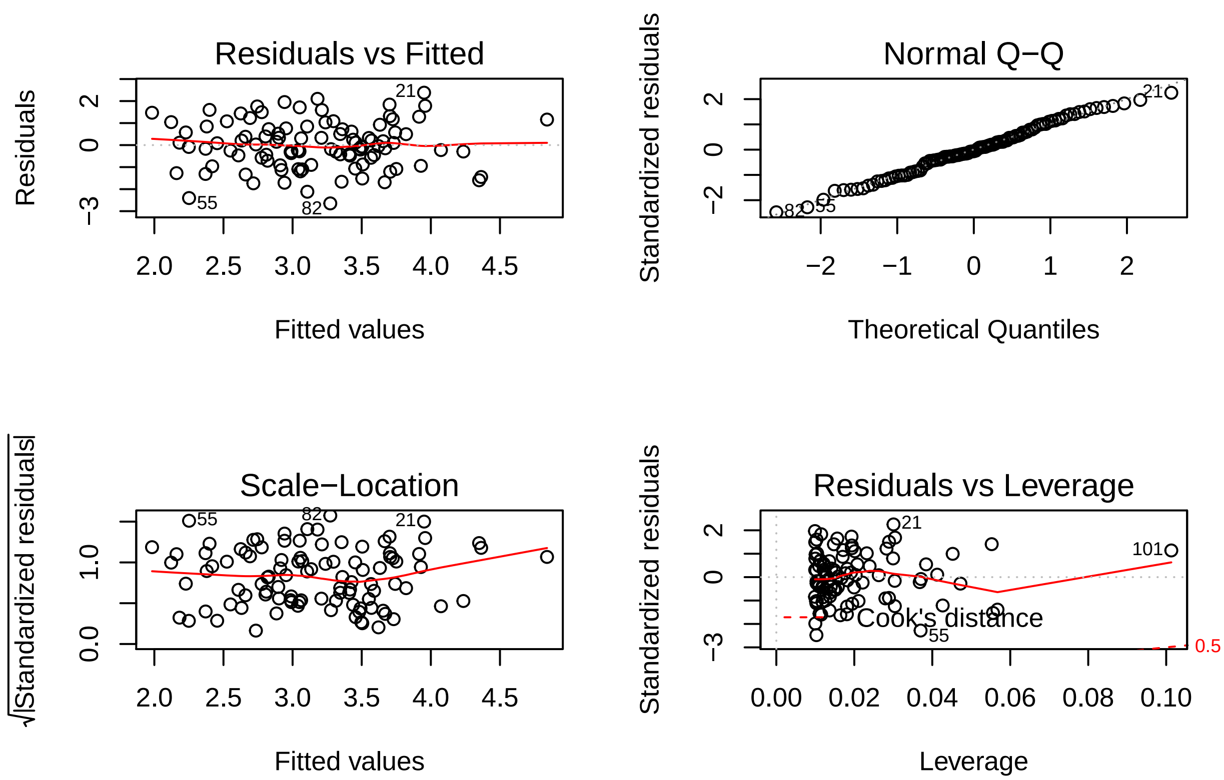
```
par(mfrow=c(2,2))
plot(lm4)
```




```
par(mfrow=c(2,2))
plot(lm5)
```




```
par(mfrow=c(2,2))
plot(lm6)
```



15

```r
rm(list=ls())
boston = Boston
attach(boston)
```

a

Table 4:

	Dependent variable:				
	crim				
	(1)	(2)	(3)	(4)	(5)
zn	-0.074^{***}				
	(0.016)				
indus		0.510^{***}			
		(0.051)			
chas			-1.893		
			(1.506)		
nox				31.249^{***}	
				(2.999)	
rm					-2.684^{***}
					(0.532)
Constant	4.454^{***}	-2.064^{***}	3.744^{***}	-13.720^{***}	20.482^{***}
	(0.417)	(0.667)	(0.396)	(1.699)	(3.364)
Observations	506	506	506	506	506
R^2	0.040	0.165	0.003	0.177	0.048
Adjusted R^2	0.038	0.164	0.001	0.176	0.046
Residual Std. Error (df = 504)	8.435	7.866	8.597	7.810	8.401
F Statistic (df = 1; 504)	21.103^{***}	99.817^{***}	1.579	108.555^{***}	25.450^{***}
Note:			*p<0.1; **p<0.05; ***p<0.01		

Isoladamente, todas as variáveis são significativas, menos **chas**

b

```r
lm1 <- lm(crim ~ ., data = boston)
summary(lm1)
```

```
##
## Call:
## lm(formula = crim ~ ., data = boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

Table 5:

	Dependent variable:			
	crim			
	(1)	(2)	(3)	(4)
age	0.108***			
	(0.013)			
dis		−1.551***		
		(0.168)		
rad			0.618***	
			(0.034)	
tax				0.030***
				(0.002)
Constant	−3.778***	9.499***	−2.287***	−8.528***
	(0.944)	(0.730)	(0.443)	(0.816)
Observations	506	506	506	506
R^2	0.124	0.144	0.391	0.340
Adjusted R^2	0.123	0.142	0.390	0.338
Residual Std. Error (df = 504)	8.057	7.965	6.718	6.997
F Statistic (df = 1; 504)	71.619***	84.888***	323.935***	259.190***
Note:				*p<0.1; **p<0.05; ***p<0.01

Table 6:

	Dependent variable:			
	crim			
	(1)	(2)	(3)	(4)
ptratio	1.152***			
	(0.169)			
black		−0.036***		
		(0.004)		
lstat			0.549***	
			(0.048)	
medv				−0.363***
				(0.038)
Constant	−17.647***	16.554***	−3.331***	11.797***
	(3.147)	(1.426)	(0.694)	(0.934)
Observations	506	506	506	506
R^2	0.084	0.148	0.208	0.151
Adjusted R^2	0.082	0.147	0.206	0.149
Residual Std. Error (df = 504)	8.240	7.946	7.664	7.934
F Statistic (df = 1; 504)	46.259***	87.740***	132.035***	89.486***
Note:				*p<0.1; **p<0.05; ***p<0.01

```
## -9.924 -2.120 -0.353  1.019 75.051
##
## Coefficients:
##               Estimate Std. Error t value        Pr(>|t|)
## (Intercept)  17.033228   7.234903   2.354        0.018949 *
## zn            0.044855   0.018734   2.394        0.017025 *
## indus        -0.063855   0.083407  -0.766        0.444294
## chas         -0.749134   1.180147  -0.635        0.525867
## nox         -10.313535   5.275536  -1.955        0.051152 .
## rm            0.430131   0.612830   0.702        0.483089
## age           0.001452   0.017925   0.081        0.935488
## dis          -0.987176   0.281817  -3.503        0.000502 ***
## rad           0.588209   0.088049   6.680 0.0000000000646 ***
## tax          -0.003780   0.005156  -0.733        0.463793
## ptratio      -0.271081   0.186450  -1.454        0.146611
## black        -0.007538   0.003673  -2.052        0.040702 *
## lstat         0.126211   0.075725   1.667        0.096208 .
## medv         -0.198887   0.060516  -3.287        0.001087 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.439 on 492 degrees of freedom
## Multiple R-squared:  0.454,  Adjusted R-squared:  0.4396
## F-statistic: 31.47 on 13 and 492 DF,  p-value: < 0.00000000000000022
```

Podemos rejeitar a hipótese nula a 5% para **zn, dis, rad, black, medv

c

d