Predizendo uma variável por outra - Regressão Linear - Parte 1

# Aula	43
☑ Preparada	☑
☑ Revisada	☑
☑ Lecionada	☑

▼ Já que estamos falando de uma associação linear…

Surgiu a idéia de representar esta associação por uma equação de reta…

…só se precisava saber como calcular esta equação de reta:

$$\hat{y} = f(x) = \beta_0 + \beta_1 x$$

Aqui, $\hat{y}$ é o valor estimado de *y* em função de *x*, β_0 é o coeficiente linear (ou intercepto) e β_1 é o coeficiente angular (ou inclinação).

▼ Equação de Regressão:

A equação da reta de regressão prediz o valor da variável resposta *y* como uma equação de reta tendo *x* como variável explicativa.

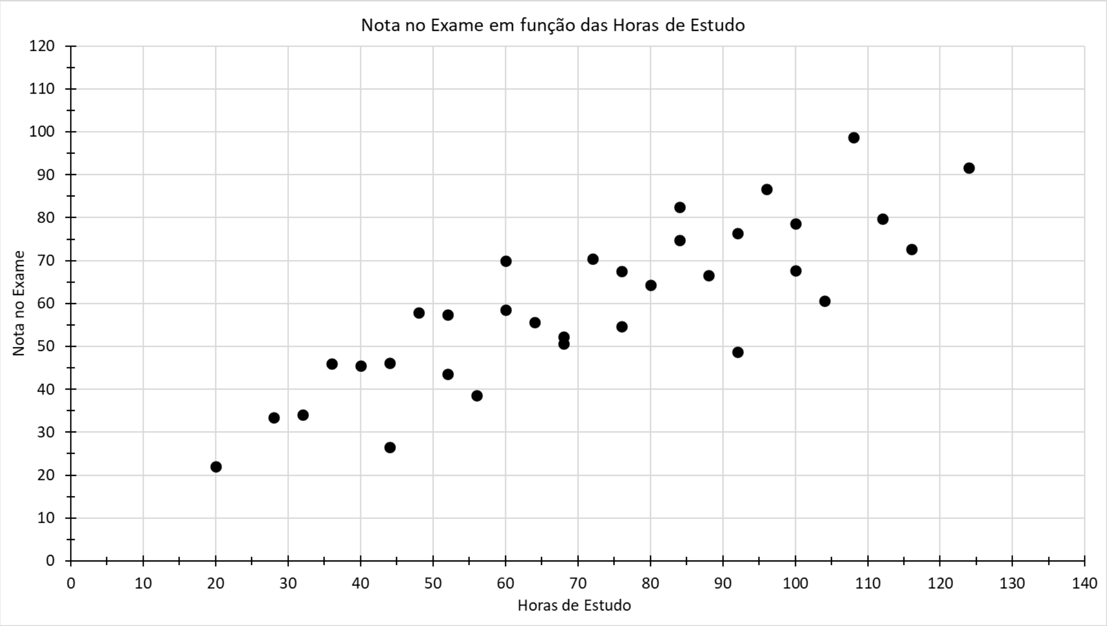
▼ O exemplo das notas no exame em função das horas de estudo:

▼ Imagine as seguintes notas em um exame em função da quantidade de horas de estudo:

Horas de Estudo	Nota no Exame
68	52,32
52	43,60
92	48,64
60	58,48
76	54,64
100	67,68
36	45,92
20	22,08
84	74,72
44	26,56
28	33,44
80	64,32
64	55,60
104	60,64
72	70,48
88	66,64
112	79,68
48	57,92
32	34,08
96	86,72
56	38,56
40	45,44
92	76,32
76	67,60
116	72,64

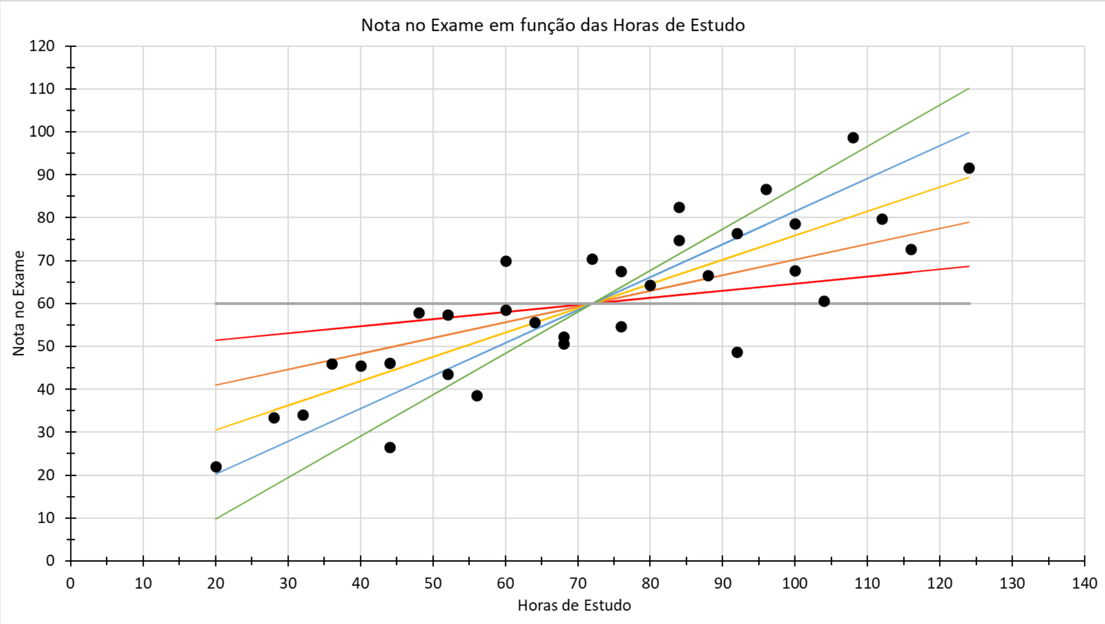
84	82,48
100	78,64
124	91,68
60	69,92
44	46,08
108	98,72
68	50,56
52	57,44

▼ Vamos ver o Gráfico de Dispersão dos dados:



É nítida a presença de uma associação linear positiva, por que, em geral, quanto mais horas de estudo, maior a nota no exame.

▼ Mas afinal, qual seria a melhor equação de reta para representar estes dados?



▼ Resposta

- No caso, a reta amarela é a reta que melhor representa os dados. Mas… Com base em que critério???

▼ A escolha da melhor reta tem a ver com os resíduos (ou erros).

▼ Os resíduos (ou erros) são a diferença entre o valor predito e o valor real

$$e = y - \hat{y}$$

▼ Podemos pensar no erro médio da reta ideal de regressão:

$$\overline{e} = \frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})$$

Mas o erro médio na reta ideal vai ter o mesmo problema do desvio médio, por que a soma vai ser igual a zero.

Precisamos nos livrar das componentes negativas, então vamos minimizar o quadrado dos erros. Por isso este

método é conhecido como Método dos Mínimos Quadrados.

Então aqui queremos minimizar a soma dos quadrados dos erros, escolhendo β_0 e β_1 tais que minimizem o quadrado do erro total, ou seja, a expressão:

$$[y - (\beta_0 + \beta_1 x)]^2$$

Não vou fazer a dedução matemática aqui, vamos apenas estudar as propriedades…