Regressão Linear: Uma Introdução

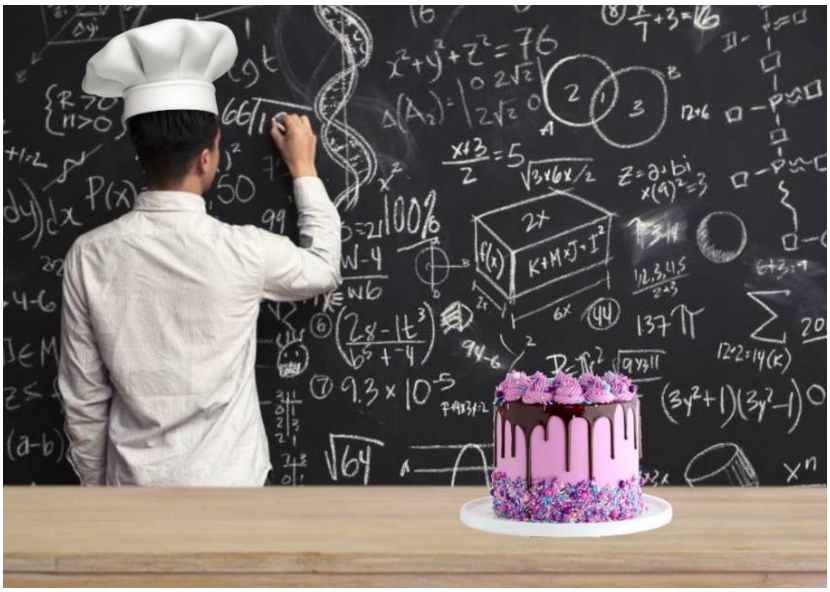
Olá, caro leitor!

Hoje vamos falar de um assunto de estatística que você, entusiasta de ciência de dados, com certeza já ouviu falar: regressão linear! É um nome bonito que pode causar curiosidade em alguns e arrepios em outros, mas que na verdade é uma técnica simples e bem interessante de análise de dados, então vem comigo e vamos aprender mais sobre ela!

Para introduzir o assunto, vamos imaginar um cenário hipotético:

Marcelo é um confeiteiro, ele abriu um pequeno comércio na sua cidade para vender os seus deliciosos bolos confeitados, mas com o passar do tempo ele percebeu que o seu comércio não estava sendo muito lucrativo, pois boa parte dos bolos que ele fazia não eram vendidos e ele acabava tendo que os jogar fora. O plot twist da história é que Marcelo também era estatístico e resolveu elaborar uma hipótese: e se eu conseguir relacionar o número de bolos vendidos em um dia com a quantidade de clientes que visitam a minha loja diariamente?

O que Marcelo estava prestes a fazer, meu caro leitor, era uma **regressão linear simples** entre a **variável dependente** número de bolos vendidos e a **variável independente** número de visitas a loja, vamos adiante.



Marcelo, então resolveu anotar diariamente quantos clientes entravam na loja e quantos bolos eram vendidos, e fez a tabela a seguir:

Visitas	Vendas
36	5
22	3
43	6
10	2
41	5
33	4
20	3
50	6
35	5
25	3

Podemos chamar os dados dessa tabela de **dados amostrais**.

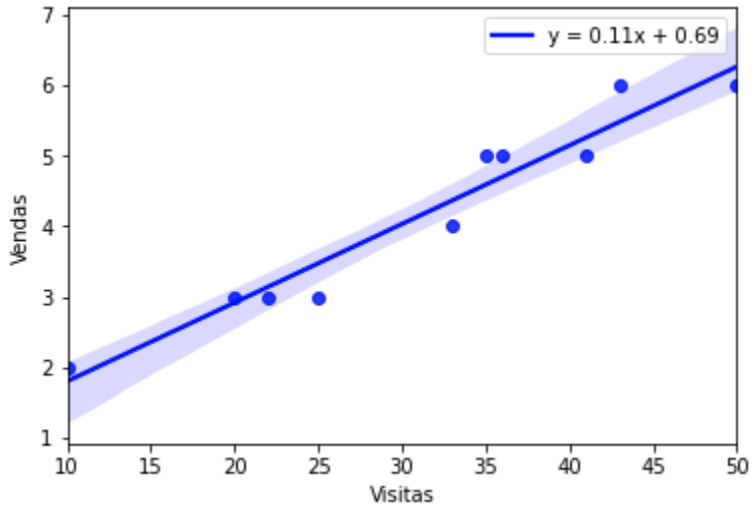
Depois de 10 dias Marcelo decidiu que tinha a quantidade de dados suficiente, colocou os dados em um arquivo csv, e fez o código em python a seguir:

12 lines (8 sloc) | 344 Bytes

```
1  import pandas as pd
2  import seaborn as sns
3  from scipy import stats
4
5  df = pd.read_csv('bolos_vendidos.csv')
6
7  slope, intercept, r_value, p_value, std_err = stats.linregress(df['Visitas'],df['Vendas'])
8
9  ax = sns.regplot(x="Visitas", y="Vendas", data=df, color='b',
10  line_kws={'label':"y = {0:.2f}x + {1:.2f}".format(slope,intercept)})
11
12 ax.legend()
```

O código acima utiliza as bibliotecas Pandas, Seaborn e Scipy. Na linha 5 é lido o csv com os dados amostrais utilizando o Pandas, na linha 7, é feita a regressão linear com a função linregress do Scipy e por último o gráfico é configurado, na linha 9, com a ajuda do Seaborn.

Com ele o gráfico abaixo foi gerado:

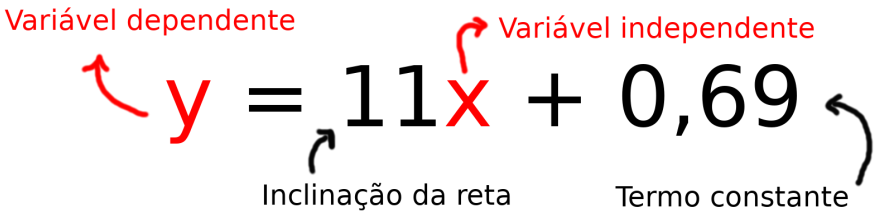


A reta que aparece no gráfico é uma **estimativa** do valor da variável dependente (y), dado uma variável independente (x), descrita pela equação: y = 0.11x + 0,69. Com essa equação Marcelo pode responder perguntas como:

- Se minha loja tiver 100 visitas em um dia, quantos bolos eu vou vender?
- Se o número de visitas à minha loja crescer 50% quantos bolos eu vou vender?

Claro que não é uma resposta exata, mas com isso Marcelo já conseguirá uma boa **aproximação** e consequentemente irá economizar e evitar bastante desperdício!

Vamos olhar mais de perto a equação gerada:



Além das já citadas variáveis dependente e independente, que são as incógnitas da equação, temos dois outros termos: a **inclinação da reta** e o **termo constante** (ou intercepto), esses são os **parâmetros da equação**, o papel da regressão linear simples é encontrar esses parâmetros de modo que a reta gerada passe o mais próximo possível de todos os pontos dos dados amostrais.

Você pode estar pensando: mas essa análise é muito básica e se eu quisesse uma análise mais sofisticada com mais fatores que possam afetar a quantidade de bolos vendidos?

Bom, esse caso podemos fazer uma **análise de regressão múltipla**, onde cada um dos fatores que afetam as vendas seriam variáveis independentes da equação associadas a um “peso”, e, em vez de uma reta como estimativa, teríamos um plano ou hiperplano dependendo da quantidade de variáveis independentes usadas.

Legal, né? Algumas outras aplicações para a regressão linear simples são:

- Estimar o salário de um funcionário baseado em anos de experiência.
- Prever o rendimento de uma colheita baseado na quantidade de chuva.

Entre muitos outros! Agora que aprendeu um pouco sobre regressão linear, consegue pensar em mais alguns?

Conclusão

No Turing Talks de hoje tivemos uma pequena introdução sobre regressão linear, técnica que será essencial na sua jornada em ciência de dados. Caso esteja pronto para dar o próximo passo e se aprofundar mais sobre o tema, sugiro o artigo Modelos de Predição | Regressão Linear.

Por fim, não deixe de acompanhar o Grupo Turing no Facebook, Linkedin, Instagram e, claro, nossos posts do Medium!

Bons estudos e até a próxima!

REFERÊNCIAS:

KUMARI, Riya. Simple Linear Regression: Applications, Limitations & Examples. **Analytics Steps**, 2020. Disponível em: <https://www.analyticssteps.com/blogs/simple-linear-regression-applications-limitations-examples>. Acesso em: 3 de junho de 2021.

RAMOS, Raniere. Regressão Linear Simples: O que é? Para que serve? Como funciona?. **O Estatístico**, 2020. Disponível em: <https://oestatistico.com.br/regressao-linear-simples/>. Acesso em: 3 de junho de 2021.

Seaborn: anotar a equação de regressão linear. CoreDump, 2017. Disponível em: <https://pt.coredump.biz/questions/45902739/seaborn-annotate-the-linear-regression-equation>. Acesso em: 3 de junho de 2021.