

Regressão Linear com Estudo de Caso



Conteudista: Prof. Esp. Rafael Tasinato

Revisão Textual: Esp. Pérola Damasceno

Objetivos da Unidade:

- Desenvolver, aprender a aplicar os conceitos e técnicas de regressão;
- Resolver um problema real de negócios com um projeto de análise de dados usando regressão;
- Adquirir a capacidade de identificar o modelo adequado, avaliar o seu desempenho e comunicar os seus resultados.



 Material Teórico



 Material Complementar



 Referências



Material Teórico

O que é Análise de Regressão?

Análise de regressão, ou simplesmente **regressão**, busca saber de que maneira uma variável pode prever o comportamento da outra, ou seja, variação que y tem para x .

Por exemplo: Qual o impacto nas **vendas de imóveis** (y), se tiver um aumento na **taxa juros** (x)?



Figura 1 – Venda de imóveis

Fonte: Getty Images

Qual o melhor **horário** (**y**) para viajar de ônibus e **metrô** (**x**) em determinada cidade?



Figura 2 – Metrô e ônibus

Fonte: Getty Images

Terei mais vendas (**y**) se tiver mais movimento em frente à loja (**x**)?



Figura 3 – Centro comercial

Fonte: Getty Images

Qual o consumo de energia elétrica de uma residência (y) com alteração da temperatura de uma determinada cidade (x)?

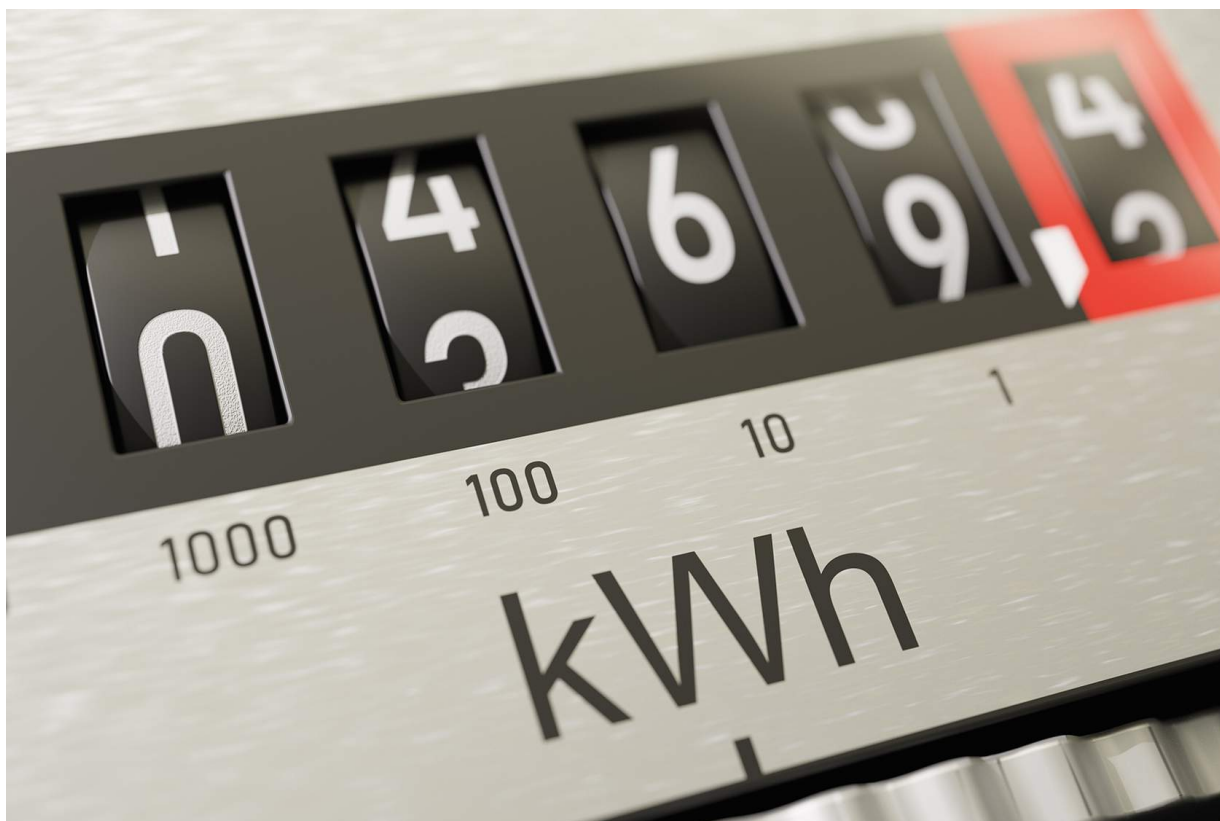


Figura 4 – Consumo de energia elétrica

Fonte: Getty Images

Como podemos observar estas questões são comuns no nosso dia a dia, temos as respostas para algumas delas, de forma empírica, e um paralelo com alguns gestores do mundo corporativo, comumente o fazem diariamente, e correm o risco de uma decisão errada falir uma empresa, gerar perda de capital e/ou representação no mercado.

Para auxiliar nas tomadas de decisões destes gestores ou de outros profissionais, é efetuado o levantamento da Análise de Dados Exploratórios com Regressão ou, em outras palavras, separamos os dados, tratamos e identificamos a relação entre eles com a possibilidade de preverem as possíveis consequências de suas decisões.

Introdução à Regressão Linear Simples: o que é, para

que Serve e Exemplos de Aplicações

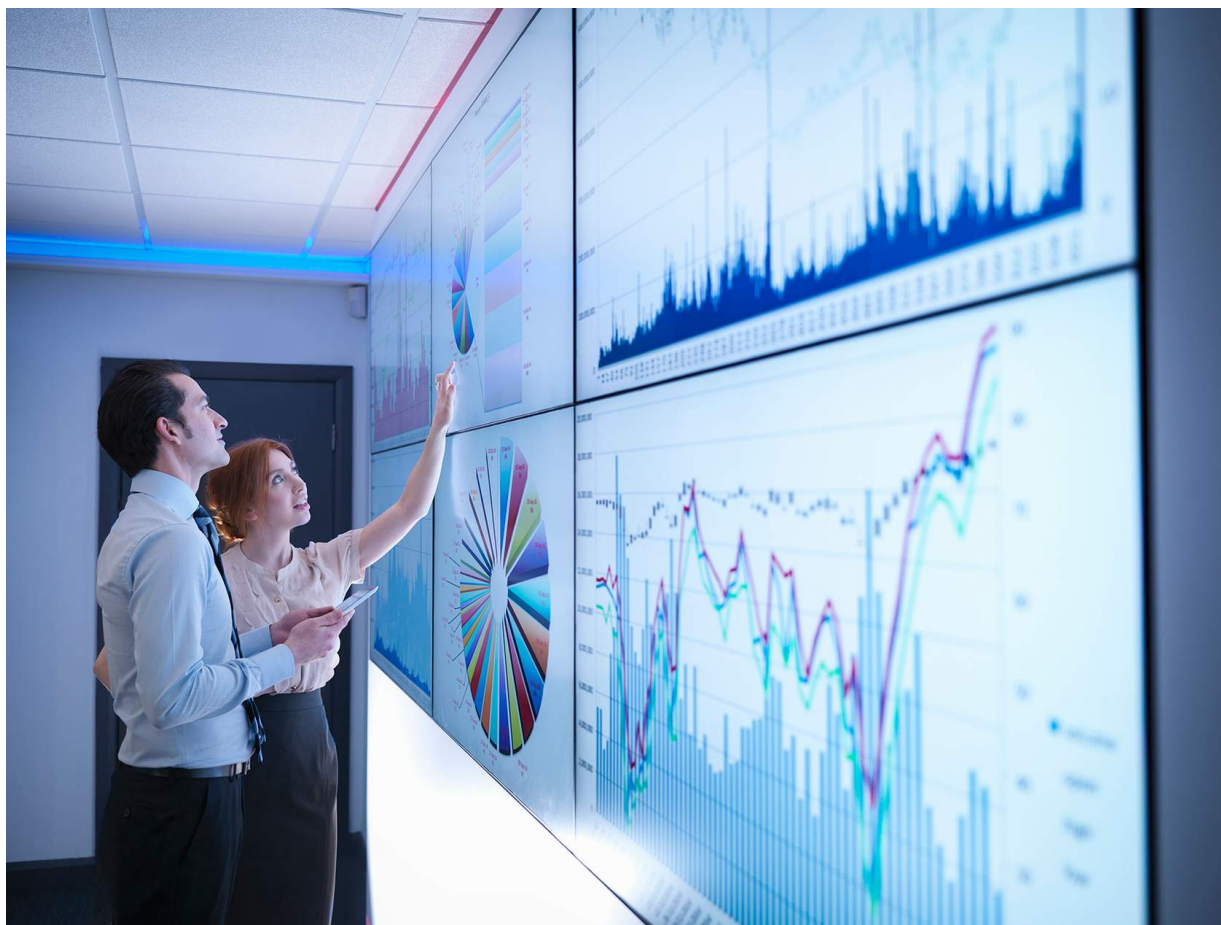


Figura 5 – Regressão linear

Fonte: Getty Images

O que é?

Regressão linear é um método estatístico que permite estudar a relação entre uma variável dependente (y) e uma variável independente (x), ou seja, como uma variável se comporta em função de outra (LOCK, 2017).

Larson e Farber (2006) definem uma reta de regressão, também chamada de reta de melhor ajuste, como a reta para a qual a soma dos quadrados dos resíduos é um mínimo.

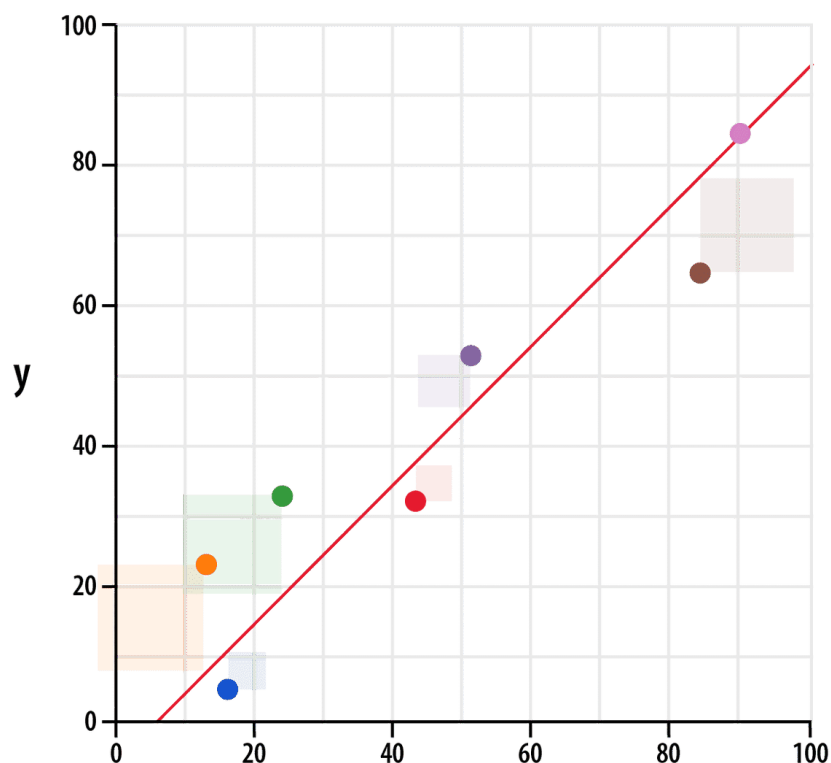


Figura 6 – Regressão linear e quadrados

Fonte: Reprodução

#ParaTodosVerem: um gráfico com eixo x e y, onde o eixo y é na vertical e o x na horizontal. No olhar do observador, da esquerda para a direita, inicia-se em 0 e termina na numeração 100, com intervalos numéricos de 20. No gráfico, apresenta pontos dispersos nos gráficos com um padrão linear crescente. Os pontos estão dispersos, porém é possível identificar uma reta passando entre eles e o espaço do ponto até a reta, apresentando diversos quadrados de diferentes tamanhos e cores. O ponto que está diretamente sobre a reta não apresenta representação gráfica de quadrado até a reta. No gráfico, podemos identificar uma reta entre os pontos e apresentar a distância entre eles até a reta. Fim da descrição.

Para que Serve?

Podemos usar a regressão linear para analisar se a altura de uma pessoa está relacionada ao seu peso, ou se o tempo de estudo influencia na nota de uma prova. A regressão linear simples usa apenas uma variável independente, enquanto a regressão linear múltipla usa mais de uma.

Nessa Unidade, vamos nos concentrar na regressão linear simples.

Exemplos de Aplicações

Para realizar uma análise de regressão linear, precisamos escolher e coletar dados que sejam relevantes para o problema que queremos investigar.

- Os dados devem ser quantitativos, ou seja, expressos em números, e devem ter uma possível relação linear, ou seja, que possam ser aproximados por uma reta;
- Os dados podem ser obtidos de diversas fontes, como pesquisas, experimentos e observações;
- É importante que os dados sejam confiáveis, representativos e tenham uma quantidade suficiente para permitir uma análise adequada.

Depois de coletar os dados, podemos ajustar um modelo de regressão linear aos mesmos, ou seja, encontrar uma equação que descreva a melhor reta que se adapte aos pontos no gráfico.

Essa reta é chamada de reta de regressão e tem a forma $y = ax + b$, onde a é o coeficiente angular e b é a interceptação. O coeficiente angular indica a inclinação da reta e o quanto y varia em relação ax . A interceptação indica o valor de y quando x é zero. Para encontrar os valores de a e b podemos usar diferentes métodos, como o método dos mínimos quadrados desenvolvido por Carl Friedrich Gauss (1777 - 1855) e proposto por Adrien-Marie Legendre (1752-1833), que minimiza a soma dos quadrados das distâncias verticais entre os pontos e a reta.

A partir do modelo de regressão linear, podemos interpretar os resultados e tirar conclusões sobre a relação entre as variáveis.

Podemos verificar se há uma relação entre elas, ou seja, se elas variam juntas em uma mesma direção (correlação positiva) ou em direções opostas (correlação negativa).

Podemos também medir a intensidade dessa correlação pelo coeficiente de correlação R^2 , que varia de -1 a 1 :

- Quanto mais próximo de -1 ou 1 , mais forte é a correlação;
- Quanto mais próximo de zero, mais fraca é a correlação.

Além disso, podemos avaliar a qualidade do ajuste do modelo aos dados e testar hipóteses sobre os parâmetros do modelo. Uma forma de avaliar a qualidade do ajuste é pelo coeficiente de determinação (R^2), que indica quanto da variação de y é explicada pela variação de x . O R^2 varia de 0 a 1 , e quanto mais próximo de 1 , melhor é o ajuste. Outra forma é pelo erro padrão da estimativa (S), que indica o desvio médio dos valores observados em relação aos valores estimados pela reta

Quanto menor o s , melhor é o ajuste.

Para testar hipóteses sobre os parâmetros do modelo, podemos usar o teste t , que compara o valor estimado de um parâmetro com um valor nulo (geralmente zero) e verifica se há evidência estatística para rejeitar ou não essa hipótese nula. Por exemplo, podemos testar se o coeficiente angular é diferente de zero, o que indicaria que há uma relação linear entre as variáveis.

Por fim, podemos aplicar a regressão linear simples a problemas reais de diferentes áreas, como Economia, Biologia, Engenharia ou nas demais áreas de conhecimento. A regressão linear simples pode ser usada para fazer previsões, estimar tendências e avaliar relações causais.

No entanto, é preciso ter cuidado ao aplicar o modelo e verificar se ele é adequado.

Como Ajustar uma Reta aos Dados: Métodos de Mínimos Quadrados e Critérios de Qualidade do Ajuste

Métodos dos Mínimos Quadrados

Mínimos quadrados é um método matemático que busca encontrar a melhor reta que se ajusta a um conjunto de dados, minimizando a soma dos quadrados das diferenças entre o valor estimado pela reta e o valor observado nos dados.

Essas diferenças são chamadas de resíduos e podem ser representadas por:

$$E_i = y_i - (a + bx_i)$$

Sendo:

- y_i o valor observado;
- a o coeficiente linear;
- b o coeficiente angular;
- x_i o valor da variável independente.

Para encontrar os valores de a e b que minimizam a soma dos quadrados dos resíduos, podemos derivar a função:

$$J(a,b) = \sum_{i=1}^N \epsilon_i^2$$

Em relação a a e b e igualar a zero, obtém-se o sistema:

$$\frac{\partial J}{\partial a} = -2 \sum_{i=1}^N (y_i - bx_i - a) = 0$$

$$\frac{\partial J}{\partial b} = -2 \sum_{i=1}^N x_i (y_i - bx_i - a) = 0$$

Resolvendo esse sistema, podemos obter as fórmulas para a e b em função dos dados:

$$a = \frac{n \cdot \sum x \cdot y - \sum x \cdot \sum y}{n \sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y}{n} - a \cdot \frac{\sum x}{n}$$

Onde \bar{x} e \bar{y} são as médias dos valores de x e y , respectivamente.

Essas são as fórmulas dos mínimos quadrados que permitem encontrar a reta que melhor se ajuste aos dados; já os critérios da qualidade dos mínimos quadrados são baseados na ideia de que o método deve encontrar a melhor reta que se ajusta aos dados, minimizando a soma dos quadrados dos resíduos.

Os resíduos são as diferenças entre os valores observados e os valores estimados pela reta. Quanto menor for a soma dos quadrados dos resíduos, melhor será o ajuste da reta aos dados, sendo necessário encontrar valores para a e b , ou seja, vamos estimar a inclinação da reta

utilizando uma amostra aleatória de dados correspondentes a intersecção de x e y . A inclinação nos fornece o efeito em y da mudança de uma unidade em x (CHEIN, 2019).

Saiba Mais

É importante analisar o gráfico e ver se existem dados nos extremos, pois estes dados isolados podem prejudicar a nossa análise.

Pontos de *Outlier* e Interferência na Reta do Gráfico

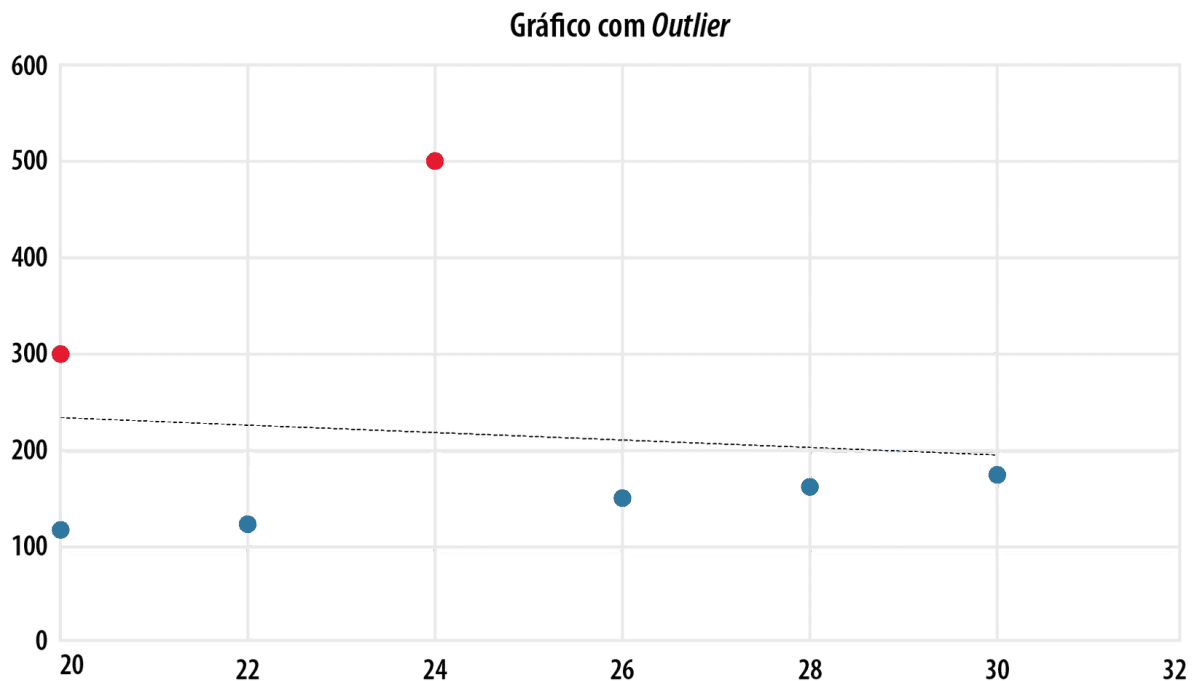


Figura 7 – Efeitos sobre a reta, com *Outlier*

#ParaTodosVerem: gráfico de dispersão apresentando uma sequência na reta, que sofre a interferência de dois *outliers*. A representação gráfica apresenta uma reta na horizontal. Há duas retas, X e Y, sendo a Y com início em 0 e término no número 600; já a reta X tem início no 20 e término no número 32. Os intervalos da primeira reta (Y, na vertical) é de 100, já na reta horizontal (X) é de 2. Apresenta uma sequência de pontos da cor azul no parâmetro y, próximo do número 100 e com uma progressão de x do 20 até o ponto 30, com 5 pontos, sendo estes de forma aproximada (Y e X). Ponto 1: 101 – 20; Ponto 2: 103 – 22; Ponto 3: 101 – 26; Ponto 4: 105 – 26; e o Ponto 5: 198 – 30. Apresenta dois pontos (os *outliers*) marcados em vermelho: Ponto vermelho 1: 300,20 e Ponto vermelho 2: 500,24. Por serem apenas dois pontos e muito fora do padrão da reta, nós devemos retirar do nosso estudo, pois atrapalham a nossa análise e o estudo da reta. Fim da descrição.

No gráfico com *outlier*, apresentam dois pontos que saem do padrão da reta, e estes pontos podem atrapalhar as nossas análises da reta, por isso é importante retirá-los do cálculo, uma vez que estes pontos indicam eventos fora do padrão.

Existem alguns critérios que podem ser usados para avaliar a qualidade dos mínimos quadrados, tais como:

- O coeficiente de determinação (R^2), que mede a proporção da variância total dos dados, que é explicada pela reta de regressão. Quanto mais próximo de 1 for o valor de R^2 , melhor será o ajuste da reta aos dados;
- O teste de significância dos coeficientes, que verifica se os coeficientes linear e angular da reta são diferentes de zero. Se os coeficientes forem significativos, indica que há uma relação linear entre as variáveis. Se os coeficientes não forem significativos, isso significa que não há evidência de uma relação linear entre as variáveis;
- A análise dos resíduos verifica se os resíduos seguem uma distribuição normal com média zero e variância constante. Se os resíduos não seguirem essas características, isso pode indicar que o modelo linear não é adequado para os dados ou que há algum problema nos dados, como *outliers* ou heterocedasticidade;
- **A matriz de covariância dos coeficientes mede a precisão dos coeficientes estimados pelo método:** quanto menor for a covariância entre os coeficientes, mais confiáveis serão as estimativas;
- O teste F verifica se o modelo linear é significativo como um todo, comparando a variância explicada pelo modelo com a variância não explicada. Se o teste F for significativo, isso significa que o modelo linear é melhor do que um modelo nulo, que não usa nenhuma variável explicativa;
- O teste de Durbin-Watson verifica se há autocorrelação nos resíduos, ou seja, se os resíduos estão relacionados entre si de forma sequencial. Se houver autocorrelação nos resíduos, isso pode indicar que o modelo linear não captura toda a informação dos dados ou que há algum problema nos dados, como tendência ou sazonalidade.

Existem outros critérios que podem ser usados dependendo do contexto e do objetivo da análise. O importante é sempre verificar se o método dos mínimos quadrados está produzindo

resultados confiáveis e consistentes com os dados, e se o modelo linear é o mais adequado para representar a relação entre as variáveis.

Leitura

Ordinary Least Squares Regression

Exemplo visual no gráfico do site.

Clique no botão para conferir o conteúdo.

ACESSE

Podemos ver o porquê dos quadrados, e que nos indica a possibilidade de erro entre o dado e a reta: quanto menor o quadrado, menor é o erro da reta.

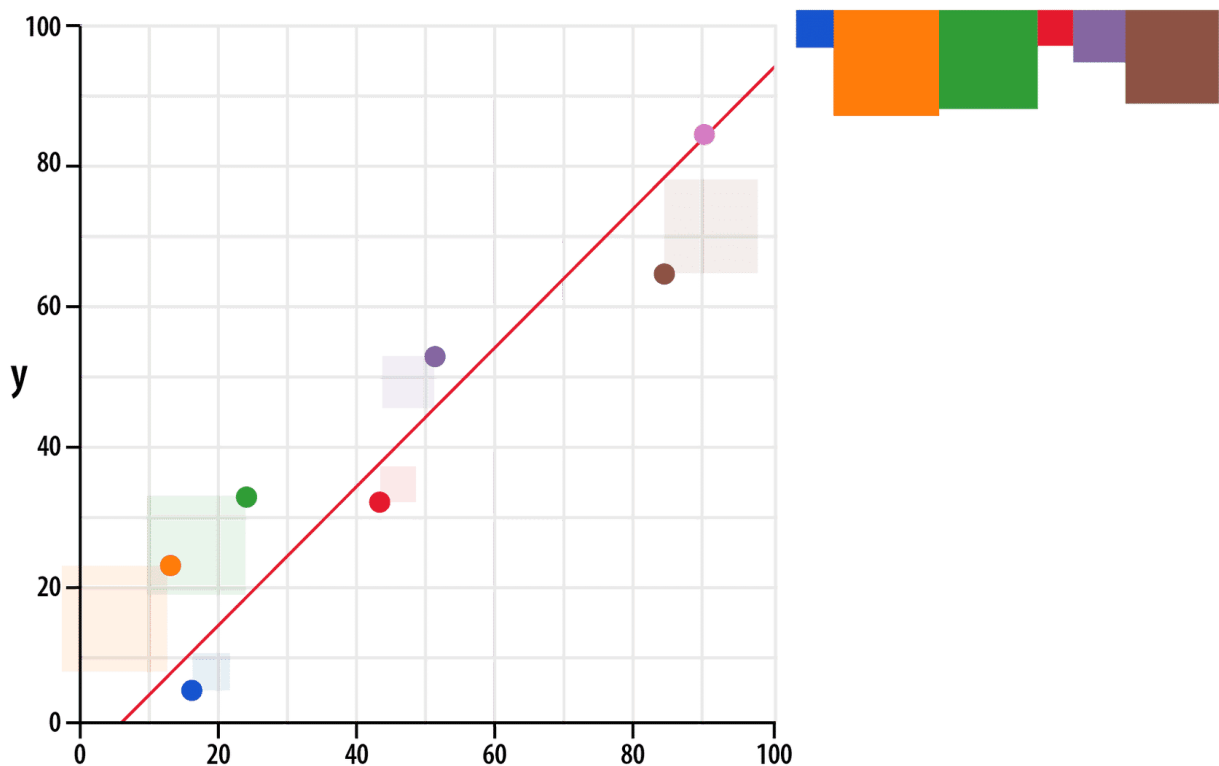


Figura 8 – Regressão linear e quadrados

Fonte: Reprodução

#ParaTodosVerem: gráfico mostra o eixo (x) e o eixo (y), sendo que o y sempre deve ser adicionado na vertical e o x na horizontal. Ao fundo, apresenta uma tela quadriculada, onde os pontos de intersecção estão marcados e de cada ponto marcado saem representação de quadrados até o ponto da reta. O gráfico apresenta que quanto maior a distância do ponto da reta, menor é o quadrado; e quanto mais próximo do ponto da reta, menor é o quadrado. Na lateral esquerda do gráfico, consta a representação da diferença do tamanho de cada quadrado para a reta final. Fim da descrição.

Como Encontrar a Equação da Reta de Regressão: Cálculo do Coeficiente Angular e da Interceptação em y

Para encontrar a equação da reta de regressão, você precisa calcular o coeficiente angular e a interceptação em y da reta que melhor se ajusta aos dados. A equação da reta de regressão tem a forma $y = ax + b$ onde a é o coeficiente angular e b é a interceptação em y.

A fórmulas:

$$a = \frac{n \cdot \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$
$$b = \bar{y} - a\bar{x}$$

Com estes dados podemos calcular e chegar na fórmula da reta:

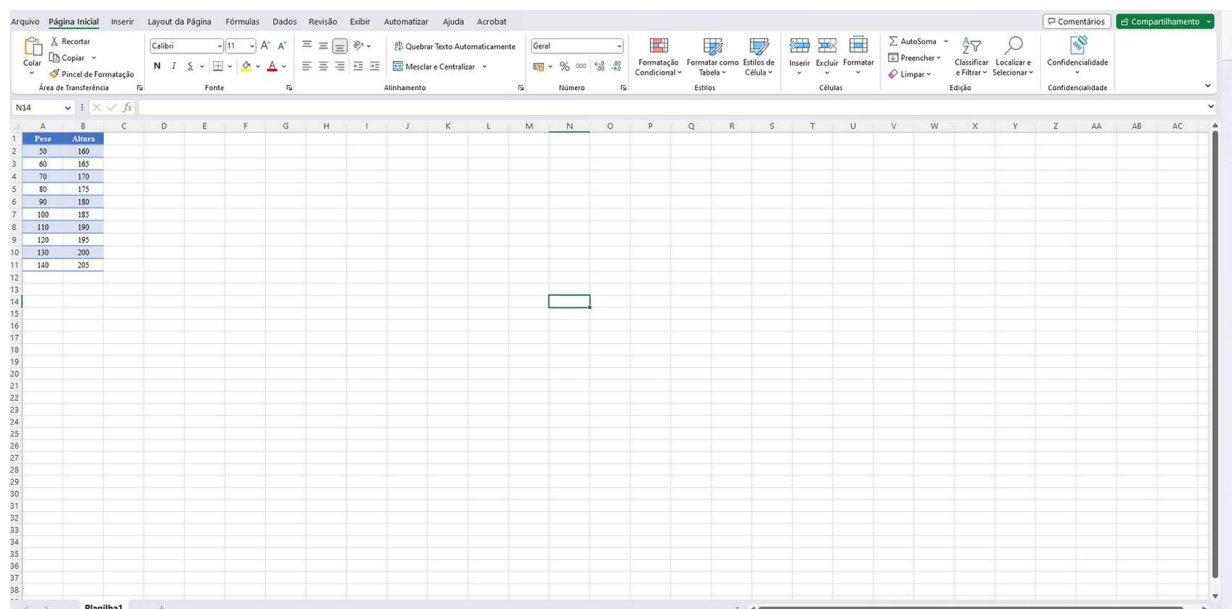
$$y = ax + b$$

Usando o Excel

Usando o *Excel*, podemos encontrar a equação da reta linear e o R, de uma forma simples, seguindo estes passos:

1

Crie uma tabela com os dados das variáveis dependente e independente.



The screenshot shows the Microsoft Excel interface with the 'Planilha1' worksheet. A table is created with two columns: 'Peso' and 'Altura'. The data is as follows:

| | Peso | Altura |
|----|------|--------|
| 1 | 50 | 160 |
| 2 | 60 | 165 |
| 3 | 70 | 170 |
| 4 | 80 | 175 |
| 5 | 90 | 180 |
| 6 | 100 | 185 |
| 7 | 110 | 190 |
| 8 | 120 | 195 |
| 9 | 130 | 200 |
| 10 | 140 | 205 |

Figura 9 – Tabela com os dados das variáveis dependente e independente

Fonte: Reprodução

#ParaTodosVerem: planilha do *Excel* aberta onde nas colunas A e B temos medidas de altura e peso, da linha 2 a linha 11. Fim da descrição.

2

Selecione toda a tabela e, depois, na aba “Inserir”, escolha o gráfico de dispersão;

• **Passo:** Inserir > Gráficos > Dispersão.

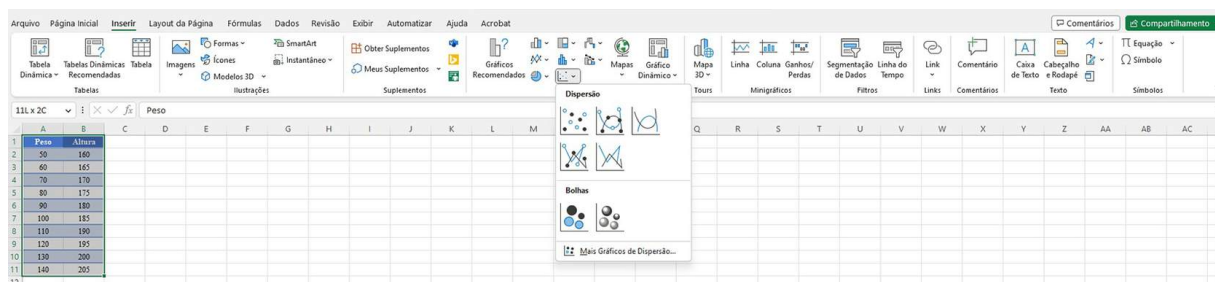


Figura 10 – Escolhendo o gráfico de dispersão

Fonte: Reprodução

#ParaTodosVerem: tabela do *Excel* com células selecionadas, aberta na guia inserir, na aba de gráficos, selecionando a opção dispersão. Fim da descrição.

3

Selecione o gráfico e clique na guia *Design* > Adicionar elemento gráfico > Linha de tendência > Linear;

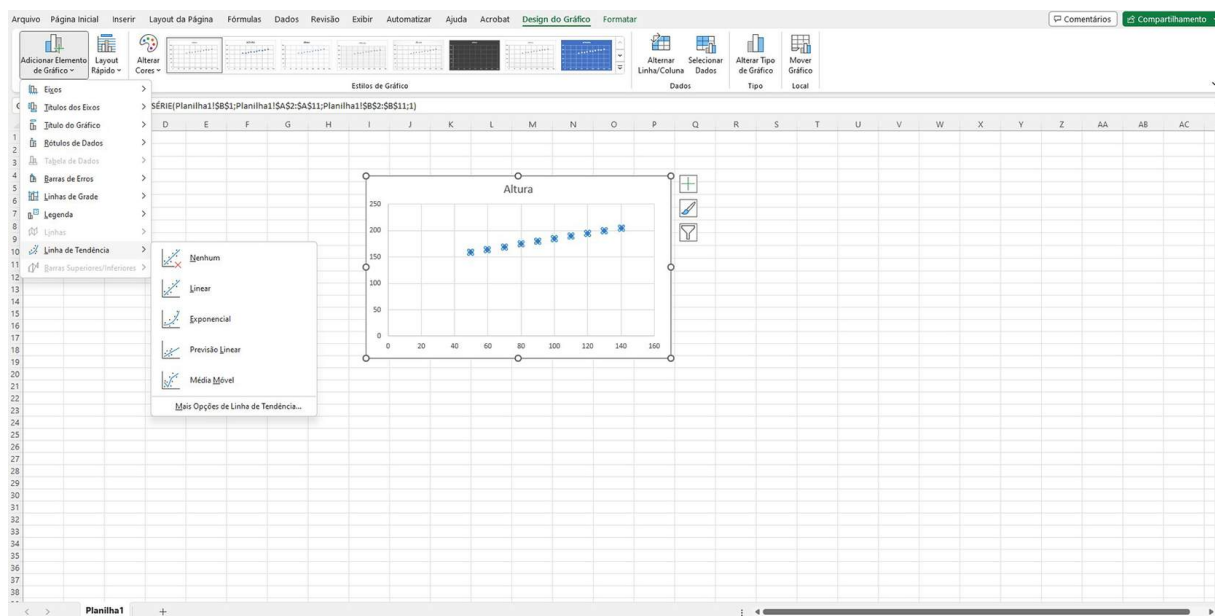


Figura 11 – Adicionando elemento gráfico

Fonte: Reprodução

#ParaTodosVerem: planilha do *Excel* aberta na guia “*Design do Gráfico*”, na guia “*Adicionar elemento do gráfico*”, com o cursor na opção “*linha de tendência*”, selecionando a opção “*Linear*”. Fim da descrição.

4

Marque a opção para exibir a equação e o R^2 no gráfico.

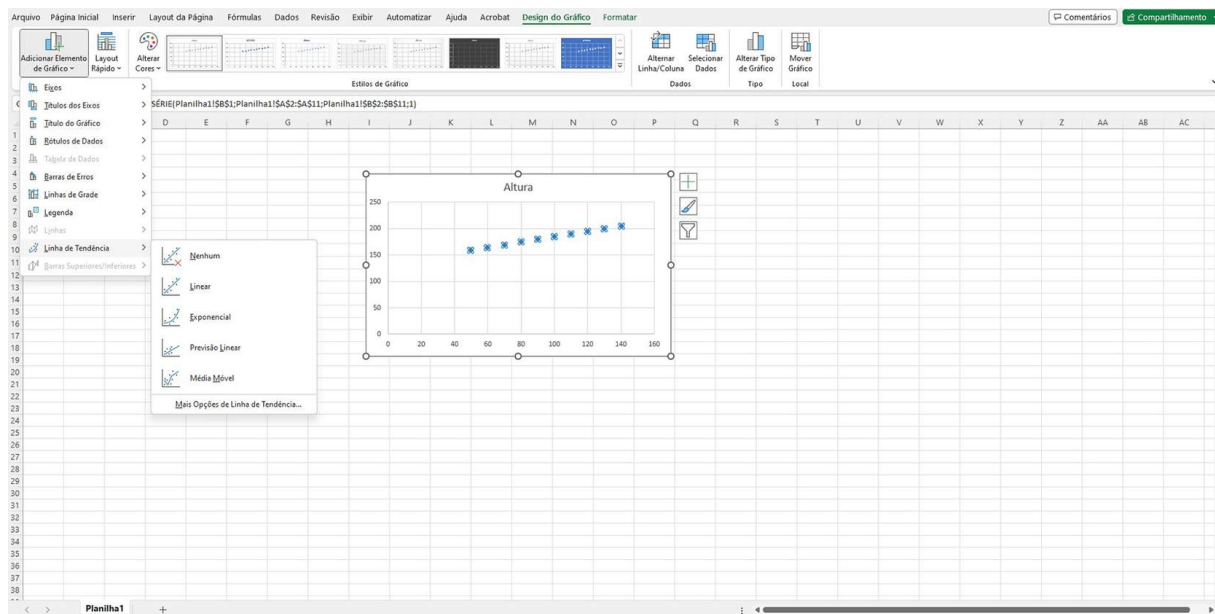


Figura 12 – Exibindo a equação e o R2 no gráfico

Fonte: Reprodução

#ParaTodosVerem: planilha de *Excel* exibindo caixa de texto da equação. Fim da descrição.

A equação da reta de regressão será do tipo $y = ax + b$, onde a é o coeficiente angular e b é a interceptação em y .

Saiba Mais

Pode usar a ferramenta **Análise de Dados do Excel** para obter mais informações sobre a regressão linear, como os intervalos de confiança e os testes de hipóteses. Para isso, você precisa habilitar a **Análise de**

Dados na guia Dados > Análise de Dados. Depois, você pode escolher a opção Regressão e preencher os campos solicitados.

Por que Fazer Previsões Usando a Reta de Regressão, Intervalos de Confiança, Teste de Hipóteses?

Ao realizar previsões, é importante ter em mente que podemos utilizar intervalos de confiança e testes de hipóteses.

Intervalos de confiança são valores dentro da qual há previsão é esperada com determinado nível de confiança. Normalmente, é utilizado a taxa de erro (variabilidade dos dados) e a estimativa pontual da reta de regressão. Larson e Farber (2006, p. 277) definem a “estimativa pontual como um valor único estimado para um parâmetro populacional μ é a média amostral \bar{x} ”.

Teste de hipóteses: em regressão linear, um teste de hipóteses pode ser usado para testar se os coeficientes da reta de regressão são diferentes de zero. Se um coeficiente for estatisticamente significativo, isso indica que a variável independente tem uma relação significativa com a variável dependente e pode ser útil na previsão.

Podemos entender que tanto os intervalos de confiança quanto os testes de hipóteses desempenham um papel importante na obtenção de previsões confiáveis usando a reta de regressão. Os intervalos de confiança fornecem uma medida da incerteza em torno da previsão, enquanto os testes de hipóteses ajudam a avaliar a significância estatística das variáveis independentes na previsão.

Estudo de Caso

O prefeito estava preocupado com o futuro da sua cidade.

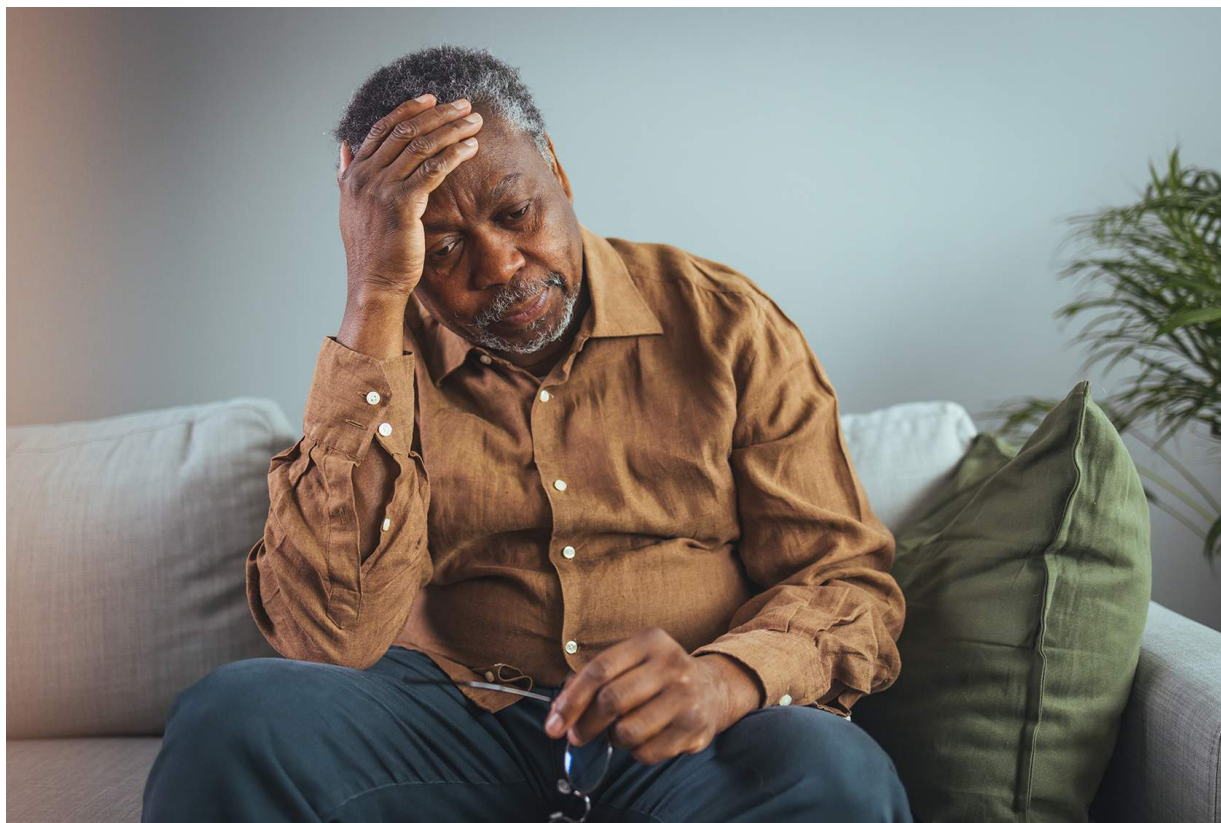


Figura 13 – Preocupação

Fonte: Getty Images

O homem apresentado na Figura 13 está observando o gráfico sobre o consumo de energia de uma cidade, pois o verão estava chegando e que as temperaturas poderiam chegar a 45 graus, o que aumentaria muito o consumo de energia dos aparelhos de ar-condicionado.

Ele tinha em mãos uma tabela que mostrava a relação entre a temperatura e o consumo de energia na cidade:

Tabela 1 – Temperatura e consumo de energia

| Temperatura (°C) | Consumo (KWh) |
|---------------------|------------------|
| 20 | 118 |
| 22 | 125 |
| 24 | 140 |
| 26 | 153 |
| 28 | 162 |
| 30 | 175 |
| 32 | 187 |

Teme-se de que se a temperatura chegar a 45 graus e que o consumo de energia pode causar um apagão na rede elétrica. O prefeito não quer que isso aconteça, pois prejudicaria a qualidade de vida dos seus eleitores e a sua popularidade.

Observando o gráfico, podemos perceber que há uma tendência de aumento do consumo de energia elétrica à medida que a temperatura aumenta.

Como podemos identificar qual a trajetória do aumento para que o prefeito possa tomar a melhor decisão?

Para quantificar essa relação, podemos usar a regressão linear simples para encontrar a equação da reta que melhor se ajusta aos pontos.

A equação tem a forma:

$$y = Ax + B$$

Onde y é o consumo de energia elétrica, x é a temperatura ambiente, a é o coeficiente angular da reta (que indica a inclinação) e b é a interceptação da reta com o eixo y (que indica o valor de quando x é zero).

Para encontrar os valores de A e B , podemos usar o método dos mínimos quadrados, que consiste em minimizar a soma dos quadrados das distâncias verticais entre os pontos e a reta.

Tabela 2 – Relação entre temperatura (x) e consumo (y)

| Temperatura (x) | Consumo (y) | $x.y$ | x^2 |
|------------------------|--------------------|-------|-------|
| 20 | 118 | 2360 | 400 |
| 22 | 125 | 2750 | 484 |
| 24 | 140 | 3360 | 576 |
| 26 | 153 | 3978 | 676 |
| 28 | 162 | 4536 | 784 |
| 30 | 175 | 5250 | 900 |
| 32 | 187 | 5984 | 1024 |

| Temperatura (x) | Consumo (y) | $x.y$ | x^2 |
|------------------------|--------------------|----------------|---------------|
| $\Sigma 182$ | $\Sigma 1060$ | $\Sigma 28218$ | $\Sigma 4844$ |

A Tabela 2 mostra a relação entre temperatura (x) e consumo (y) para sete valores diferentes de x . Ela também mostra os produtos de x e y , e os quadrados de x . A Tabela nos permite ver como o consumo aumenta à medida que a temperatura aumenta, e como os produtos e quadrados mudam de acordo com este aumento. A Tabela tem quatro colunas e oito linhas, incluindo os cabeçalhos e as somas.

A primeira coluna mostra os valores de x , que variam de 20 a 32 em incrementos de 2. A segunda coluna mostra os valores correspondentes de y , que variam de 118 a 187. A terceira coluna mostra os produtos de x e y , que variam de 2360 a 5984. A quarta coluna mostra os quadrados de x , que variam de 400 a 1024. A última linha mostra as somas de cada coluna, que são 182, 1060, 28218 e 4844 respectivamente.

Ainda revela que há uma correlação positiva entre temperatura e consumo, pois ambos aumentam juntos. Ela também mostra que os produtos de x e y aumentam mais rápido do que os quadrados de x , pois a diferença entre eles fica maior. A Tabela será usada para calcular o coeficiente angular e o intercepto de uma equação linear que modela a relação entre temperatura e consumo.

Esse método resulta nas seguintes fórmulas:

$$a = \frac{n \cdot \sum x \cdot y - \sum x \cdot \sum y}{n \cdot \sum x^2 - (\sum x)^2}$$

$$b = \frac{\sum y}{n} - a \frac{\sum x}{n}$$

Onde n é o número de observações, $\sum x$ é a soma dos valores de x , $\sum y$ é a soma dos valores de y , $\sum xy$ é a soma dos produtos de x e y e $\sum x^2$ é a soma dos quadrados de x .

Aplicando essas fórmulas aos dados da tabela, obtemos:

$$a = \frac{(7.28218 - 182.1060)}{7.4844 - (182)^2}$$

$$a = \frac{197526 - 192920}{33908 - 33124}$$

$$a = \frac{4606}{784}$$

$$a = 5,875$$

Após calcular o a , que é o coeficiente angular da reta e que indica a inclinação, vamos calcular o valor de b :

$$b = \frac{\sum y}{n} - a \frac{\sum x}{n}$$

$$B = \frac{1060}{7} - 5,875 * \frac{182}{7}$$

$$B = \frac{1060}{7} - 5,875 * \frac{182}{7}$$

$$B = 151,4286 - 5,875 * 26$$

$$B = 151,4286 - 152,75$$

$$B = -1,32143$$

Portanto, a equação da reta é:

$$y = 5,875.x - 1,32143$$

Podemos traçar essa reta no gráfico de dispersão para verificar o ajuste.

Adicionar um Gráfico com Reta

A partir da equação da reta, podemos fazer previsões sobre o consumo de energia elétrica para diferentes valores de temperatura.

Por exemplo, se quisermos saber qual seria o consumo esperado para uma temperatura de 23°C, basta substituir x por 23 na equação:

$$y = 5,875.(23) - 1,32143$$

$$y \approx 132$$

Calcular o coeficiente de determinação (R^2), que mede o grau de explicação da variável dependente pela variável independente.

O R^2 varia entre 0 e 1, sendo que quanto mais próximo de 1, melhor é o ajuste da reta aos dados.

Para calcular o R^2 : Encontraremos a equação da reta que se ajusta aos dados usando o método dos mínimos quadrados.

Já fizemos isso antes e encontramos a seguinte equação:

$$y = 5,875.x - 1,32143$$

Essa equação permite estimar o valor de y para cada valor de x .

Outro exemplo: para $x = 20$ temos:

$$y = 5,875.(20) - 1,32143$$

$$\hat{y} = 117,15 - 1,32143$$

$$\hat{y} = 115$$

Esse é o valor estimado de y pela reta quando x é 20.

Vamos fazer o mesmo para os outros valores de x e obter os seguintes valores estimados de y :

Tabela 3 – Soma dos Quadrados dos Resíduos

| x | y | \hat{y} | $(y - \hat{y})^2$ |
|--------------|-----|-----------|-------------------|
| 20 | 118 | 115,82 | 4,75 |
| 22 | 125 | 127,929 | 8,58 |
| 24 | 140 | 139,679 | 0,10 |
| 26 | 153 | 151,429 | 2,47 |
| 28 | 162 | 163,179 | 1,39 |
| 30 | 175 | 174,929 | 0,01 |
| 32 | 187 | 186,679 | 0,10 |
| SQRes: 17,40 | | | |

Agora, podemos calcular a soma dos quadrados dos resíduos (SQRes), que é a soma dos quadrados das distâncias verticais entre os pontos e a reta. Para isso, precisamos subtrair o valor observado de y pelo valor estimado de y e elevar ao quadrado.

Por exemplo, para o primeiro ponto, temos:

$$(y - \hat{y})^2 = (118 - 115,82)^2$$

$$(y - \hat{y})^2 = (2,18)^2$$

$$(y - \hat{y})^2 = 4,7524$$

Faremos o mesmo para os outros pontos e somar todos os resultados.

Vamos obter:

$$SQRes = \Sigma(y - \hat{y})^2$$

$$SQRes = 4,7524 + 8,57 + 0,10 + 8,58 + 0,10 + 2,47 + 1,39$$

$$SQRes = 17,40$$

Esse é o valor da soma dos quadrados dos resíduos: $SQRes$ 17,40.

Apenas precisamos encontrar a Soma dos Quadrados Totais (SQT), que é a soma dos quadrados das distâncias verticais entre os pontos e a média de y .

Para isso, precisamos calcular a média de y usando a seguinte fórmula:

$$\bar{y} = \Sigma y / n$$

Onde n é o número de observações.

Usando os valores da tabela, teremos:

$$\bar{y} = \Sigma y / n$$

$$\bar{y} = (118 + 125 + 140 + 153 + 162 + 175 + 187) / 7$$

$$\bar{y} = 1060 / 7$$

$$\bar{y} = 151,43$$

Esse é o valor da média de y .

Agora, você pode calcular a soma dos quadrados totais (SQT), que é a soma dos quadrados das distâncias verticais entre os pontos e a média de y .

Então, iremos subtrair o valor observado de y pela média de y e elevar ao quadrado.

Por exemplo, para o primeiro ponto, temos:

$$(y - \bar{y})^2 = (118 - 151,43)^2$$

$$(y - \bar{y})^2 = (-33,43)^2$$

$$(y - \bar{y})^2 = 1117,62$$

Podemos fazer o mesmo para os outros pontos e somar todos os resultados.

Obteremos:

$$SQT = \sum (y - \bar{y})^2$$

$$SQT = 117,47 + 698,47 + 11,43 + 130,61 + 2,47 + 111,76 + 555,61 + 1265,33$$

$$SQT = 3881,71$$

Esse será o valor da soma dos quadrados totais.

E, finalmente, já podemos calcular o R^2 usando a seguinte fórmula:

$$R^2 = 1 - (SQRes / SQT)$$

Substituindo os valores que encontramos:

$$R^2 = 1 - (17,40 / 3881,71)$$

$$R^2 = 1 - 0,00448$$

$$R^2 = 0,99552$$

Esse é o valor do R^2 para essa regressão linear. Indica que a reta explica cerca de 99,55% da variação dos dados. Podemos analisar que este valor de R^2 é relativamente alto, o que significa que a reta se ajusta muito bem aos dados, pois um valor de R^2 muito baixo indicaria uma pior correlação linear entre as variáveis.

No entanto, o valor de R^2 , por si só, não é suficiente para avaliar a qualidade de uma regressão linear, pois também precisamos considerar outros fatores, como o erro padrão da regressão, os intervalos de confiança, os testes de hipóteses e a análise de resíduos.

Além disso, o valor de R^2 depende do contexto e do propósito da análise. Às vezes, um valor de R^2 baixo pode ser aceitável se a reta ainda for útil para fazer previsões ou testar teorias.



Material Complementar

Indicações para saber mais sobre os assuntos abordados nesta Unidade:

Vídeo

Como Fazer Regressão Linear em *Python* (*Statsmodels* e *Scikit-learn*)

Leitura

Como Fazer Regressão Linear no *Excel*

Clique no botão para conferir o conteúdo.

ACESSE

IBM: Regressão Linear

Clique no botão para conferir o conteúdo.

ACESSE

AWS: O que é Regressão Linear?

Clique no botão para conferir o conteúdo.

ACESSE



Referências

AULA 3.6 | Regressão Linear Simples. SELAU, L. [S.l.]. 21/10/2020. 1 vídeo (14 min.). Publicado pelo canal Probabilidade e Estatística UFRGS. Disponível em: <<https://www.youtube.com/watch?v=PvHqrfKPYS0>>. Acesso em: 04/08/2023.

BONAFINI, F. C. **Estatística**. São Paulo: Pearson, 2012. (*e-book*)

CHEIN, F. **Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas**. Brasília – DF: ENAP, 2019.

CRESPO, A. A. **Estatística fácil**. 19. ed. São Paulo: Saraiva, 2009. (*e-book*)

FABER, L. **Estatística aplicada**. 4. ed. São Paulo: Pearson, 2010. (*e-book*)

LARSON, R.; FARBER, E. **Elementary Statistics**. [S.l.]: Prentice Hall, 2006.

LOCK, R. H. **Estatística: revelando o poder dos dados**. Rio de Janeiro: LTC, 2017.

MCCLABE, J. T.; BENSON, P. G.; SINCICH, T. **Estatística para administração e economia**. São Paulo: Pearson, 2009. (*e-book*)

MOORE, D. S. **A estatística básica e sua prática**. 7. ed. Rio de Janeiro: LTC 2017. (*e-book*)

NEUFELD, J. L. **Estatística aplicada à Administração usando Excel**. São Paulo: Pearson, 2003. (e-book)

REGRESSÃO linear simples. CARRERA, J. M. [S.l.]. 19/12/2020. 1 vídeo (221 min.). Publicado pelo canal Finanças 101. Disponível em: <<https://youtu.be/tsj2zBbYdVY>>. Acesso em: 26/04/2023.

REVISÃO sobre regressão linear. *Khan Academy*. [S.d.]. Disponível em: <<https://pt.khanacademy.org/math/statistics-probability/describing-relationships-quantitative-data/introduction-to-trend-lines/a/linear-regression-review>>. Acesso em: 04/08/2023.