

Regressão Simples

Curso Ninja de R para Ciências da Saúde

Henrique Pinto Gomide

08-10-2014

Antes de começar

- ▶ Esta aula é baseada no material do site OpenIntro.
- ▶ O laboratório foi traduzido pelo Professor Erikson Kaszubowski.
Larga a mão de ser ruim, agradeça-o.
- ▶ Downloads
- ▶ Banco de dados
- ▶ Texto base

Motivação

Abrir banco de dados

```
download.file("http://www.openintro.org/stat/data/mlb11.RData",  
             load("mlb11.RData"))
```

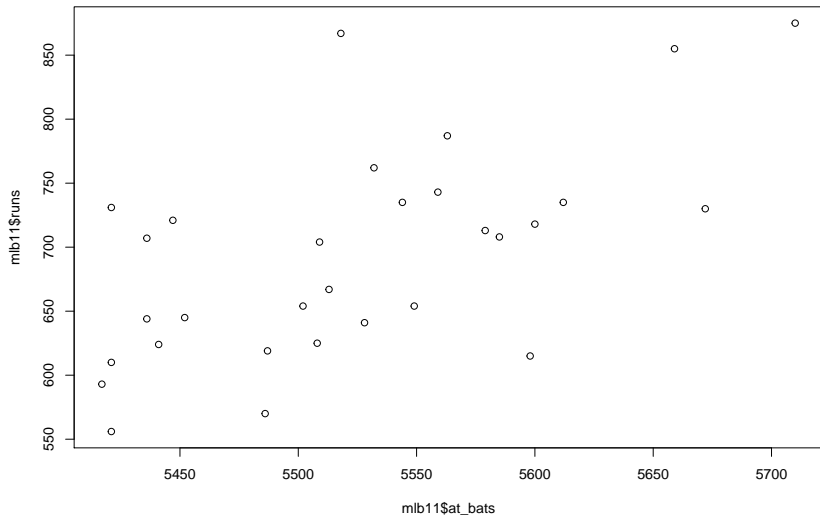
Exercícios - 1

Que tipo de gráfico você utilizaria para mostrar a relação entre *runs* (pontos) e alguma outra variável numérica? Crie um gráfico dessa relação utilizando a variável *at_bats* como preditora.

A relação parece ser linear? Se você soubesse o valor de *at_bats* (vez ao taco) de um time, você se sentiria confiante para utilizar um modelo linear para prever o número de pontos (*runs*)?

Respostas

```
plot(mlb11$at_bats, mlb11$runs)
```



R: Parece que sim...

Exercícios - 2

Qual o coeficiente de correlação, ou seja, a força de associação destas duas variáveis?

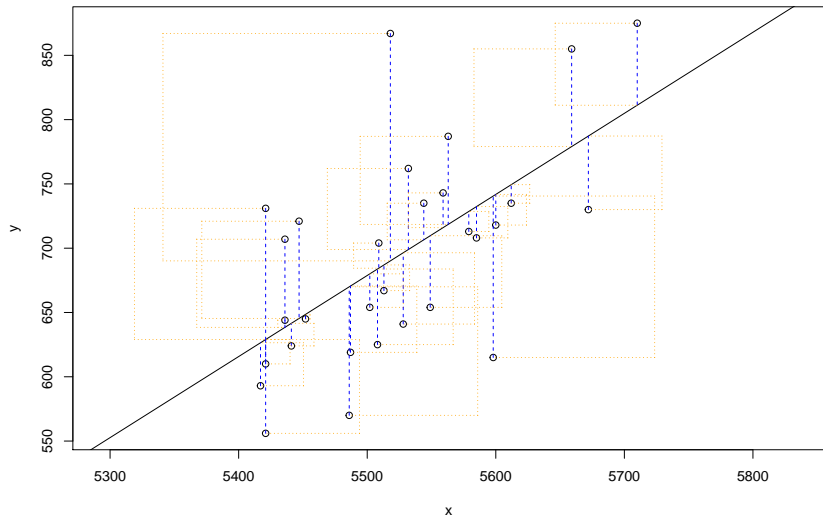
Resposta

```
cor(mlb11$at_bats, mlb11$runs)
```

```
## [1] 0.6106
```


Escolher uma melhor linha

```
plot_ss(mlb11$at_bats, mlb11$runs, showSquares = TRUE)
```



Click two points to make a line.

Função *lm*

```
m1 <- lm(runs ~ at_bats, data = mlb11)
```

Resultados

```
summary(m1)
```

```
##
```

```
## Call:
```

```
## lm(formula = runs ~ at_bats, data = mlb11)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max  
## -125.6  -47.0  -16.6    54.4   176.9
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)  
## (Intercept) -2789.243    853.696   -3.27  0.00287 **  
## at_bats      0.631      0.155     4.08  0.00034 ***
```

```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

```
##
```

```
## Residual standard error: 66.5 on 28 degrees of freedom
```

Exercício

1. Ajuste um modelo com a variável *homeruns* para prever *runs*. Escreva a equação da linha. O que o coeficiente angular nos diz?

Resposta

```
m2 <- lm(runs ~ homeruns, data = mlb11)
summary(m2)
```

```
##
```

```
## Call:
```

```
## lm(formula = runs ~ homeruns, data = mlb11)
```

```
##
```

```
## Residuals:
```

```
##      Min       1Q   Median       3Q      Max
```

```
## -91.61 -33.41   3.23  24.29 104.63
```

```
##
```

```
## Coefficients:
```

```
##              Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  415.239      41.678    9.96 1.0e-10 ***
```

```
## homeruns      1.835       0.268    6.85 1.9e-07 ***
```

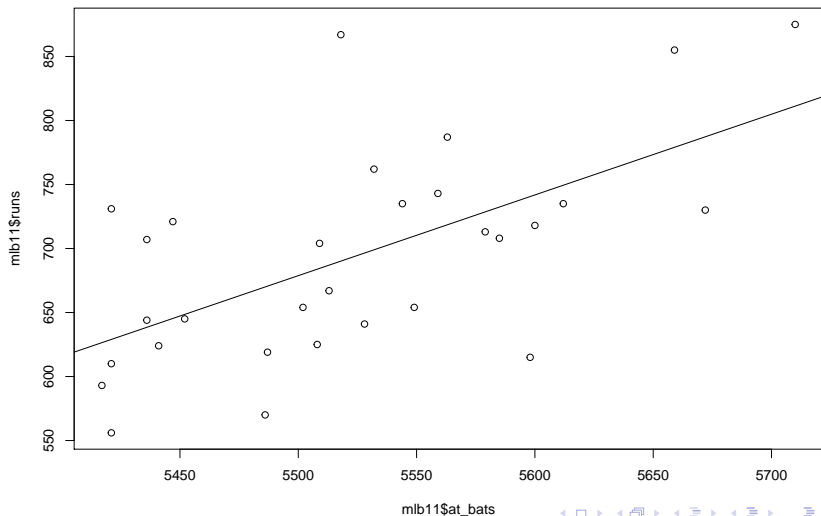
```
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
```

```
##
```

Predição e erro de predição

```
plot(mlb11$runs ~ mlb11$at_bats)
abline(m1)
```



Exercício

Se o técnico do time (dificilmente o Joel Santanta) visse a linha de regressão dos mínimos quadrados, quantos pontos (runs) ele prediria para um time com 5.578 vezes ao taco (*at_bat*)?

Esse valor superestima ou subestima o valor real? Por quanto erra?

Resposta

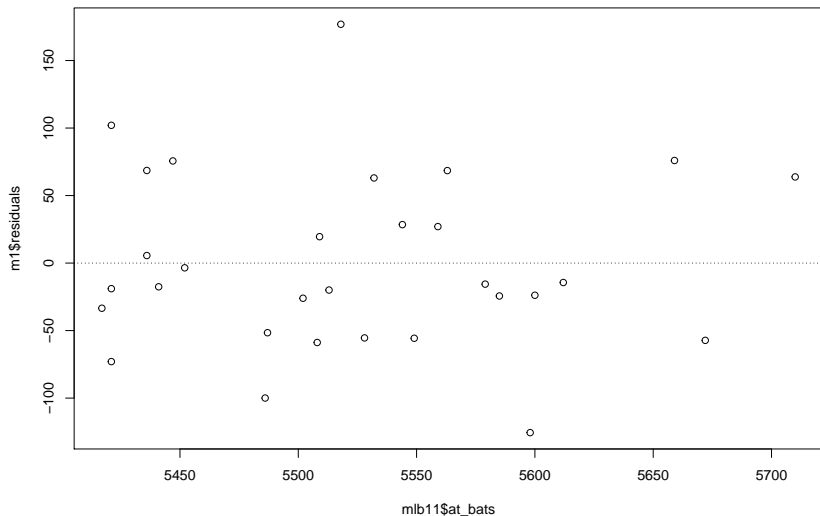
Fórmula $\text{runs} \sim -2789.2429 + 0.6305 * a_tbats$

```
runs <- -2789.2429 + 0.6305 * 5578  
runs
```

```
## [1] 727.7
```


Diagnósticos do modelo - Linearidade

```
plot(m1$residuals ~ mlb11$at_bats)  
abline(h = 0, lty = 3)
```



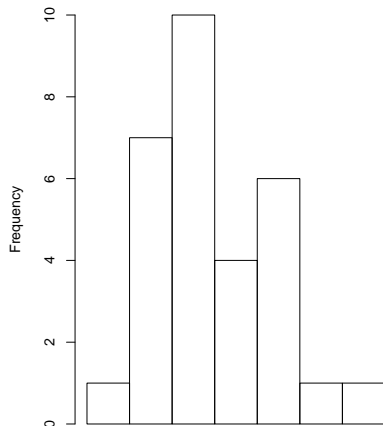
Diagnósticos do modelo - Resíduos distribuídos normalmente

- ▶ Parte do pressuposto: $\text{Constructo} = \text{Mensuração} + \text{Erro}$
- ▶ A condição de distribuição normal está atendida?

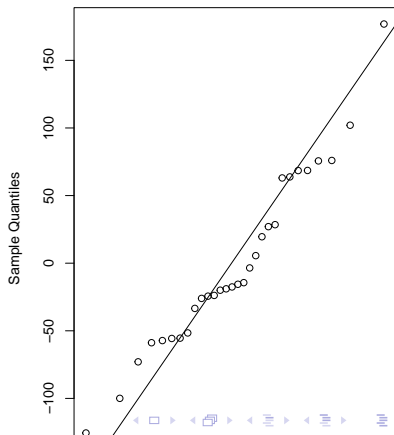
Resíduos distribuídos normalmente

```
par(mfrow = c(1,2))  
hist(m1$residuals)  
qqnorm(m1$residuals)  
qqline(m1$residuals)
```

Histogram of m1\$residuals



Normal Q-Q Plot



Diagnósticos do modelo - Variância

```
plot(m1$residuals ~ mlb11$at_bats)
abline(h = 0, lty = 3)
```

