Regressão linear simples

Ivan Bezerra Allaman

October 28, 2014

Introdução

- Foi visto na aula anterior que o coeficiente de correlação de Pearson é utilizado para mensurar o grau de associação entre duas variáveis quantitativas.
- E se o nosso interesse for ir além disso, ou seja, se estivermos interessados em saber o quanto o aumento no número de horas de treinamento de um empregado irá reduzir o número de acidentes daquele empregado?
- Ou ainda, o aumento em uma hora de sono irá aumentar em quanto o tempo de reação de uma pessoa?
- E se estivéssimos interessados em prever o tempo de reação de uma pessoa para uma determinada quantidade de horas de sono. Como fazer?
- Para responder a estas perguntas utilizaremos a análise de regressão linear simples.

Objetivo

- Estudar a relação funcional entre duas variáveis quantitativas.
- Estabelecer um modelo para entender a relação funcional entre as variáveis.
- Fazer predições como o modelo ajustado principalmente para valores que não foram observados na amostra.

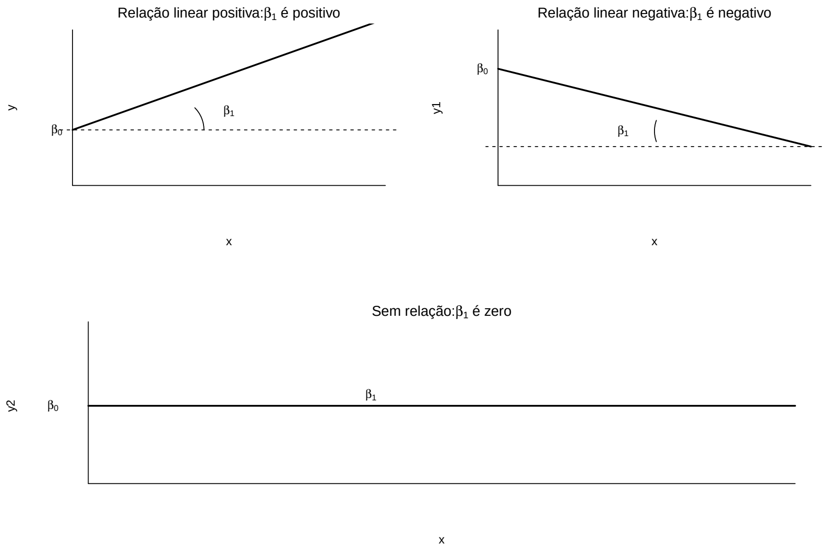
O modelo

- O modelo matemático que estabele a relação funcional entre duas variáveis é definido como:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

em que:
y= é a variável dependente
β_0= é o coeficiente linear ou intercepto da reta de regressão
β_1= é o coeficiente angular ou inclinação (declive) da reta de regressão
x= é a variável independente
ε= é o erro aleatório referente a variabilidade em y quem não pode ser explicada pela variável x.

Possíveis retas de regressão linear simples



Entendendo o comportamento de β_0 e β_1

- Acesse aqui

Ajuste da regressão

- Agora que já entendemos como funciona os parâmetros de uma regressão, chegou a hora de ajustarmos um modelo de regressão aos dados provenientes de uma amostra.
- Observemos a seguinte situação
  - Suponhamos que o dono de uma rede de restaurantes esteja interessado em saber a relação entre a quantidade de estudantes que almoçam em seus restaurantes com o lucro obtido trimestralmente.
  - Uma amostra de dez restaurantes foi coletado e os dados podem ser visualizados a seguir:

```
##    restaurante estudantes vendas_trimestrais
## 1            1          2                 58
## 2            2          6                105
## 3            3          8                 88
## 4            4          8                118
## 5            5         12                117
## 6            6         16                137
## 7            7         20                157
```

Método dos mínimos quadrados

- A idéia é encontrar valores de **b** e **a** que faça com que a reta de regressão passe na menor distância possível entre os pontos observados, minimizando o máximo possível o erro e fazendo com que o modelo explique o máximo possível a variabilidade dos dados.
- As letras **b** e **a** são os estimadores dos parâmetros β_0 e β_1 respectivamente.
- Primeiramente vamos demonstrar o método de maneira interativa, ou seja, vamos tentar encontrar valores de **b** e **a** que minimiza a soma de quadrados do erro.
- Acesse aqui
- Matematicamente o que queremos é minimizar a seguinte quantidade:

$$\sum \varepsilon^2 = \sum (y - \beta_0 - \beta_1 x)^2$$

- Qual é o mínimo que desejamos para $\sum \varepsilon^2$?

Exemplo

- Voltando ao exemplo anterior, vamos obter as estimativas de a e b para elaborarmos uma equação de regressão.

```r
#-------------------- opção 1 - na unha!--------------------
sumxy = sum(estudantes*vendas_trimestrais)
sumxsumy = sum(estudantes)*sum(vendas_trimestrais)
n = length(estudantes)

sumx2 = sum(estudantes^2)
sumx = sum(estudantes)

# Portanto, b é igual:

b = (sumxy - sumxsumy/n)/(sumx2 - (sumx)^2/n)
b
```

```
## [1] 5
```

```r
# Calculando a, tem-se:
```