Iniciação à Estatística

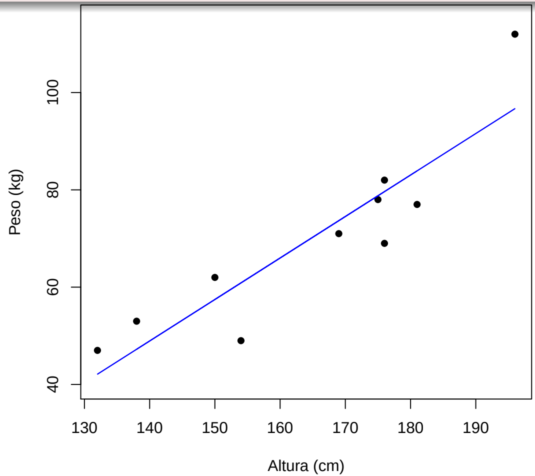
Prof. Fernando de Souza Bastos
fernando.bastos@ufv.br

Departamento de Estatística
Universidade Federal de Viçosa
Campus UFV - Viçosa

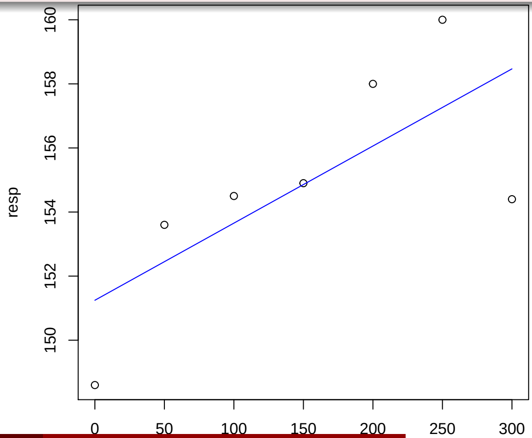
Sumário

Em certas situações podemos estar interessados em descrever a relação entre duas variáveis, e também predizer o valor de uma a partir de outra. Por exemplo, se sabemos a altura de um certo estudante, mas não o seu peso, qual seria um bom chute para o peso deste estudante?

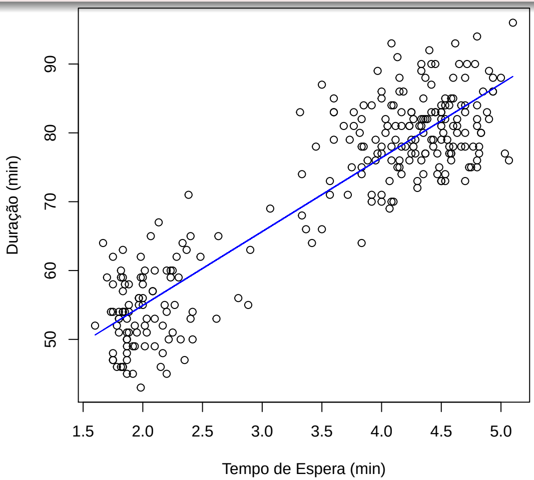
```r
x = c(176, 154, 138, 196, 132, 176, 181, 169, 150, 175)
y = c(82, 49, 53, 112, 47, 69, 77, 71, 62, 78)
plot(x,y,xlab="Altura (cm)",ylab="Peso (kg)",
      pch=16, ylim = c(40,115))
lines(x, fitted(lm(y ~ x)), col="blue")
```



```
doses <- c(0, 50, 100, 150, 200, 250, 300)
resp <- c(148.6, 153.6, 154.5, 154.9, 158, 160, 154.4)
reglin <- lm(resp ~ doses)
plot(doses, resp) #(variável indep. primeiro)
lines(doses, fitted(reglin), col="blue")#acrescenta
a reta de regressão ajustada
```



```
#Tempo de espera entre erupções e a duração da erupção
fit <- lm(waiting~eruptions, data=faithful)
plot(faithful)
lines(faithful$eruptions, fitted(fit), col="blue")
```



Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

(i) Altura dos pais e altura dos filhos;

Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

(i) Altura dos pais e altura dos filhos;

(ii) Renda semanal e despensas de consumo;

Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

(i) Altura dos pais e altura dos filhos;
(ii) Renda semanal e despensas de consumo;
(iii) Variação dos salários e taxa de desemprego;

Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

(i) Altura dos pais e altura dos filhos;
(ii) Renda semanal e despensas de consumo;
(iii) Variação dos salários e taxa de desemprego;
(iv) Demanda dos produtos de uma firma e publicidade;

Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

(i) Altura dos pais e altura dos filhos;
(ii) Renda semanal e despensas de consumo;
(iii) Variação dos salários e taxa de desemprego;
(iv) Demanda dos produtos de uma firma e publicidade;

Sob dois pontos de vista:

(i) Explicitando a forma dessa relação: regressão.

Objetivos:

Estudar a relação linear entre duas variáveis quantitativas. Exemplos:

(i) Altura dos pais e altura dos filhos;
(ii) Renda semanal e despensas de consumo;
(iii) Variação dos salários e taxa de desemprego;
(iv) Demanda dos produtos de uma firma e publicidade;

Sob dois pontos de vista:

(i) Explicitando a forma dessa relação: regressão.
(ii) Quantificando a força dessa relação: correlação.

Importante:

Uma relação estatística por sí propria não implica uma causa, para atribuir causa, devemos invocar alguma teoría!

Importante:

Uma relação estatística por sí propria não implica uma causa, para atribuir causa, devemos invocar alguma teoría!

Uma regressão espúria é uma relação estatística existente entre duas variáveis, mas onde não existe nenhuma relação causa-efeito entre elas. Essa relação estatística pode ocorrer por pura coincidência ou por causa de uma terceira variável. Ou seja, neste último caso, pode ocorrer que as variáveis X e Y sejam correlacionadas porque ambas são causadas por uma terceira variável Z.

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). Determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer cinco situações:

(i) (A) causa realmente (B);

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). Determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer cinco situações:

1. (A) causa realmente (B);
2. (B) pode ser a causa de (A);

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). Determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer cinco situações:

(i) (A) causa realmente (B);
(ii) (B) pode ser a causa de (A);
(iii) Um terceiro factor (C) pode ser causa tanto de (A) como de (B);

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). Determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer cinco situações:

(i) (A) causa realmente (B);
(ii) (B) pode ser a causa de (A);
(iii) Um terceiro factor (C) pode ser causa tanto de (A) como de (B);Pode ser uma combinação das três situações anteriores. Por exemplo, (A) causa (B) e ao mesmo tempo (B) causa também (A);

Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). Determinar se existe de fato uma relação de causalidade requer investigação adicional pois podem acontecer cinco situações:

(i) (A) causa realmente (B);
(ii) (B) pode ser a causa de (A);
(iii) Um terceiro factor (C) pode ser causa tanto de (A) como de (B);Pode ser uma combinação das três situações anteriores. Por exemplo, (A) causa (B) e ao mesmo tempo (B) causa também (A);
(iv) A correlação pode ser apenas uma coincidência, ou seja, os dois eventos não têm qualquer relação para além do fato de ocorrerem ao mesmo tempo.

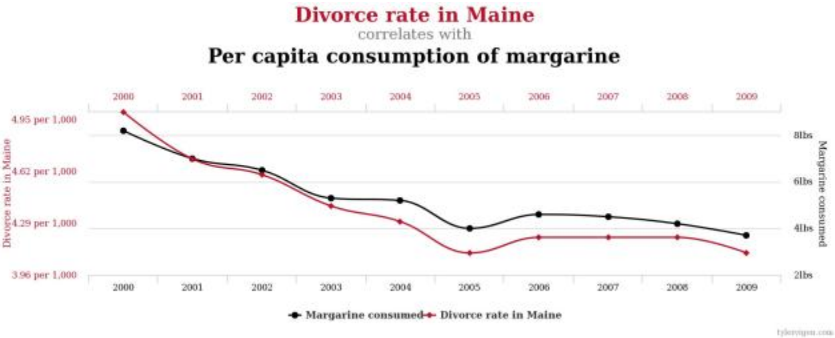
Exemplos:

"Quanto maiores são os pés de uma criança, maior a capacidade para resolver problemas de matemática. Portanto, ter pés grandes faz ter melhores notas em matemática".

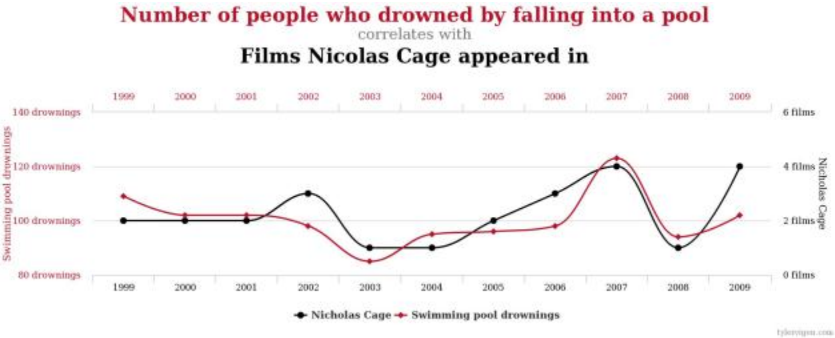
Exemplos:

“Vários estudos apontavam inicialmente que as mulheres em menopausa que recebiam terapia de substituição hormonal (TSH) tinham também um menor risco de doença coronária, o que levou à ideia de que a TSH conferia protecção contra a doença coronária. No entanto, estudos controlados e randomizados (mais rigorosos), feitos posteriormente, mostraram que a TSH causava na verdade um pequeno mas significativo aumento do risco de doença coronária. Uma reanálise dos estudos revelou que as mulheres que recebiam a TSH tinham também uma maior probabilidade de pertencer a uma classe socioeconómica superior, com melhor dieta e hábitos de exercício. A utilização da TSH e a baixa incidência de doença coronária não eram causa e efeito, mas o fruto de uma causa comum".

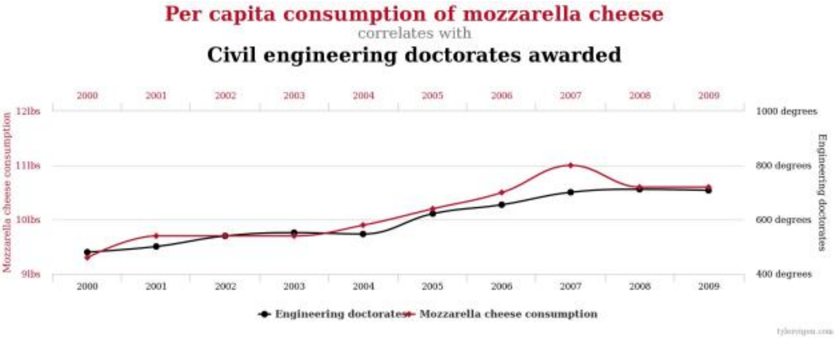
“Quanto menos as pessoas se divorciam em Maine (EUA), menor fica o consumo de margarina naquele Estado".



Deveríamos banir Nicolas Cage do cinema para evitar o afogamento de pessoas? O primeiro gráfico nos dá o número de pessoas afogadas (linha vermelha) e as aparições do Nicolas Cage em filmes (linha preta).



Consumo de muçarela (linha vermelha) e doutorados obtidos em engenharia civil (linha preta)



De forma geral, um modelo estatístico pode ser escrito da seguinte forma:

$$Y = \text{componente determinística} + \text{componente aleatória}$$

existem diversas maneiras de específicar essas componentes. Começaremos com uma regressão linear simples.

Uma regresão linear simples tem como objetivo aproximar uma variável resposta Y através de uma função linear de uma variável de interesse, ou seja,

$$Y = f(X, \beta) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$$

No qual assume-se que:

(i) $E(\varepsilon) = 0$

Uma regresão linear simples tem como objetivo aproximar uma variável resposta Y através de uma função linear de uma variável de interesse, ou seja,

$$Y = f(X, \beta) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$$

No qual assume-se que:

(i) $E(\varepsilon) = 0$
(ii) $V(\varepsilon) = \sigma^2$ (Homocedásticidade)

Uma regresão linear simples tem como objetivo aproximar uma variável resposta Y através de uma função linear de uma variável de interesse, ou seja,

$$Y = f(X, \beta) + \varepsilon = \beta_0 + \beta_1 X + \varepsilon$$

No qual assume-se que:

(i) $E(\varepsilon) = 0$
(ii) $V(\varepsilon) = \sigma^2$ (Homocedásticidade)
(iii) $Cov(\varepsilon_i, \varepsilon_j) = 0$

em outras palavras, os erros tem média zero, variância constante e são não correlacionados.

A variável preditora X pode vir de diversas fontes:

- inputs quantitativos (valores reais, medidas)
- tranformação de variável quantitativas (log() ,$\sqrt{()}$,etc)
- inputs qualitativos ("dummy"e.x. genêro, classes)

Dessa forma, um modelos de regressão consiste em 4 passos:

1. Escolher o componente determinístico do modelo;

A variável preditora X pode vir de diversas fontes:

- inputs quantitativos (valores reais, medidas)
- tranformação de variável quantitativas ($\log()$,$\sqrt{()}$,etc)
- inputs qualitativos ("dummy"e.x. genêro, classes)

Dessa forma, um modelos de regressão consiste em 4 passos:

1. Escolher o componente determinístico do modelo;
2. Especificar a distribuição do erro;

A variável preditora X pode vir de diversas fontes:

- inputs quantitativos (valores reais, medidas)
- tranformação de variável quantitativas (log() ,$\sqrt{()}$,etc)
- inputs qualitativos (“dummy"e.x. genêro, classes)

Dessa forma, um modelos de regressão consiste em 4 passos:

1. Escolher o componente determinístico do modelo;
2. Especificar a distribuição do erro;
3. Utilizar os dados para estimar os parâmetros do modelo;

A variável preditora X pode vir de diversas fontes:

- inputs quantitativos (valores reais, medidas)
- tranformação de variável quantitativas (log() ,$\sqrt{()}$,etc)
- inputs qualitativos (“dummy"e.x. genêro, classes)

Dessa forma, um modelos de regressão consiste em 4 passos:

1. Escolher o componente determinístico do modelo;
2. Especificar a distribuição do erro;
3. Utilizar os dados para estimar os parâmetros do modelo;
4. Avaliar o modelo estatístico;

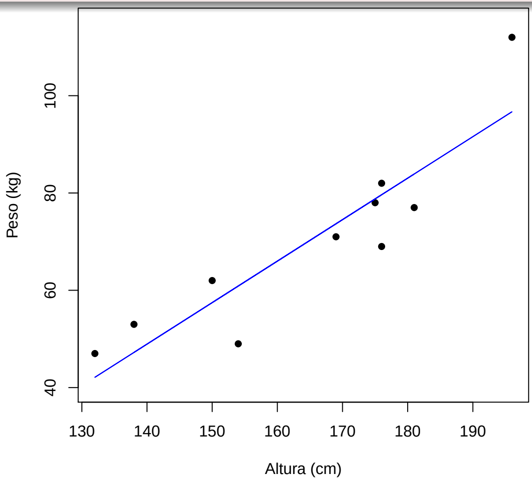
Os dados para a análise de regressão e correlação simples são da forma:

$$(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$$

Com base nos dados constrói-se o diagrama de dispersão, que deve exibir uma tendência linear para que se possa usar a regressão linear. Este diagrama permite decidir empiricamente:

- Se um relacionamento linear entre as variáveis X e Y deve ser assumido;
- Se o grau de relacionamento linear entre as variáveis é forte ou fraco, conforme o modo como se situam os pontos em redor de uma reta imaginária que passa através do enxame de pontos.

```
#Encontre o modelo de Regressão Linear que melhor se
ajusta aos dados
x = c(176, 154, 138, 196, 132, 176, 181, 169, 150, 175)
y = c(82, 49, 53, 112, 47, 69, 77, 71, 62, 78)
```



```
#Encontre o modelo de Regressão Linear que melhor se
ajusta aos dados
x <-  c(176, 154, 138, 196, 132, 176, 181, 169, 150, 175)
y <-  c( 82,  49,  53, 112,  47,  69,  77,  71,  62,  78)
Reg <- lm(y~x)
Reg

Call:
lm(formula = y ~ x)


Coefficients:
(Intercept)            x
   -70.4627       0.8528
```

```
#library(texreg)
summary(Reg)

lm(formula = y ~ x)
Residuals:
     Min       1Q   Median       3Q      Max
-11.8746  -5.8428   0.7893   4.8001  15.3061
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -70.4627    24.0148  -2.934 0.018878 *
x             0.8528     0.1448   5.889 0.000366 ***
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
Residual standard error: 8.854 on 8 degrees of freedom
Multiple R-squared:  0.8126,Adjusted R-squared:  0.7891
F-statistic: 34.68 on 1 and 8 DF,  p-value: 0.0003662
```

	Model 1
(Intercept)	-70.46^{*}
	(24.01)
x	0.85^{***}
	(0.14)
R^2	0.81
Adj. R^2	0.79
Num. obs.	10
RMSE	8.85

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Tabela: Statistical models

Raiz Quadrada do Erro Quadrático Médio

ROOT MEAN SQUARE ERROR (RMSE)

A medida de erro mais comumente usada para aferir a qualidade do ajuste de um modelo é a chamada RAIZ DO ERRO MÉDIO QUADRÁTICO. Ela é a raiz do erro médio quadrático da diferença entre a predição e o valor real. Podemos pensar nela como sendo uma medida análoga ao desvio padrão.

R^2

Representa a porcentagem de variação na resposta que é explicada pelo modelo. Ele é calculado como 1 menos a razão da soma dos quadrados dos erros (que é a variação que não é explicada pelo modelo) pela soma total dos quadrados (que é a variação total no modelo).

Use R^2 para determinar se o modelo ajusta bem os dados. Quanto mais alto o valor de R^2 melhor o modelo ajusta seus dados. O valor de R^2 está sempre entre 0 e 100

R^2

Representa a porcentagem de variação na resposta que é explicada pelo modelo. Ele é calculado como 1 menos a razão da soma dos quadrados dos erros (que é a variação que não é explicada pelo modelo) pela soma total dos quadrados (que é a variação total no modelo).
Use R^2 para determinar se o modelo ajusta bem os dados. Quanto mais alto o valor de R^2 melhor o modelo ajusta seus dados. O valor de R^2 está sempre entre 0 e 100

R^2

Use R^2 para determinar se o modelo se ajusta bem aos dados. Quanto mais alto o valor de R^2 melhor o modelo ajusta seus dados. O valor de R^2 está sempre entre 0 e 100%.

Considere as seguintes questões ao interpretar o valor de R^2 :

O R^2 sempre aumenta quando você adiciona mais preditores a um modelo. Por exemplo, o melhor modelo de cinco preditores terá sempre um R^2 que é pelo menos tão elevado quanto o melhor modelo de quatro preditores. Portanto, R^2 é mais útil quando for comparado a modelos do mesmo tamanho.

Considere as seguintes questões ao interpretar o valor de R^2 :

O R^2 sempre aumenta quando você adiciona mais preditores a um modelo. Por exemplo, o melhor modelo de cinco preditores terá sempre um R^2 que é pelo menos tão elevado quanto o melhor modelo de quatro preditores. Portanto, R^2 é mais útil quando for comparado a modelos do mesmo tamanho.

Amostras pequenas não fornecem uma estimativa precisa da força da relação entre a resposta e os preditores. Se você precisar que R^2 seja mais exato, deve usar uma amostra maior (geralmente, 40 ou mais).

R^2 é apenas uma medida de o quão bem o modelo ajusta os dados. Mesmo quando um modelo tem um R^2 elevado, você deve verificar os gráficos de resíduos para conferir se o modelo satisfaz os pressupostos do modelo.

R^2 Ajustado

O R^2 ajustado é a porcentagem de variação na resposta que é explicada pelo modelo, ajustada para o número de preditores do modelo em relação ao número de observações.

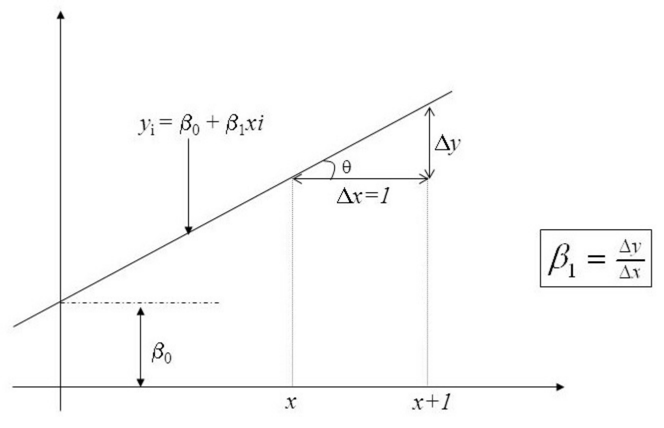
R^2 Ajustado

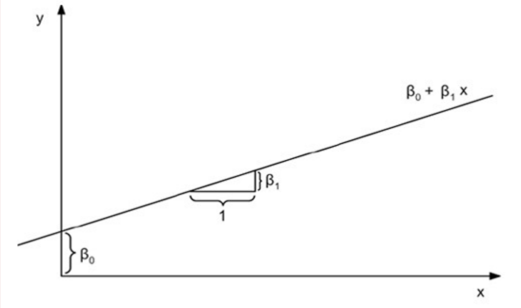
O R^2 ajustado é a porcentagem de variação na resposta que é explicada pelo modelo, ajustada para o número de preditores do modelo em relação ao número de observações.

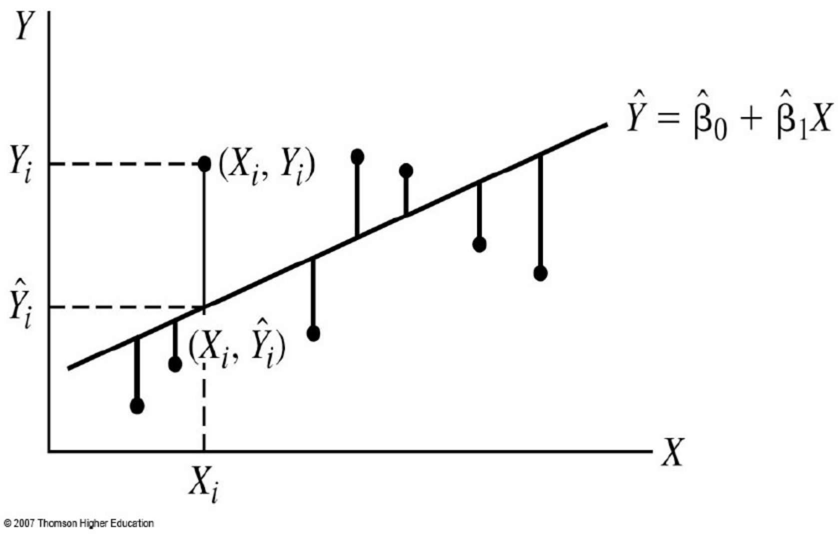
Interpretação:

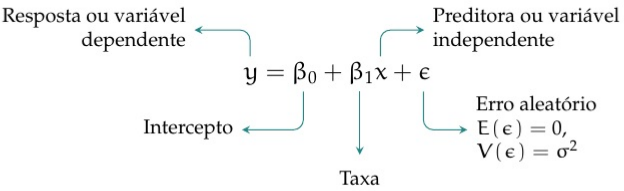
Use o R^2 ajustado quando desejar comparar modelos que têm diferentes números de preditores. R^2 sempre aumenta quando você adiciona um preditor ao modelo, mesmo quando não existe uma verdadeira melhoria ao modelo. O valor de R^2 ajustado incorpora o número de preditores no modelo para ajudá-lo a escolher o modelo correto.

O parâmetro β_0 é chamado intercepto ou coeficiente linear e representa o ponto em que a reta regressora corta o eixo dos $y's$, quando $x = 0$. Já o parâmetro β_1 representa a inclinação da reta regressora e é dito coeficiente de regressão ou coeficiente angular. Além disso, temos que para um aumento de uma unidade na variável x, o valor $E(Y|x)$ aumenta β_1 unidades. A interpretação geométrica dos parâmetros β_0 e β_1 pode ser vista na próxima Figura.


$y_i = \beta_0 + \beta_1 xi$
Δy
θ
$\Delta x = 1$
$\beta_1 = \frac{\Delta y}{\Delta x}$
β_0
x
$x+1$


y
$\beta_0 + \beta_1 x$
β_1
1
β_0
x


Y
Y_i
$\hat{Y}_i$
(X_i, Y_i)
$(X_i, \hat{Y}_i)$
$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$
X
X_i
© 2007 Thomson Higher Education


Resposta ou variável dependente
Preditora ou variável independente
$y = \beta_0 + \beta_1 x + \epsilon$
Intercepto
Taxa
Erro aleatório
$E(\epsilon) = 0$,
$V(\epsilon) = \sigma^2$

LINEARIDADE

(o modelo linear descreve corretamente a relação funcional entre X e Y) Se esse pressuposto for violado a estimativa do erro aumentará, já que os valores observados não se aproximarão dos valores preditos (local onde passará a reta). Pressuposto fundamental já que essa regressão é um modelo linear.

NORMALIDADE

Normalidade dos resíduos é esperada para que não existam tendências e que a estatística F funcione de forma correta.

VARIÂNCIAS HOMOGÊNEAS

As variâncias dentro de cada grupo é igual (ou pelo menos aproximadamente) àquela dentro de todos os grupos. Desta forma, cada tratamento contribui de forma igual para a soma dos quadrados.

Se os pressupostos forem atendidos fica mais fácil afirmar que os resultados da análise são devido aos efeitos testados. Além disso, a confiabilidade do teste aumenta, já que se terá certeza que não há tendências nos resultados.

Iniciemos com um exemplo. Um investigador deseja estudar a possível relação entre o salário (em mil reais) e o tempo de experiência (em anos completos) no cargo de gerente de agências bancárias de uma grande empresa. Os dados coletados são lidos no R:

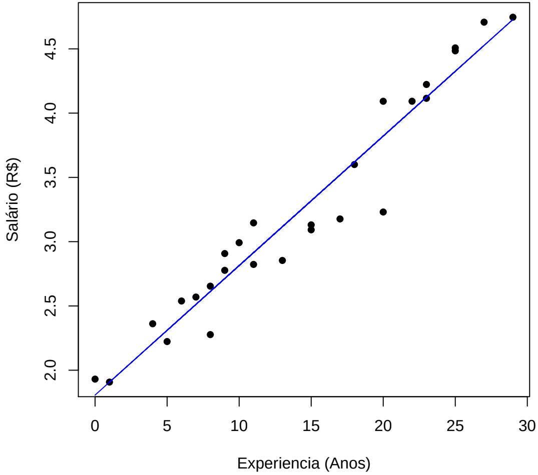
```
dados <- read.table("Exp_Salario.txt",
                     sep = "", dec = ".", header = TRUE)
names(dados) <- c("X","Y")
```

Iniciemos com um exemplo. Um investigador deseja estudar a possível relação entre o salário (em mil reais) e o tempo de experiência (em anos completos) no cargo de gerente de agências bancárias de uma grande empresa. Os dados coletados são lidos no R:

```
dados <- read.table("Exp_Salario.txt",
                    sep = "", dec = ".", header = TRUE)
names(dados) <- c("X","Y")
```

são considerados 27 pares de observações correspondentes à variável resposta Salário e à variável explicativa Experiência, para cada um dos gerentes da empresa.

```r
plot(X,Y,xlab="Experiencia (Anos)",ylab="Salário (R$)",
      pch=16)
lines(X, fitted(lm(Y ~ X)), col="blue")
```



```
cor(X,Y)
```

```
[1] 0.9704137
```

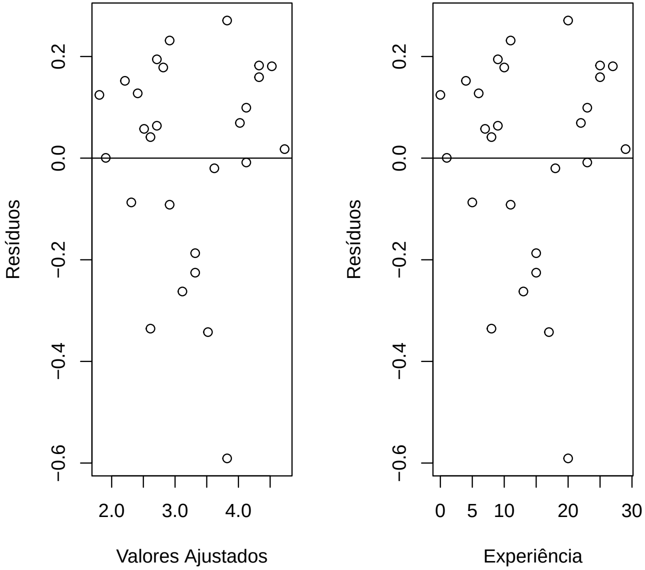
Observe que o R retornou o valor 0.9704137 o que evidencia uma forte relação linear entre as variáveis em estudo. Para avaliar se esse resultado é significativo, pode-se realizar um Teste de Hipóteses para o Coeficiente de Correlação.

```
cor.test(X,Y)
```

```
Pearson's product-moment correlation
t = 20.096, df = 25, p-value < 2.2e-16
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9353175 0.9865989
```

Para avaliar as suposições de que os erros possuem variância constante e são não correlacionados entre si, construa os gráficos de "Resíduos versus Valores Ajustados da Variável Resposta" e "Resíduos versus Valores da Variável Explicativa".

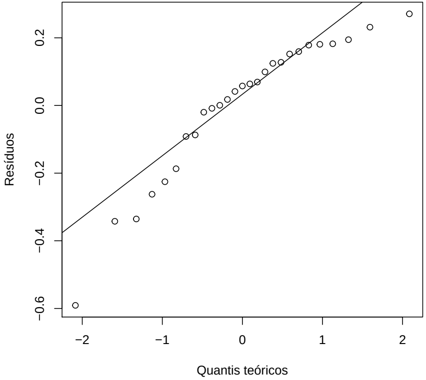
```r
m0 <- lm(Y~X)
plot(fitted(m0),residuals(m0),xlab="Valores Ajustados",
      ylab="Resíduos")
abline(h=0)
plot(X,residuals(m0),xlab="Experiência",ylab="Resíduos")
abline(h=0)
par(mfrow=c(1,2))
```



observa-se a violação da suposição de homocedasticidade dos erros.

observa-se ainda a violação da suposição de que os erros aleatórios têm distribuição Normal. Via gráfico qqplot abaixo:

```
qqnorm(residuals(m0), ylab="Resíduos",
       xlab="Quantis teóricos",main="")
qqline(residuals(m0))
```

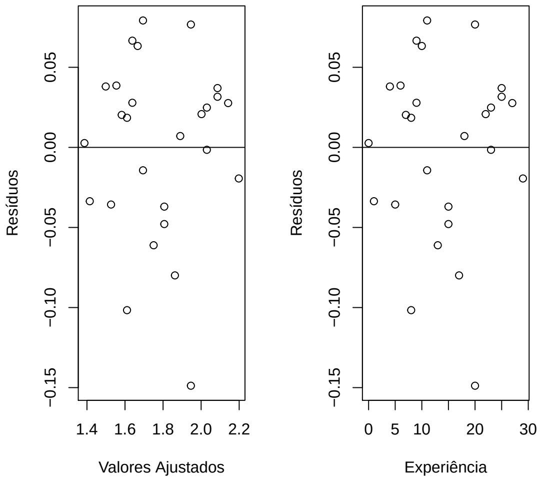


```
shapiro.test(residuals(m0))

Shapiro-Wilk normality test


data:  residuals(m0)
W = 0.9012, p-value = 0.01425
```

Calculamos $\sqrt{Y}$ e reaplicamos o modelo linear



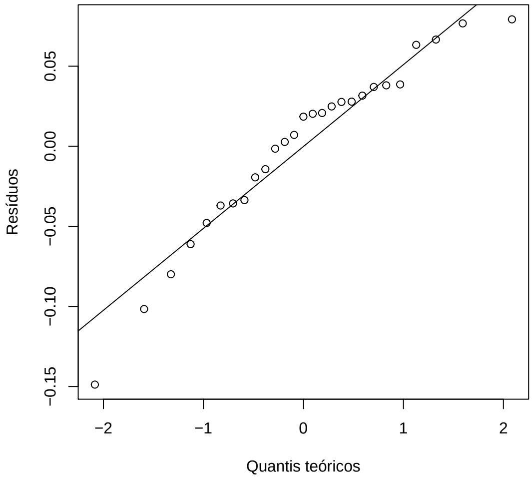
Calculamos $\sqrt{Y}$ e reaplicamos o modelo linear

observa-se a não violação da suposição de homocedasticidade dos erros.

```
Shapiro-Wilk normality test


data:  residuals(m1)
W = 0.94139, p-value = 0.1319
```

```r
qqnorm(residuals(m1), ylab="Resíduos",
        xlab="Quantis teóricos",main="")
qqline(residuals(m1))
```



Calculamos $\sqrt{Y}$ e reaplicamos o modelo linear

```
dados <- read.table("Exp_Salario.txt",
                     sep = "", dec = ".", header = TRUE)
names(dados) <- c("X","Y")
attach(dados)
m1 <- lm(sqrt(Y)~X)
m1

Call:
lm(formula = sqrt(Y) ~ X)

Coefficients:
(Intercept)            X
    1.38680      0.02797
```

Calculamos $\sqrt{Y}$ e reaplicamos o modelo linear

```
summary(m1)

lm(formula = sqrt(Y) ~ X)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.38680    0.02165   64.06   <2e-16 ***
X            0.02797    0.00133   21.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'
Residual standard error: 0.05606 on 25 degrees of freedom
Multiple R-squared:  0.9465,Adjusted R-squared:  0.9444
F-statistic: 442.3 on 1 and 25 DF,  p-value: < 2.2e-16
```