Faculdade XPe

Capítulo 06

ESTATÍSTICA PARA CIÊNCIA DE DADOS

Arnaldo Satoru Gunzi

2024

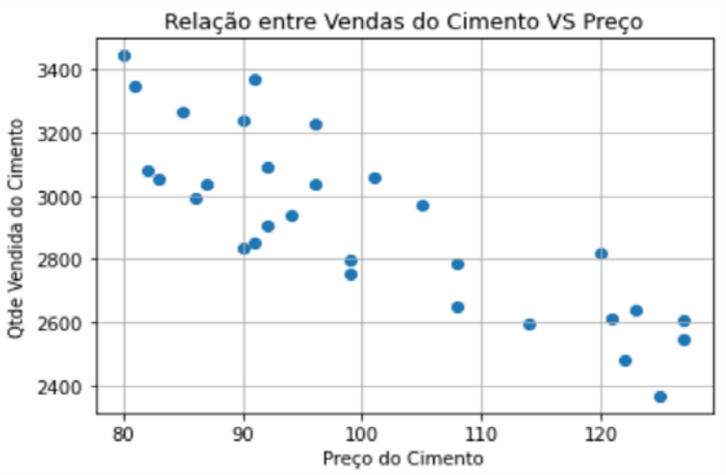
Regressão Linear Simples

A Regressão Linear permite gerar um modelo matemático através de uma reta que explique a relação entre variáveis. No caso mais simples, teremos a relação entre uma variável explicativa X e uma variável resposta Y:

$$\hat{Y} = \beta_0 + (\beta_1 * X_1) + \varepsilon$$

Onde β_0 é o termo de intercepto (em outras palavras, é o valor de Y quando X = 0). β_1 é a inclinação da reta, representa a mudança média prevista em y resultante do aumento de uma unidade em X. ε é um termo erro aleatório com média μ zero e variância σ^2 constante.
Suponha que temos dados sobre o preço do cimento e suas vendas.

Figura 15 - Relação entre preço e vendas

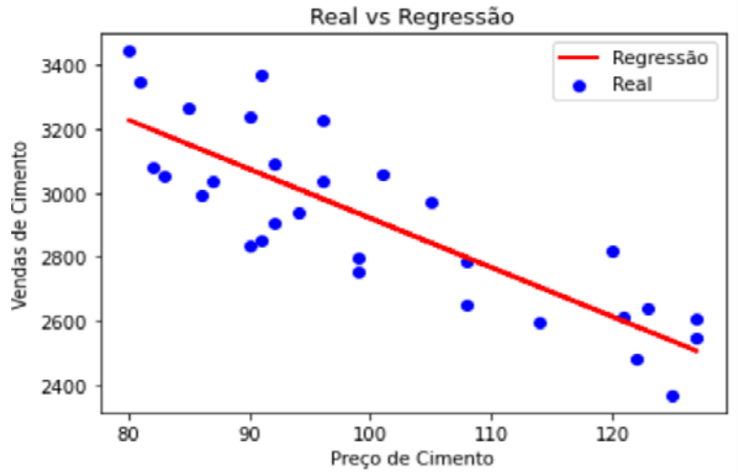


Existe relação entre o preço do cimento e a quantidade vendida? Como o preço explica as vendas? Posso utilizar o preço para prever as vendas? Visualmente, dá para estabelecer uma relação entre ambas (quanto maior o preço, menor a quantidade vendida). A regressão linear é a que fornece a melhor reta que descreve tal relação, baseada na minimização do erro quadrático entre a reta e os dados disponíveis.

Regressão Linear Simples e Regressão Linear Múltipla

A regressão linear simples é quando temos apenas uma variável preditora e uma variável resposta. É o exemplo do preço e suas vendas. Nosso modelo considera apenas o preço para explicar e prever a variação nas vendas. Vamos ajustar uma reta de regressão para modelarmos estatisticamente a relação entre o preço do cimento e a quantidade de vendas.

Figura 16 - Ajustando uma regressão linear entre o preço do cimento e suas vendas



Conforme já analisamos anteriormente, a correlação entre as duas variáveis é negativa, ou seja, quando uma aumenta (preço) a outra diminui (vendas). Então, nossa reta é decrescente. Temos também a equação de primeiro grau que descreve a reta, ou seja, descreve a relação entre o preço e as vendas.

$$\text{Vendas de Cimento} = 458{,}13 + (-15{,}372 * \text{Preço do Cimento})$$

Onde 458,13 é o intercepto B0 e -15,372 é o coeficiente angular B1. Ou seja, a cada real que aumenta no preço, as vendas caem, em média, em 15,372 unidades.

Regressão linear múltipla

Voltando ao nosso exemplo de vendas do cimento em função de seu próprio preço, podemos incluir mais uma variável em nosso estudo. Suspeita-se que na nossa empresa, o PIB do país no momento da venda também influencie. Portanto, estamos trabalhando com três variáveis, pois temos duas variáveis preditoras (preço do cimento, PIB) e uma variável resposta (quantidade vendida do cimento). Ou seja, um modelo de regressão linear múltipla.
A regressão linear múltipla pode comportar p variáveis preditoras ao invés de somente uma:

$$\hat{Y} = \beta_0 + (\beta_1 * X_1) + (\beta_2 * X_2) + \cdots + (\beta_p * X_p) + \varepsilon$$

A equação de regressão múltipla ficou:

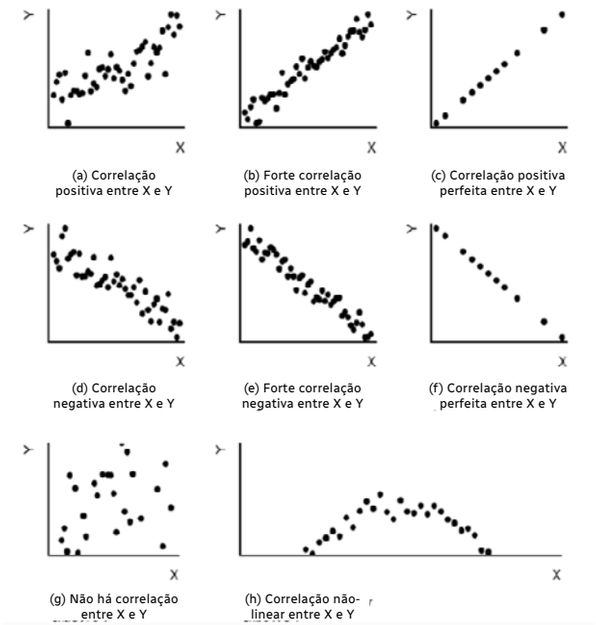
$$\text{Vendas do cimento} = 4238{,}31 + (-16{,}16 * \text{Preço}) + (200{,}78 * \text{PIB})$$

Observe que o coeficiente do Preço do Cimento no modelo múltiplo ficou ligeiramente diferente do coeficiente no modelo simples. Isso é comum, pois adicionamos mais uma variável e os coeficientes são estimados em conjunto.

Correlação e causalidade

Para medir a força da correlação entre duas variáveis quantitativas, pode ser utilizado o coeficiente de correlação de Pearson.
O coeficiente de correlação de Pearson é um valor numérico que vai de -1 a 1. Quanto mais próxima de 1, mais forte será a correlação entre as duas variáveis de forma positiva (quando uma variável aumenta a outra também aumenta). Quanto mais próximo de -1 for a correlação, mais forte será a correlação de forma negativa (quando uma variável aumenta a outra diminui).

Figura 17 - Analisando a correlação pelo gráfico de dispersão



Para mensurar a força da correlação entre um par de variáveis, podemos calcular o coeficiente de correlação linear de Pearson com a seguinte equação:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left[\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]\left[\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}\right]}$$

Os pacotes computacionais atuais já calculam o valor mencionado, então não nos preocuparemos com a fórmula.
O coeficiente de correlação (denotado pela letra grega rho) entre o preço do cimento e suas vendas é ρ=-0,83. Vamos utilizar a tabela abaixo para nos orientar na interpretação desse valor.

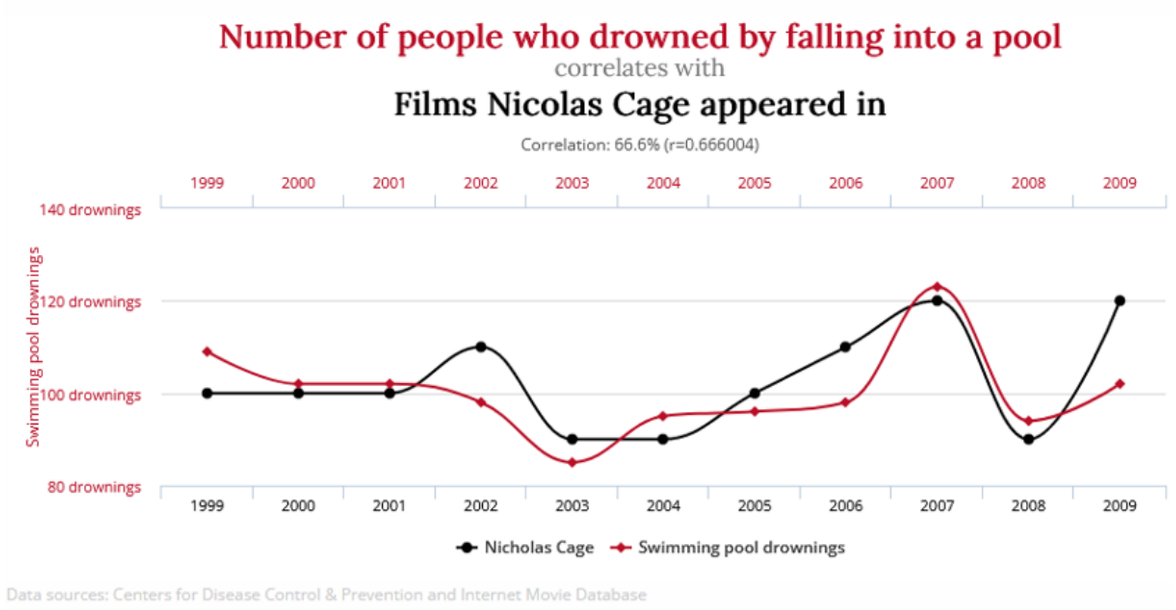
Figura 18 - Interpretando valores do coeficiente de correlação

Valor do Coeficiente	Interpretação
Entre 0,10 e 0,29	Correlação positiva fraca
Entre 0,30 e 0,49	Correlação positiva moderada
Entre 0,5 e 1	Correlação positiva alta
Entre -0,10 e -0,29	Correlação negativa fraca
Entre -0,30 e -0,49	Correlação negativa moderada
Entre -0,5 e -1	Correlação negativa alta

Fonte: COHEN, Multiple Commitment in the workplace: an integrative approach, 2003

Um cuidado é que Correlação não implica causalidade: Só porque duas variáveis estão correlacionadas não significa que uma causa a outra.
Exemplo clássico: Existe uma correlação entre o número de pessoas que se afogam em piscinas e o número de filmes estrelados por Nicolas Cage em um ano. Claramente, Nicolas Cage não causa afogamentos; isso é apenas uma coincidência.

Figura 19 - Numero de pessoas que se afogam em piscinas x número de filmes estrelados por Nicolas Cage



https://www.tylervigen.com/spurious-correlations

Explicação vs Predição

Podemos utilizar a regressão linear para explicação dos dados passados, conforme foi colocado.
Outro fator bastante positivo da regressão linear é que o algoritmo nos traz informações inferenciais, ou seja, podemos extrapolar conclusões para a população a partir da amostra. Portanto, podemos utilizá-lo para modelagem preditiva.

Diagnóstico do Ajuste do Modelo de Regressão Linear

Ao ajustar um modelo de regressão linear sobre um conjunto de dados, duas métricas importantes são o R^2 (R quadrado) e o R^2 ajustado. O R^2 é uma medida percentual, ou seja, varia de zero a um, e nos diz o quanto da variação da variável resposta o modelo ajustado explica. O R2 também é chamado de coeficiente de determinação.

Figura 20 - Exemplo de R^2 e R^2 ajustado

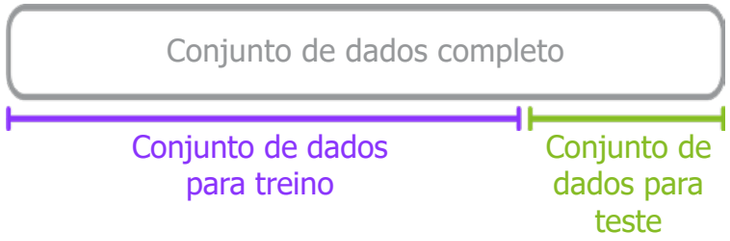
```
R-squared: 0.753
R2 ajustado: 0.735
```

O R^2 obtido foi de 0,753 (75,3%). Podemos interpretar que o modelo ajustado com esses preditores, consegue explicar 75,3% das vendas do cimento. A parte não explicada é devido a fatores aleatórios e a variáveis não inclusas no modelo. Resumindo, quanto maior o R^2, melhor.
O R^2 ajustado é uma métrica interessante para comparar modelos, pois ela penaliza a adição de novas variáveis preditoras. Sempre que adicionarmos uma nova variável preditora no modelo, o R^2 irá aumentar automaticamente. Entretanto, se essa nova variável não contribuir de forma significativa para o modelo, o R^2 ajustado irá diminuir.

Overfitting

A fim de utilizar o modelo para realizar predições, é extremamente aconselhado separar um percentual do conjunto de dados para ajustar a regressão e testar o modelo num outro percentual, que ficou de fora do treino. Esse procedimento é chamado de hold-out e é importante para evitar o overfitting, que é quando um modelo estatístico se super ajusta aos dados de treinamento apresentando um R^2 altíssimo, porém, quando chegam dados que ele ainda não conhece, o modelo tem dificuldade em acertar nas previsões.

Figura 21 - Metodologia hold-out



Pílulas de conhecimento

Aos alunos interessados na álgebra linear por trás do algoritmo: https://www.youtube.com/watch?v=TtY_ASnw60M&t=1s.

Outras metodologias e métricas pra validar a capacidade preditiva de um algoritmo podem ser acessadas no link: https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-machine-learning-tips-and-tricks#model-selection.

Estatística Computacional – Regressão Linear

Vide link a seguir para acompanhar os códigos para regressão linear com Python.
Link do Github: https://github.com/asgunzi/Estatistica_Analise_Dados

Faculdade XPe

xpeducacao.com.br