Predizendo uma variável por outra - Regressão Linear - Parte 2

# Aula	44
☑ Preparada	☑
☑ Revisada	☑
☑ Lecionada	☑

▼ Propriedades da equação do Método dos Mínimos Quadrados

A equação de reta obtida pelo Método dos Mínimos Quadrados apresenta tanto resíduos positivos quanto resíduos negativos.

A equação de reta obtida pelo Método dos Mínimos Quadrados zera a soma dos resíduos.

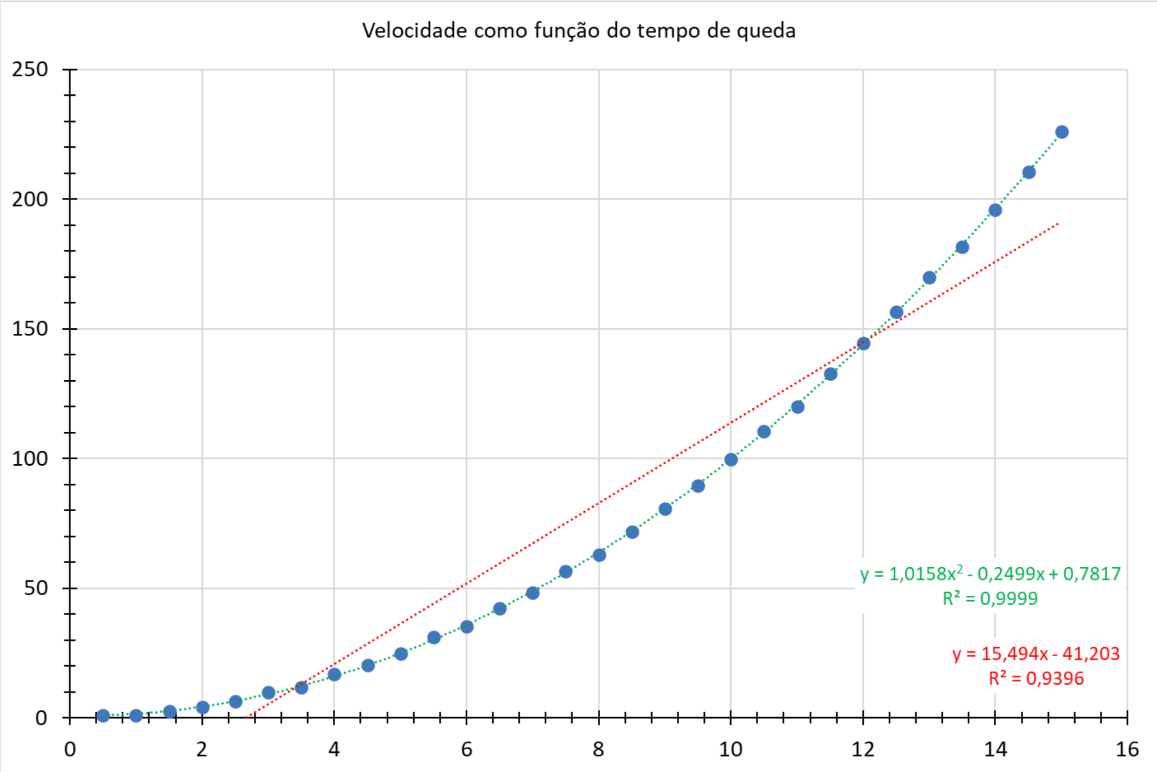
A equação de reta obtida pelo Método dos Mínimos Quadrados minimiza o soma dos quadrados dos resíduos.

A equação de reta obtida pelo Método dos Mínimos Quadrados passa pelo ponto médio de coordenadas $(\overline{x}, \overline{y})$.

A equação de reta obtida pelo Método dos Mínimos Quadrados pode ser generalizada para dar origem à Regressão Linear Múltipla:

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + (...) + \beta_n x_n$$

A equação de reta obtida pelo Método dos Mínimos Quadrados **é linear nos seus coeficientes β_i**. Podemos tratar de relações não lineares com as variáveis x_i.



▼ Fórmula do Método dos Mínimos Quadrados

Lembrando que estamos falando da equação de reta **de 1 variável preditora**:

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

▼ A inclinação β_1 pode ser calculada com

$$\beta_1 = r_{(x,y)} \frac{s_y}{s_x}$$

Ou seja, a inclinação tem a ver com a correlação e com os desvios padrão de *x* e *y*.

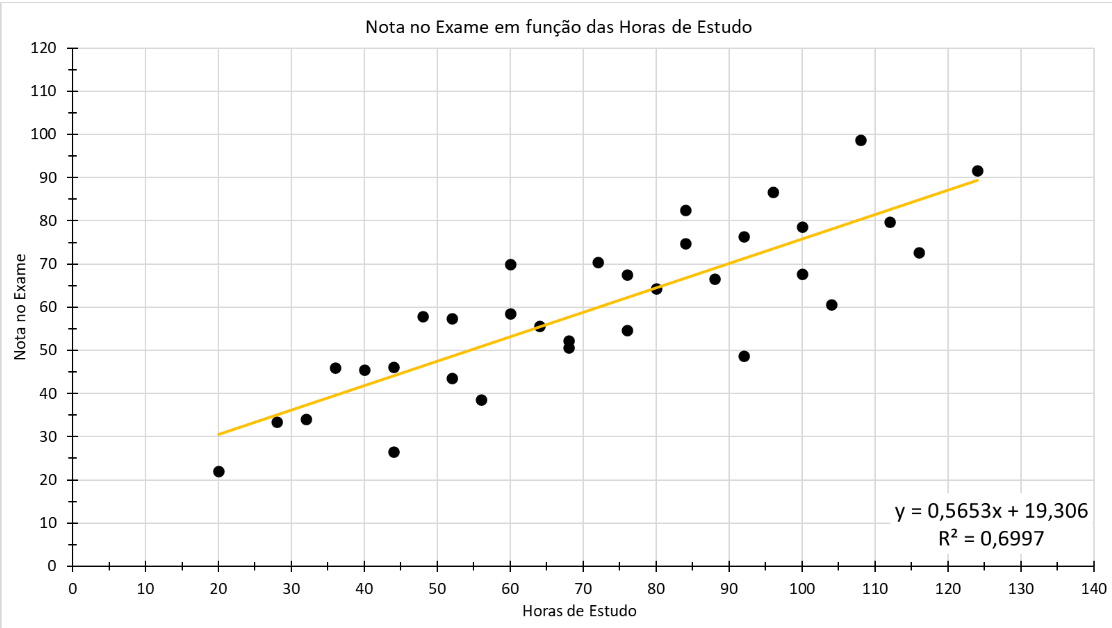
Na prática, os softwares não dependem do cálculo da correlação nem do desvio padrão de *y*, podendo calcular mais rapidamente a inclinação β_1 por

$$\beta_1 = \frac{1}{(n-1)s_x^2} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

- **A partir da inclinação β_1 podemos calcular o intercepto β_0 com**

$$\beta_0 = \overline{y} - \beta_1 \overline{x}$$

- **Do exemplo anterior**



$$\beta_0 = 19,306 \qquad \beta_1 = 0,5653 \qquad \hat{y} = 19,306 + 0,56$$

Afinal, e o R²?

- **Primeira interpretação**

<u>No caso específico de uma equação de reta</u>, R^2 é exatamente o quadrado da correlação entre *x* e *y*. Ou seja.

$$R^2 = r^2_{(x,y)}$$

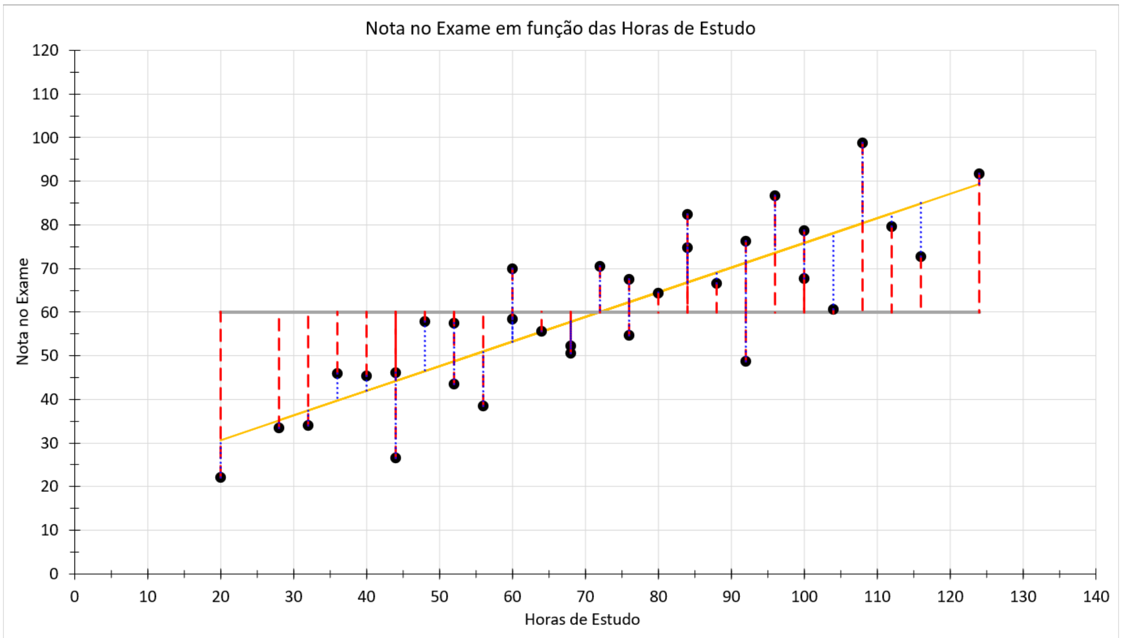
Inclusive o quadrado na notação vem exatamente daí. Vemos pela fórmula que o R^2 varia entre 0 e 1 (ou entre 0% e 100%).

Segunda interpretação

Para modelos em geral, o R^2 é uma função entre a soma de quadrados de resíduos do modelo e a soma de quadrados de distância da média de *y* dada por

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{\sum_{i=1}^{n}(y_i - \overline{y})^2}$$

- A fórmula acima foi colorida para você reparar na distância entre os dados e a média de *y* em vermelho e na distância entre os dados e o modelo de *y* em azul.



Importantíssimo:

O R^2 é uma medida de redução do erro que se tem ao usar o modelo ao invés de usar a média. No exemplo acima, o R^2 de 0,6997 significa que erramos 70% menos usando o modelo para prever *y* do que se fizéssemos a previsão de *y* pela média.

Curiosidade:

Se você propuser no exemplo do gráfico acima uma equação de reta decrescente (enquanto a associação dos dados é

positiva), é possível que a conta proposta na segunda interpretação tenha um resultado negativo…

Outros diagnósticos em modelos de regressão:

MAE (*Mean Absolute Error* [Erro Absoluto Médio])

$$MAE = \frac{1}{n}\sum_{i=i}^{n}|y_i - \hat{y}_i|$$

MSE (*Mean Squared Error* [Erro Quadrático Médio])

$$MSE = \frac{1}{n}\sum_{i=i}^{n}(y_i - \hat{y}_i)^2$$

RMSE (*Root Mean Squared Error* [Raiz do Erro Quadrático Médio])

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=i}^{n}(y_i - \hat{y}_i)^2}$$

Então… Quais destes indicadores vamos usar?

O MSE é expresso na unidade ao quadrado.

Já o MAE e o RMSE estão na mesma unidade que o dado original, então usamos os dois!

Faz sentido?

A comparação entre o MAE e o RMSE vai nos mostrar quantos valores extremos temos nos erros do modelo em que estamos medindo estes indicadores.