Introdução

Suponha um conjunto de dados, ou seja, um conjunto de pontos em um gráfico (diagrama de dispersão), e propõe-se desenhar uma linha que passe o mais próximo possível de todos esses pontos. A posição dessa linha, nomeada de linha de regressão, deve apresentar a tendência média desses pontos, minimizando as distâncias verticais entre a linha e os pontos (erros). Esse processo de ajuste é conhecido como Método de Mínimos Quadrados Ordinários, que busca encontrar a melhor reta para descrever a relação linear entre as variáveis.

O Método de Mínimos Quadrados Ordinários (em inglês *Ordinary Least Squares*, OLS) é uma técnica estatística que visa encontrar a linha de regressão que melhor se ajusta a um conjunto de dados. O princípio desta técnica é minimizar a soma dos quadrados dos erros ou resíduos, onde estes resíduos são as diferenças entre os valores observados da variável dependente (y) e os valores preditos pela equação da linha de regressão (Gujarati & Porter, 2011; Wooldridge, 2023). O objetivo consiste em estimar ou prever o valor médio da variável dependente (y) com base nos valores observados da variável independente (x).

A importância do OLS está no fato de ele ser uma das técnicas mais utilizadas para modelar relações entre variáveis, permitindo não apenas prever valores futuros, mas também entender a força e direção dessa relação. Em diversos campos, como economia, finanças e ciências naturais, essa técnica possibilita a tomada de decisões informadas, a identificação de padrões e a realização de inferências sobre os dados. Além disso, ao minimizar os erros, o OLS garante que as previsões sejam, em média, as mais precisas possíveis, tornando-o uma ferramenta essencial na análise de dados e modelagem estatística.

Método de Mínimos Quadrados Ordinários (OLS)

A modelagem da relação entre a variável dependente (y) e a variável independente (x) pelo método OLS é amplamente utilizado em regressão linear. A fórmula da linha de regressão ajustada é apresentada na equação 1, onde y_i representa o valor esperado (ou predito) de y para um dado valor de x no ponto i.

$$y_i = \beta_0 + \beta_1 . x_i + u_i \qquad (1)$$

Onde β_0 (intercepto, equação 2) representa o valor de y_i quando x_i é igual a zero, β_1 (coeficiente angular, equação 3) é a inclinação da linha, e u_i (equação 4) é o resíduo da regressão, a parte da variação de y_i que não pode ser explicada pela x_i.

$$\beta_0 = \bar{y} - \beta_1 . \bar{x} \qquad (2)$$

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (3)$$

$$u_i = y_i - (\beta_0 + \beta_1 . x_i) \quad (4)$$

Com $\bar{y}$ e $\bar{x}$ correspondem às médias dos valores observados da variável dependente y e da variável independente x, respectivamente.

O método OLS ajusta β_0 e β_1 de forma que a soma dos quadrados dos erros (SSE) seja mínima, promovendo assim o melhor ajuste possível aos dados. A SSE (equação 5) indica o total da variação nos dados que não é explicada pelo modelo.

$$SSE = \sum_{i=1}^{n} u_i^2 \quad (5)$$

Enquanto o SSE apresenta a magnitude dos erros na predição dos valores, o coeficiente de determinação R^2 (equação 6) quantifica a proporção da variação total da variável dependente (y) que é explicada no modelo de regressão. Ou seja, R^2 é um indicativo do quão bem a linha de regressão ajustada representa os dados observados. O valor de R^2 varia entre 0 e 1, onde 1 indica que o modelo explica perfeitamente a variabilidade dos dados, e valores próximos a 0 sugerem que o modelo tem baixa capacidade de explicação.

$$R^2 = 1 - \frac{SSE}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad (6)$$

Para a ilustração do Método de Mínimos Quadrados Ordinários (OLS), considere uma situação em que se deseja analisar a relação entre as horas que um aluno dedica ao estudo e a nota que ele obtém em uma prova. Em termos gerais, presume-se que quanto mais tempo o aluno estude, melhores sejam suas notas, pois isso reflete um maior preparo e assimilação do conteúdo. No entanto, é importante ressaltar que essa relação não é perfeitamente previsível. Fatores aleatórios, como a qualidade do estudo, o estado emocional do aluno e condições externas, assim como variações individuais de aprendizado, podem influenciar significativamente os resultados obtidos.

Neste cenário, coletaram-se 15 observações fictícias representadas na Tabela 1 e Figura 1. Os dados simulam os diferentes cenários de horas estudadas e as notas correspondentes. Essa abordagem permite avaliar a tendência média da relação entre as variáveis e determinar até que ponto as horas de estudo são capazes de prever o desempenho na prova, mesmo considerando as incertezas e variabilidades presentes no processo.

Tabela 1 - Valores das observações de horas de estudo de cada aluno e suas as respectivas notas da prova (de 0 a 100 pontos) e valores da média para cada variável.

Horas Estudadas	Notas da Prova (0 a 100)

	4,40	23,33
	9,60	48,12
	7,60	37,71
	6,40	31,34
	2,40	9,06
	2,40	10,58
	1,50	6,69
	8,80	46,09
	6,40	32,74
	7,40	33,34
	1,20	6,57
	9,70	47,88
	8,50	41,11
	2,90	15,78
	2,60	15,24
Média:	5,45	27,04

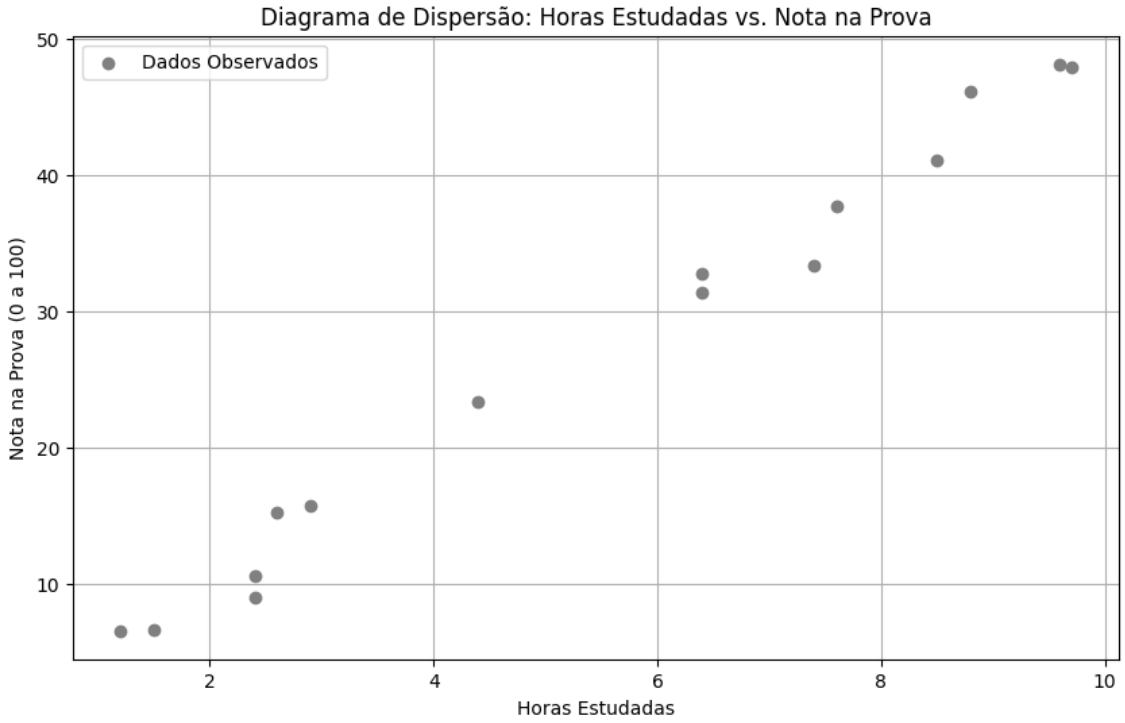

Figura 1 - Diagrama de dispersão com os valores fictícios sobre a relação das horas estudadas e nota da prova dos alunos para uma determinada matéria.

A seguir, a Figura 2 apresenta a relação entre as horas de estudo (variável independente) e as notas obtidas na prova (variável dependente). Cada ponto representa uma observação individual, e a linha de regressão é ajustada pelo método de Mínimos Quadrados Ordinários (OLS) para mostrar a tendência geral dos dados. A inclinação da linha (representada pelo coeficiente angular β_1) indica que, à medida que o número de horas estudadas aumenta, as

notas tendem a aumentar também. O coeficiente de determinação R^2 no gráfico indica o quão bem a linha de regressão ajusta os dados: um valor próximo de 1, como o mostrado (por exemplo, 0,99), sugere que a maior parte da variação nas notas pode ser explicada pelo tempo de estudo. A soma dos quadrados dos erros (SSE) mostra o quanto os pontos divergem da linha ajustada, ilustrando a precisão do ajuste.

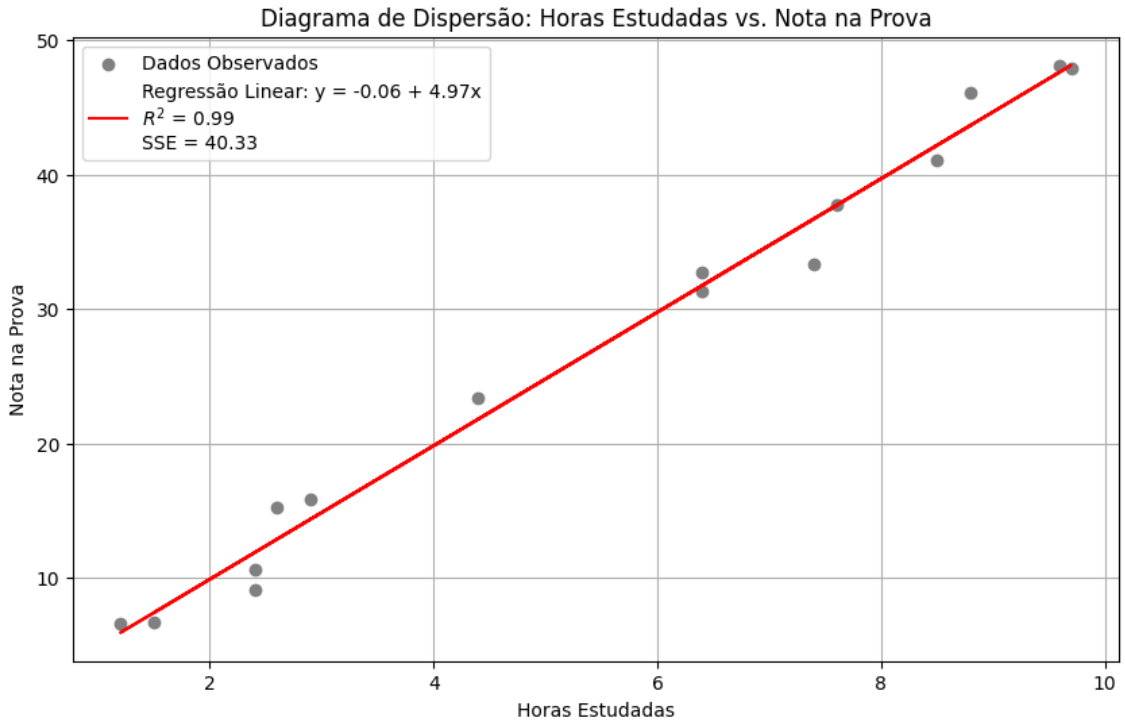

Figura 2 - Diagrama de dispersão com os valores fictícios sobre a relação das horas estudadas e nota da prova dos alunos para uma determinada matéria e com a reta de regressão (em vermelho), valores de SSE e R^2.

Referências Bibliográficas

Gujarati, D. N., & Porter, D. C. (2011). Econometria básica. ed. *Porto Alegre: AMGH*.

Wooldridge, J. M. (2023). *Introdução à econometria: uma abordagem moderna*. Cengage Learning.