

# Exemplo de regressão linear múltipla

## Enunciado (Proveniente das aulas de Modelos Estatísticos, da Pós-Graduação em Análise de Dados e Gestão de Informação da Universidade dos Açores)

Pensa-se que a energia elétrica consumida mensalmente (\\ consumo \\) na produção de um determinado produto químico está relacionada com a temperatura média ambiental (\\ temperatura \\), o número de dias do mês (\\ dias \\), a pureza média do produto (\\ pureza \\) e o número de toneladas de produto produzidas (\\ produção \\). Dados históricos sobre estas variáveis estão disponíveis no ficheiro consumo energia.txt.

### Leitura de dados

```
setwd("C:/Users/Pedro Medeiros/Desktop/Dropbox/9999999.Pós-Graduação/07.ME/02.Exercícios")
df <- read.table("consumo_energia.txt", header = TRUE)
df
```

##	consumo	temperatura	dias	pureza	producao
## 1	240	25	24	91	100
## 2	236	31	21	90	95
## 3	270	45	24	88	110
## 4	274	60	25	87	88
## 5	301	65	25	91	94
## 6	316	72	26	94	99
## 7	300	80	25	87	97
## 8	296	84	25	86	96
## 9	267	75	24	88	110
## 10	276	60	25	91	105
## 11	288	50	25	90	100
## 12	261	38	23	89	98

### Exploração inicial

#### Nomes de variáveis

```
names(df)
```

## [1]	"consumo"	"temperatura"	"dias"	"pureza"	"producao"
--------	-----------	---------------	--------	----------	------------

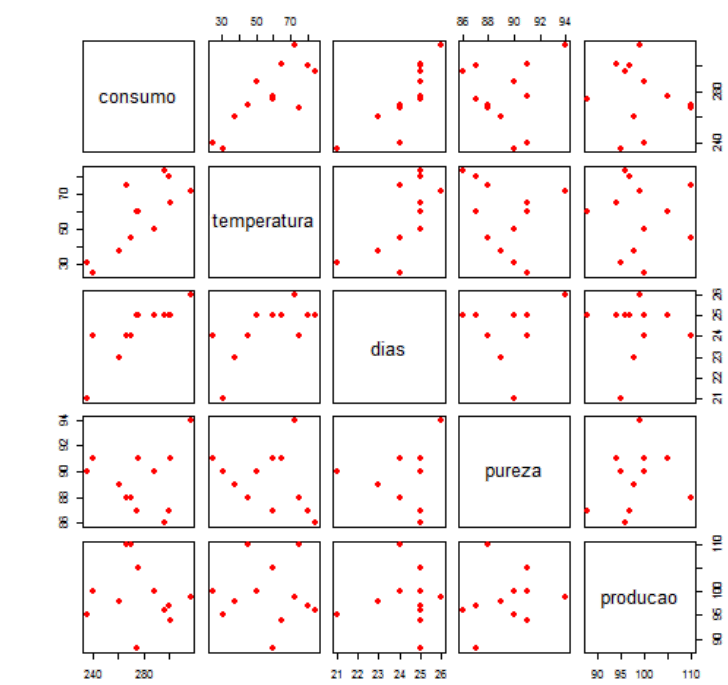
#### Sumário de variáveis

```
summary(df)
```

##	consumo	temperatura	dias	pureza
##	Min. :236	Min. :25.0	Min. :21.0	Min. :86.0
##	1st Qu.:266	1st Qu.:43.2	1st Qu.:24.0	1st Qu.:87.8
##	Median :275	Median :60.0	Median :25.0	Median :89.5
##	Mean :277	Mean :57.1	Mean :24.3	Mean :89.3
##	3rd Qu.:297	3rd Qu.:72.8	3rd Qu.:25.0	3rd Qu.:91.0
##	Max. :316	Max. :84.0	Max. :26.0	Max. :94.0
##	producao			
##	Min. : 88.0			
##	1st Qu.: 95.8			
##	Median : 98.5			
##	Mean : 99.3			
##	3rd Qu.:101.2			
##	Max. :110.0			

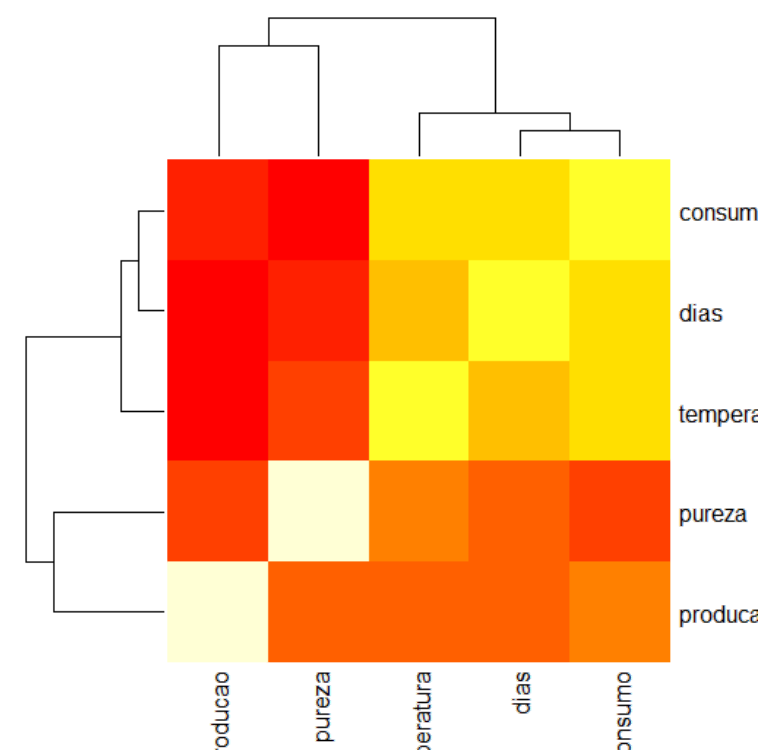
#### Scatters combinados

```
pairs(df, col = 2, pch = 19)
```



- O gráfico permite fazer as seguintes observações:
- Parece existir relação entre consumo e temperatura;
  - Parece existir relação entre consumo e número de dias; ### heatmap

```
heatmap(abs(cor(df)))
```



```
cor(df)
```

```
## consumo      consumo temperatura      dias      pureza producao
## consumo      1.00000      0.80254      0.82696      0.09285 -0.13266
## temperatura  0.80254      1.00000      0.66046 -0.28757 -0.02356
## dias         0.82696      0.66046      1.00000      0.11274 -0.02533
## pureza       0.09285     -0.28757      0.11274      1.00000      0.07891
## producao    -0.13266     -0.02356     -0.02533      0.07891      1.00000
```

O heatmap anterior confirma a maior relação entre as três variáveis.

## Regressão linear múltipla

### 3.1 Estime o modelo de regressão linear múltipla.

```
lm1 <- lm(consumo ~ temperatura + dias + pureza + producao, data = df)
```

O modelo anterior considera que todas as variáveis têm influência no consumo.

### 3.2 Teste a significância global do modelo de regressão.

Existem indícios para rejeitar a hipótese nula do teste F, de que todos os parâmetros são nulos, o que indica que a relação pode ser explicada por uma regressão linear.

```
summary(lm1)
```

```
##
## Call:
## lm(formula = consumo ~ temperatura + dias + pureza + producao,
##     data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.10  -9.78   1.77   6.80  13.02
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -123.131    157.256  -0.78    0.46
## temperatura   0.757     0.279    2.71    0.03 *
## dias          7.519     4.010    1.87    0.10
## pureza        2.483     1.809    1.37    0.21
## producao     -0.481     0.555   -0.87    0.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.8 on 7 degrees of freedom
## Multiple R-squared:  0.852, Adjusted R-squared:  0.768
## F-statistic: 10.1 on 4 and 7 DF, p-value: 0.00496
```

### 3.3. Identifique os parâmetros que diferem de zero.

Apenas existem indícios para rejeitar a hipótese de parâmetro nulo para a variável temperatura, para um nível de significância de 5% ( $\alpha = 0.05$ )  
As restantes variáveis não parecem ter efeito sobre o consumo.

```
summary(lm1)
```

```
##  
## Call:  
## lm(formula = consumo ~ temperatura + dias + pureza + producao,  
##     data = df)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -14.10   -9.78    1.77    6.80   13.02   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept) -123.131    157.256  -0.78    0.46      
## temperatura  0.757      0.279   2.71    0.03 *    
## dias         7.519      4.010   1.87    0.10      
## pureza       2.483      1.809   1.37    0.21      
## producao    -0.481      0.555  -0.87    0.41      
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 11.8 on 7 degrees of freedom  
## Multiple R-squared:  0.852, Adjusted R-squared:  0.768   
## F-statistic: 10.1 on 4 and 7 DF,  p-value: 0.00496
```

### 3.4. Interprete as estimativas dos parâmetros estatisticamente significativos.

O único parâmetros estatisticamente significativo é a temperatura.  
Interpretação: Um aumento de 1 grau na temperatura média conduz a um aumento de 0.752 unidades de consumo eléctrico.

### 3.5. Indique a variação total da energia consumida mensalmente que é explicada pelo modelo de regressão.

A variação total de energia explicada pelo modelo é de 0.852.

### 3.6. Determine os ICs a 95% para os parâmetros do modelo.

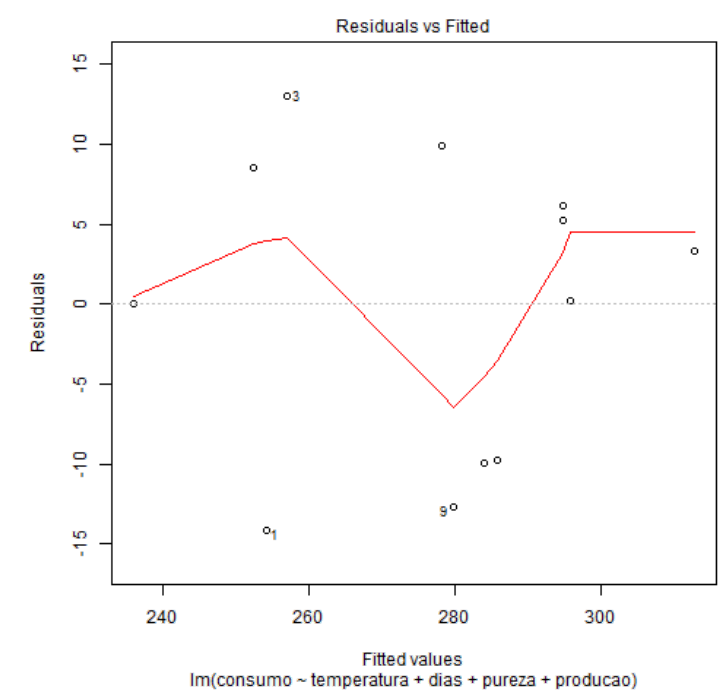
```
confint(lm1)  
  
##              2.5 %    97.5 %  
## (Intercept) -494.98273 248.7202  
## temperatura  0.09735   1.4172  
## dias        -1.96365  17.0012  
## pureza      -1.79544   6.7616  
## producao    -1.79391   0.8316
```

### 3.7 Proceda à análise de resíduos por forma a validar os pressupostos do modelo.

#### Distribuição dos resíduos

A variação dos resíduos aparenta diminuir para os valores mais altos. No entanto existem poucos dados.

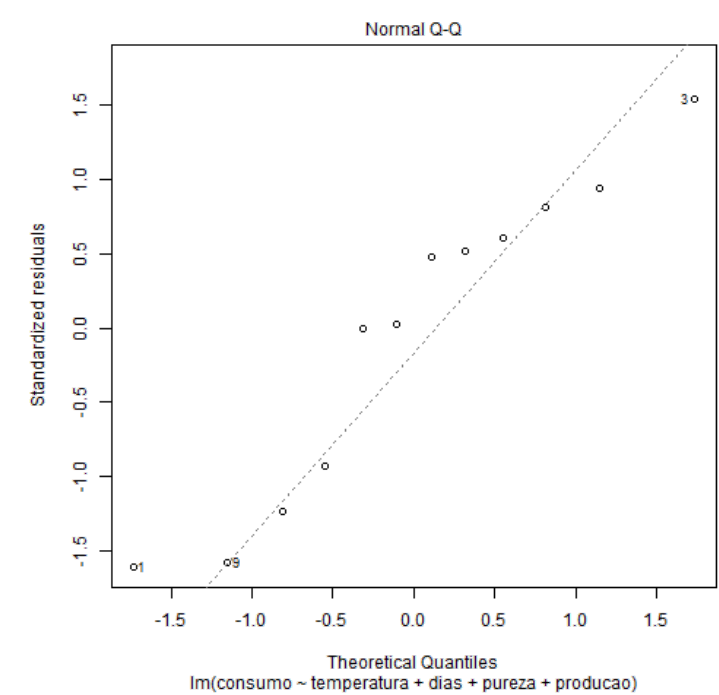
```
plot(lm1, which = 1)
```



Normalidade dos resíduos

O teste de Shapiro não indicia a rejeição da hipótese nula, de normalidade dos resíduos.  
O gráfico qqplot apresenta alguns desvios.

```
plot(lm1, which = 2)
```



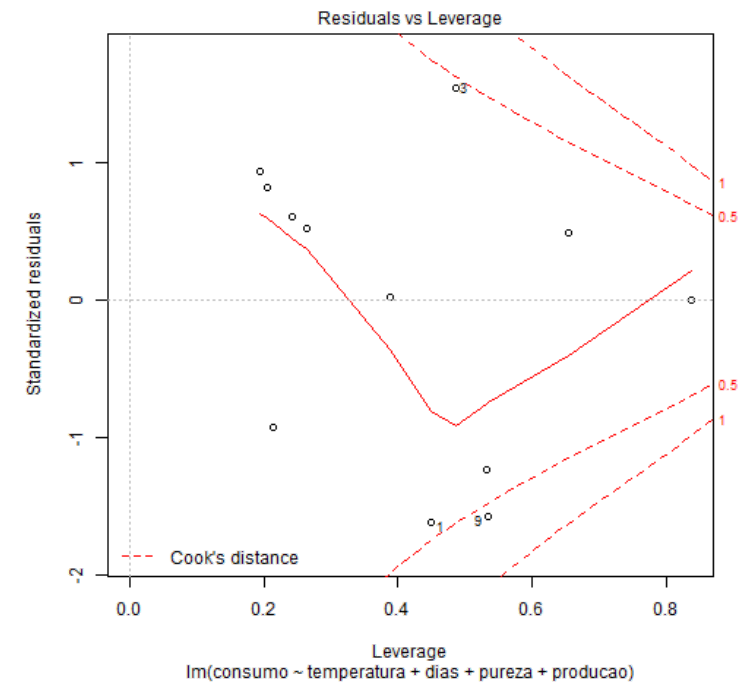
```
shapiro.test(lm1$residuals)
```

```
##
##  Shapiro-wilk normality test
##
## data:  lm1$residuals
## W = 0.915, p-value = 0.2469
```

### Deteção de valores alvanca e significativos

A observação 9 tem distância de Cook superior a 0.5. Existe uma observação com hat value próximo do valor do máximo (hat\_thresh).

```
plot(lm1, which = 5)
```



```
hat_thresh <- 2 * ((dim(df)[2]))/dim(df)[1]
which(hatvalues(lm1) > hat_thresh)
```

```
## 2
## 2
```

### Outliers

Não foi detetado nenhum outlier

```
which(rstudent(lm1) > 2)
```

```
## named integer(0)
```

### Multicolinearidade

Não foram detetados valores superiores a 5, que indiciem associação muito forte entre variáveis explicativas.

```
library(car)
```

```
## warning: package 'car' was built under R version 3.0.3
```

```
vif(lm1)

## temperatura      dias      pureza      producao
##      2.323      2.161      1.335      1.009
```

3.8. Determine uma estimativa para o consumo médio de energia quando a temperatura média ambiental é 75°F, o número de dias do mês é 24, a pureza média do produto é 90 e o número de toneladas de produto produzido é 98. Obtenha um IC a 95% para o valor

Interval = “confidence”, porque quero estimar o consumo médio da população e não o consumo da população (interval = “predict”).

```
predict(lm1, list(temperatura = 75, dias = 24, pureza = 90, producao = 98),
        interval = "conf")

##      fit      lwr      upr
## 1 290.4 272.5 308.4
```

Determinação de modelos mais simples, com representatividade semelhante

Filtragem automática pelos métodos “stepwise”, “backward” e “forward” e comparação de resultados.

Método *Stepwise*

O método indica que a variável producao poderá ser retirada do modelo sem perda de qualidade.

```
step(lm1, direction = "both")

## Start:  AIC=62.74
## consumo ~ temperatura + dias + pureza + producao
##
##              Df Sum of Sq  RSS  AIC
## - producao    1      104 1077 62.0
## <none>         1      972 62.7
## - pureza      1       262 1234 63.6
## - dias        1       488 1461 65.6
## - temperatura 1      1023 1995 69.4
##
## Step:  AIC=61.96
## consumo ~ temperatura + dias + pureza
##
##              Df Sum of Sq  RSS  AIC
## <none>         1      1077 62.0
## - pureza      1       235 1312 62.3
## + producao    1       104  972 62.7
## - dias        1       512 1589 64.6
## - temperatura 1      1001 2078 67.9

##
## Call:
## lm(formula = consumo ~ temperatura + dias + pureza, data = df)
##
## Coefficients:
## (Intercept) temperatura      dias      pureza
##    -162.135      0.749      7.691      2.343
```

Método *backward*

As conclusões são semelhantes ao método *stepwise*

```
step(lm1, direction = "backward")
```

```
## Start:  AIC=62.74
## consumo ~ temperatura + dias + pureza + producao
##
##           Df Sum of Sq  RSS  AIC
## - producao    1      104 1077 62.0
## <none>         1      972 62.7
## - pureza       1      262 1234 63.6
## - dias         1      488 1461 65.6
## - temperatura  1      1023 1995 69.4
##
## Step:  AIC=61.96
## consumo ~ temperatura + dias + pureza
##
##           Df Sum of Sq  RSS  AIC
## <none>         1      1077 62.0
## - pureza       1      235 1312 62.3
## - dias         1      512 1589 64.6
## - temperatura  1      1001 2078 67.9
```

```
##
## Call:
## lm(formula = consumo ~ temperatura + dias + pureza, data = df)
##
## Coefficients:
## (Intercept)  temperatura          dias          pureza
##      -162.135         0.749         7.691         2.343
```

Método *forward*

O método *forward* indica que se devem manter todas as variáveis.

```
step(lm1, direction = "forward")
```

```
## Start:  AIC=62.74
## consumo ~ temperatura + dias + pureza + producao
```

```
##
## Call:
## lm(formula = consumo ~ temperatura + dias + pureza + producao,
##     data = df)
##
## Coefficients:
## (Intercept)  temperatura          dias          pureza          producao
##      -123.131         0.757         7.519         2.483        -0.481
```

Teste F para comparar a qualidade dos modelos com e sem a variável produção.

Criação de modelo atualizado, sem a variável produção.  
Os parâmetros das variáveis \ ( dias \ ) e \ ( pureza \ ) continuam a não ter significado estatístico, pelo que se considera que deveriam ser retiradas da análise num caso real.

```
lm2 <- update(lm1, ~. - producao)
summary(lm2)
```



```
##
## Call:
## lm(formula = consumo ~ temperatura + dias + pureza, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.81  -6.77   3.26   7.98   9.53
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -162.135    148.315  -1.09   0.306
## temperatura   0.749     0.275    2.73   0.026 *
## dias         7.691     3.942    1.95   0.087 .
## pureza        2.343     1.774    1.32   0.223
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.6 on 8 degrees of freedom
## Multiple R-squared:  0.836, Adjusted R-squared:  0.775
## F-statistic: 13.6 on 3 and 8 DF, p-value: 0.00165
```

A comparação dos modelos indica que não existem indícios para rejeitar a hipótese nula de igualdade de qualidade dos modelos.  
**Os modelos são semelhantes** escolhendo-se, portanto, o modelo mais simples, pelo princípio da parcimónia.

```
anova(lm2, lm1)
```

```
## Analysis of Variance Table
##
## Model 1: consumo ~ temperatura + dias + pureza
## Model 2: consumo ~ temperatura + dias + pureza + producao
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1      8 1077
## 2      7  972  1      104 0.75  0.41
```

Teste F, retirando todas as variáveis, com exceção de temperatura

O teste ainda permite concluir que existem evidências para considerar os modelos equivalentes, apesar da redução do valor de  $R^2$ .

```
lm3 <- update(lm1, ~. - producao - pureza - dias)
summary(lm3)
```

```
##
## Call:
## lm(formula = consumo ~ temperatura, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -28.20  -6.60  -2.14    7.83   23.84
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  219.380     14.265   15.38  2.8e-08 ***
## temperatura   1.011     0.238    4.25  0.0017 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.3 on 10 degrees of freedom
## Multiple R-squared:  0.644, Adjusted R-squared:  0.608
## F-statistic: 18.1 on 1 and 10 DF, p-value: 0.00168
```

```
anova(lm3, lm1)
```

```
## Analysis of Variance Table
##
## Model 1: consumo ~ temperatura
## Model 2: consumo ~ temperatura + dias + pureza + producao
##   Res.Df  RSS Df Sum of Sq    F Pr(>F)
## 1     10 2340
## 2      7  972  3      1367 3.28  0.089 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

