Modelos Lineares

Prof. Steven Dutt Ross

Atividades
Aulas
Livro

Suponha que blablabla

Vamos carregar a base de dados. isso pode ser feito com o pacote *readxl*

```
library(readxl)
initech <- read_excel("~/GitHub/aulas/regressao/codigo/dados/initech.xlsx")
```

Modelo linear

Vamos ajustar um modelo linear.

Modelo de regressão linear

vamos definir um modelo de regressão linear,

$$Y = f(X) + \epsilon.$$

dessa forma temos:

$$Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

onde $\epsilon_i \sim N(0, \sigma^2)$.

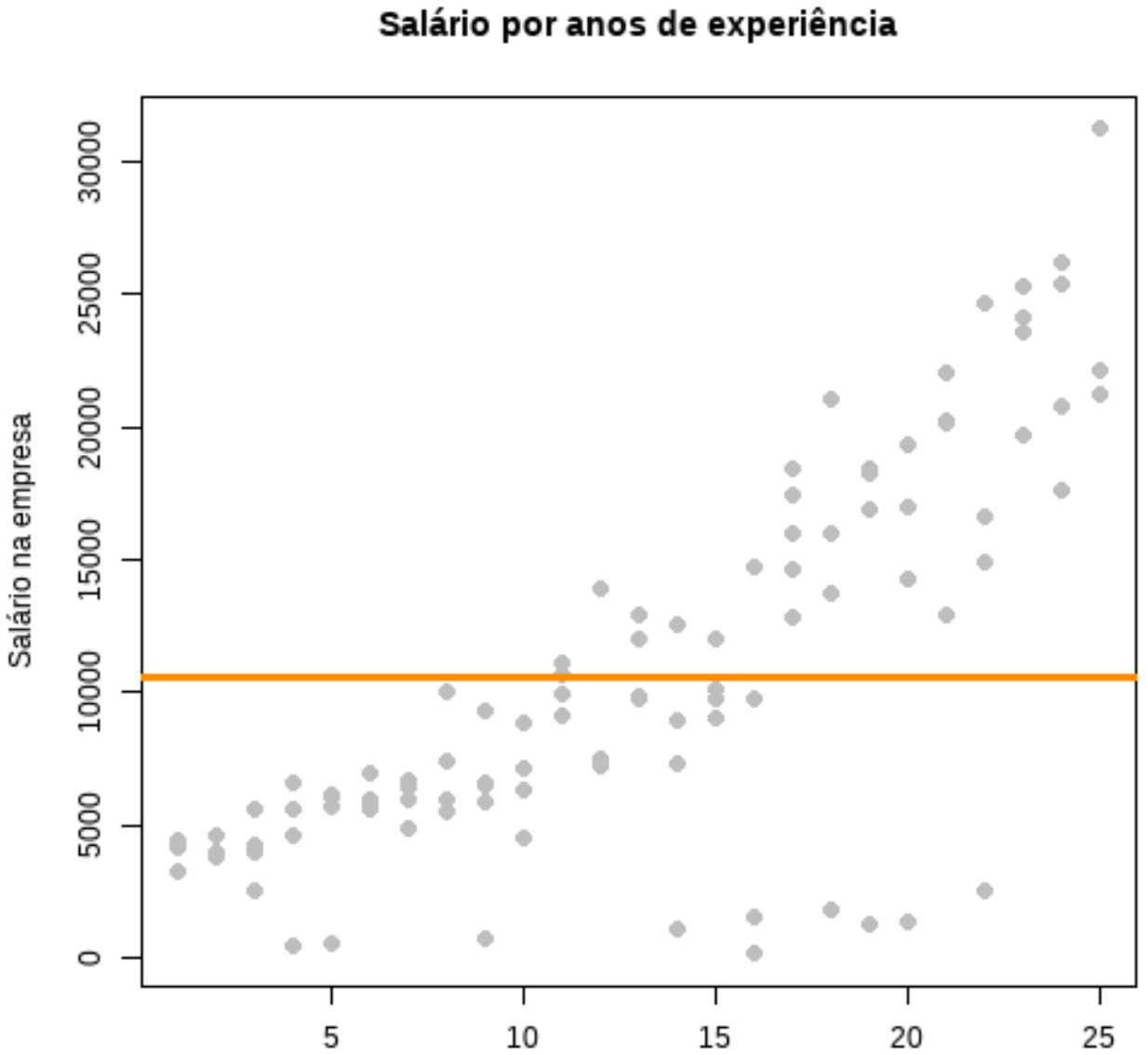
Modelo de regressão linear

A função f descreve a relação funcional entre as duas variáveis, e o termo ϵ é usado para contabilizar o erro. Isso indica que se inserirmos um determinado valor de X como entrada, nossa saída será um valor de Y, dentro de um certo intervalo de erro. Você pode pensar nisso de várias maneiras:

- Resposta = Previsão + Erro
- Resposta = Sinal + Ruído
- Resposta = Modelo + Inexplicável
- Resposta = Determinístico + Aleatório
- Resposta = Explicável + Inexplicável

Que tipo de função devemos usar para f(X) para os salários?

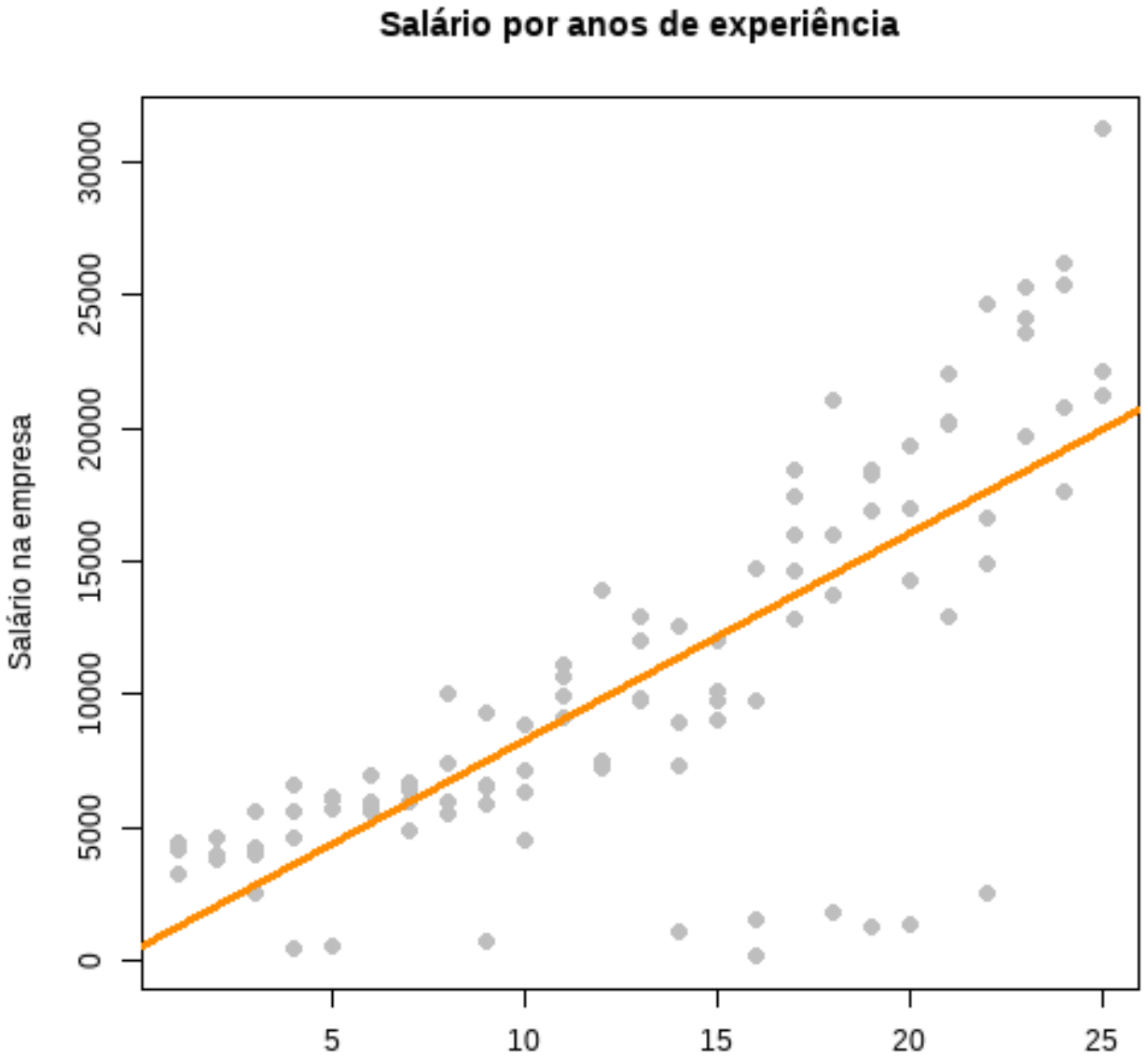
Poderíamos tentar modelar os dados com uma linha horizontal. Ou seja, o modelo para y (salário) não depende do valor de x. (Alguma função f(X)=c.) No gráfico abaixo, vemos que isso não parece fazer um trabalho muito bom. Muitos dos pontos de dados estão muito longe da linha laranja que representa c. Este é um exemplo de underfitting. A solução óbvia é fazer com que a função f(X) dependa de x (anos de trabalho).


Salário por anos de experiência
Salário na empresa
0
5000
10000
15000
20000
25000
30000
5
10
15
20
25

Também poderíamos tentar modelar os dados com uma função muito "inconstante" que tenta passar pelo maior número possível de pontos dos dados. Isso também não parece funcionar muito bem. Este é um exemplo de **overfitting**. (Observe que neste exemplo nenhuma função passará por todos os pontos, pois existem alguns valores x que possuem vários valores y possíveis nos dados.)


Salário por anos de experiência
Salário na empresa
0
5000
10000
15000
20000
25000
30000
5
10
15
20
25

Por fim, poderíamos tentar modelar os dados com uma linha bem escolhida em vez de um dos dois extremos tentados anteriormente. A linha no gráfico abaixo parece resumir muito bem a relação entre anos de experiência e salário. À medida que temos uma maior experiência de trabalho, aumentamos o nosso salário. Ainda há alguma variação sobre esta linha, mas parece capturar a tendência geral.


Salário por anos de experiência
Salário na empresa
0
5000
10000
15000
20000
25000
30000
5
10
15
20
25

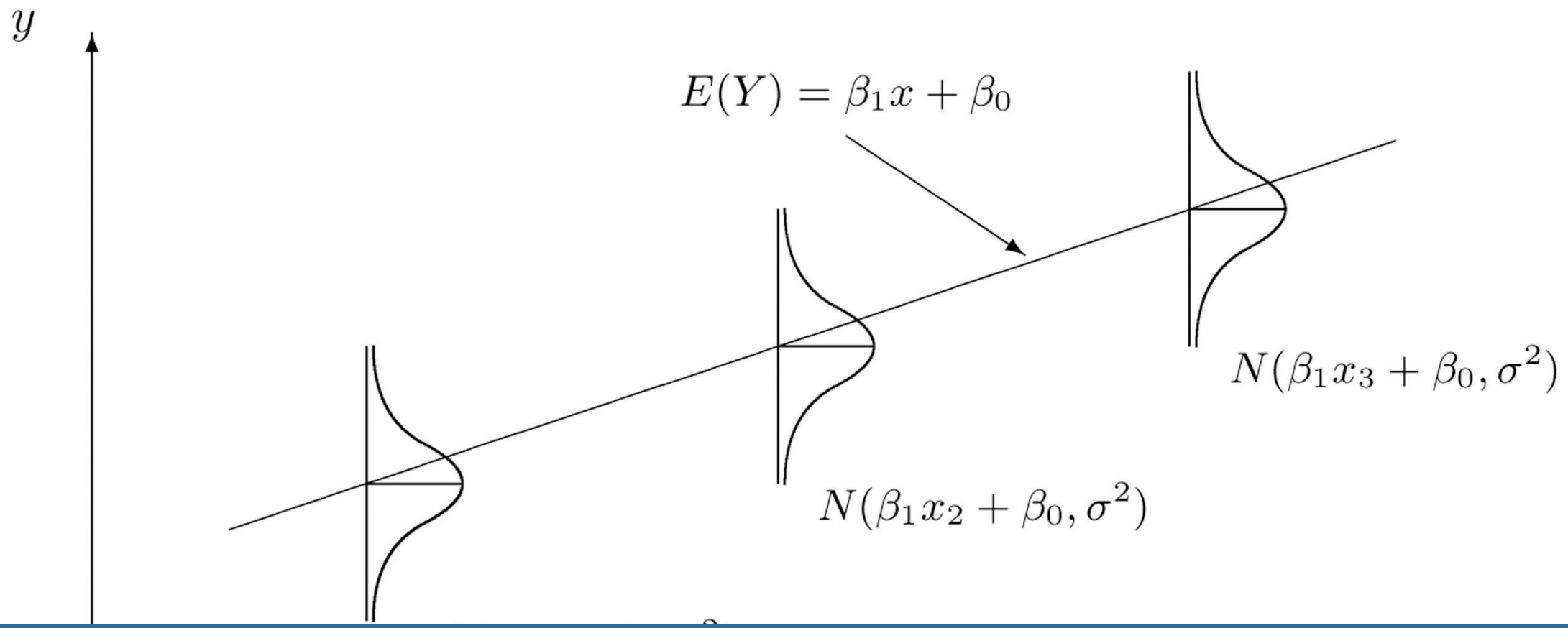
Com isso em mente, gostaríamos de restringir nossa escolha de $f(X)$ a funções *lineares* de X. Vamos escrever nosso modelo usando β_1 para a inclinação e β_0 para a intercepto,

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Os Y_i aleatórios são uma função de x_i, então podemos escrever sua média como uma função de x_i,

$$\mathrm{E}[Y_i \mid X_i = x_i] = \beta_0 + \beta_1 x_i.$$

Isso é exibido visualmente na imagem abaixo. Vemos que para qualquer valor x, o valor esperado (média) de Y é $\beta_0 + \beta_1 x$. A cada valor de x, Y tem a mesma variação σ^2.



Muitas vezes, falamos diretamente sobre os pressupostos . Eles podem ser encurtados de forma inteligente para **LINVI**.

- **L**inear. A relação entre Y e x é linear, na forma $\beta_0 + \beta_1 x$.
- **I** independente. Os erros ϵ são independentes.
- **N**ormal. Os erros, ϵ são normalmente distribuídos. Esse é o "erro" ao redor da linha segue uma distribuição normal.
- **V**ariância **I**gual. A cada valor de x, a variância de Y é a mesma, σ^2.

Fazendo previsões

Qual seria o valor esperado de salário para uma pessoa com 10 anos de experiência?

```
modelo = lm(salario ~ anos_trabalho, data = initech)
```

$$Y = 495,10 + 778,30X.$$

$$Y = 495,10 + 778,30 * 10 = 8.278,10$$

Qual seria o valor esperado de salário para uma pessoa com 20 anos de experiência?

$$Y = 495,10 + 778,30 * 20 = 16.061,10$$

sugiro olhar a função *predict* do R.

```
predict(modelo,data.frame(anos_trabalho=10))
```

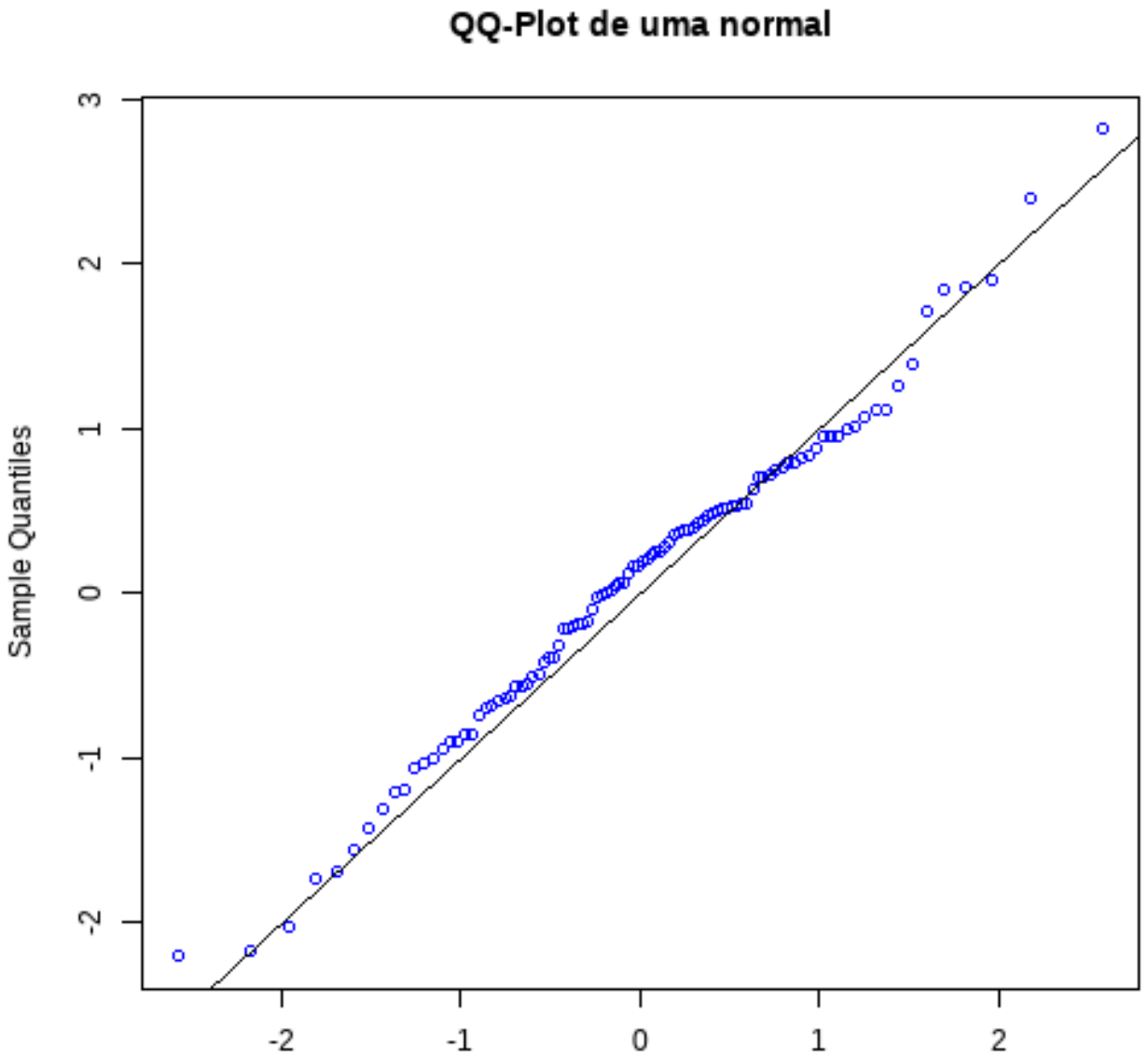
```
       1
8277.879
```

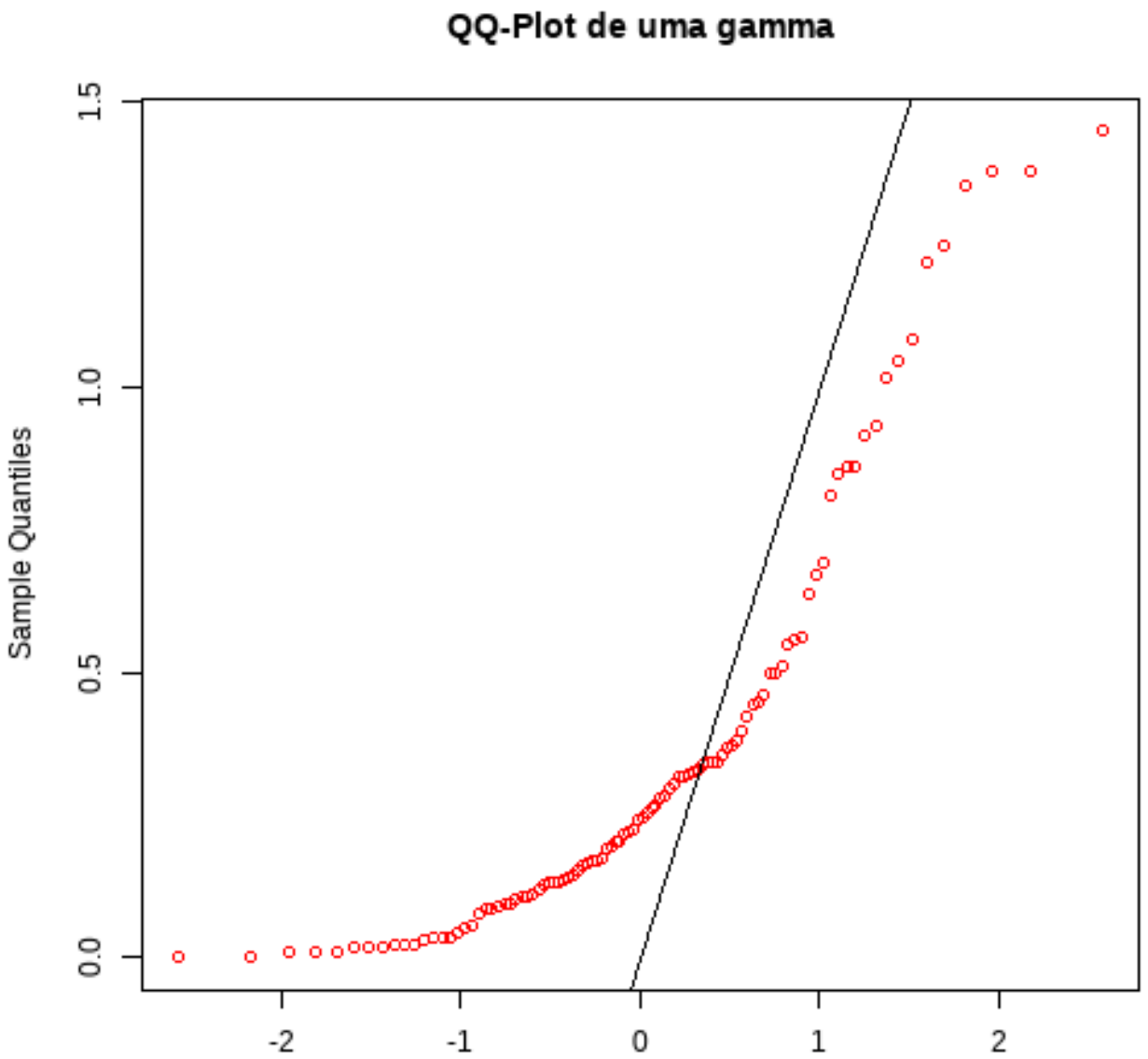
Gráfico QQ

Como descobrir se a sua variável tem uma distribuição Normal?

Um gráfico QQ é usado para avaliar visualmente quão próxima uma variável está da distribuição normal. Se os pontos caem na linha diagonal, então a distribuição dos dados podem ser consideradas perto de uma normal. Para fazer o gráfico QQ no R, podemos utilizar a função *qqnorm()*.

```
amostra_normal<-rnorm(100)
qqnorm(amostra_normal,col="red")
abline(a=0,b=1)
```


QQ-Plot de uma normal
Sample Quantiles
3
2
1
0
-1
-2
-2
-1
0
1
2


QQ-Plot de uma gamma
Sample Quantiles
0.0
0.5
1.0
1.5
-2
-1
0
1
2

Atividade: avalie quais as variáveis quantitativas das bases de dados *CARROS* e *Questionario Estresse* seguem uma distribuição Normal usando o gráfico QQ-plot.

Referências

1. BRUCE, Peter & BRUCE, Andrew **Estatística Prática Para Cientistas De Dados – 50 conceitos essenciais** Alta books, 320 p, 2019.
2. STEVENSON, Wiliam J. **Estatística aplicada à administração**. 1986.