Prever eventos olhando para o passado

Como modelos de regressão linear podem os ajudar com este dilema?

Figura 1

Uma pequena introdução

Desde os primórdios da humanidade buscamos maneiras de olhar para o futuro com olhos acurados, de modo que possamos talvez entender os fenômenos climáticos, onde um planeta pode se encontrar em sua órbita dentro de alguns dias ou até mesmo quando aquela promoção do mercado irá voltar.

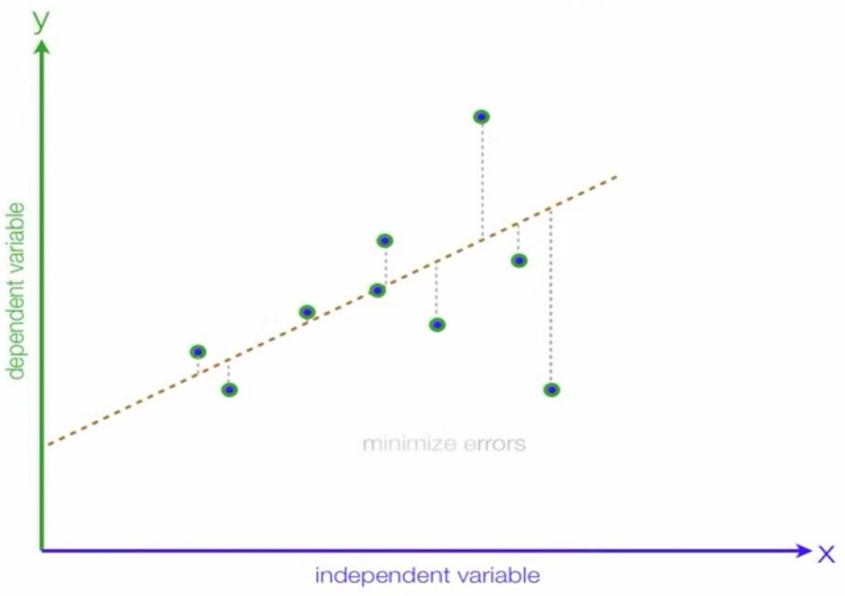
Os gregos com seus Oráculos de Delfos e as avós olhando os restos de café em xícaras propuseram seus modelos, porém quem trouxe para nós a realidade das previsões foram Adrien-Marie Legendre (1752 – 1833) e Johann Carl Friedrich Gauss (1777 – 1855) com o método dos mínimos quadrados (Least Squares Regression Line) – posteriormente este método foi amplamente divulgado e utilizado por Lambert Adolphe Jacques Quetelet (1796 – 1874).

O método dos Quadrados Mínimos

Imagine um plano, onde no eixo X possuímos diversos pontos que representam a variável independente e, no eixo Y temos a variável dependente.

O objetivo aqui é procurar uma reta de forma que trace um caminho onde sua equação $y = ax + b$ mais se aproxime dos valores já pontuados.

Cada reta traçada no plano possui uma certa distância entre os pontos, chamaremos esta distância de resíduo. O método apresentado por Legendre e Gauss consiste em somarmos o quadrado de cada resíduo e encontrar, deste modo, a reta que represente a menor soma possível.



A reta que representa a menor soma dos quadrados

Como a regressão linear pode nos ajudar?

Imagine que você pretende comprar um imóvel, deste modo, lhe interessa entender quais fatores mais afetam o preço deste. Pensando em diversos fatores, você acabou pensando em alguns nos quais acredita que sejam relevantes para seu objetivo, sendo eles: a localização, o tamanho do imóvel, se há garagem, proximidade de alguma estação metroviária e a quantidade de banheiros.

A partir de algumas regressões lineares, faz-se possível entender qual/quais delas afetam mais ou afetam menos a variável preço do imóvel, o que, obviamente, permite que você ache o imóvel com maior custo-benefício, além de prever se os imóveis que você pretende comprar se encontram dentro da faixa de preço daqueles que possuem suas mesmas características.

Implementação em Python

```
# Importando algumas bibliotecas para trabalharmos com elas
import numpy as np
import matplotlib.pyplot as plt
from sklearn.linear_model import LinearRegression
```

```
# Agora declararemos dois arrays de floats
x = np.array([1.32, 1.44, 2.21, 2.54, 1.13, 1.33, 1.78, 1.83, 1.99, 1.1, 2.1]).
 ↪reshape(-1, 1)

y = np.array([62.2, 43.2, 37.3, 58.6, 58.57, 64.15, 61,88, 58.11, 68.1, 61.3])
```

```
# Primeiro inicializamos um modelo, ainda não treinado.
model = LinearRegression()

# Após isso, vamos treiná-lo, isso significa que procuramos encontra a equação␣
 ↪da reta que relaciona melhor x e y
model.fit(x, y)
```

```
LinearRegression()
```

```
# Declarando o coeficiente de inclinação da reta
a = model.coef_

# Declarando o ponto de interceptação da reta com o eixo y
b = model.intercept_



print(f"y = {a}X + {b}")
```

```
# Declarando o coeficiente de inclinação da reta
a = model.coef_

# Declarando o ponto de interceptação da reta com o eixo y
b = model.intercept_



print(f"y = {a}X + {b}")
```

```
y = [-4.90319682]X + 68.4148185792788
```

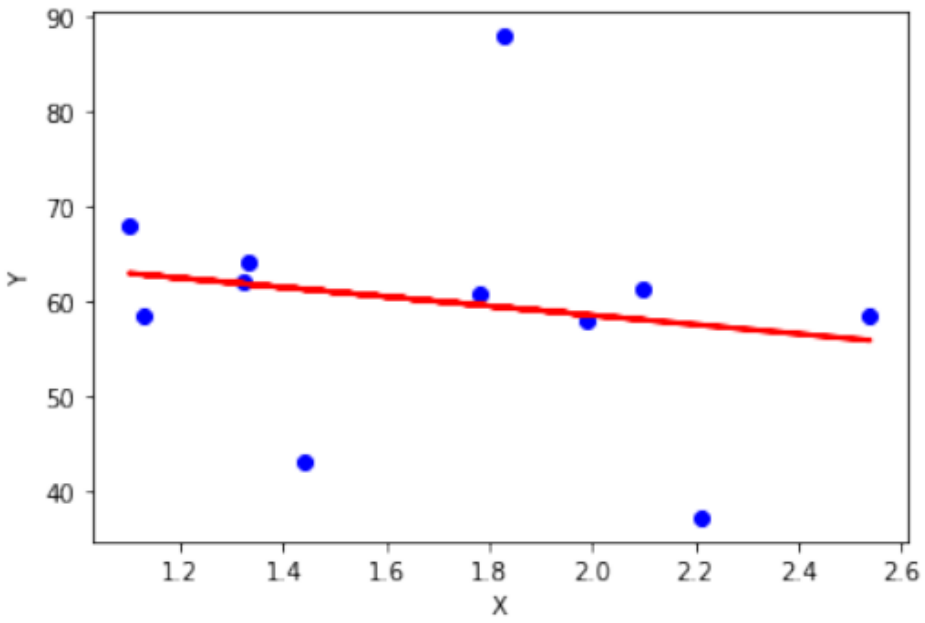
```
# Realizando a previsão de alguns pontos de Y com base em X
y_pred_model = model.predict(x)
print(y_pred_model)
```

```
[61.94259877 61.35421515 57.5787536  55.96069865 62.87420617 61.8935668
 59.68712823 59.44196839 58.6574569  63.02130207 58.11810525]
```

```
# Iremos então plotar o gráfico da regressão linear
plt.scatter(x, y,  color='blue')
plt.plot(x, model.predict(x), color='red', linewidth=2)

# Adicionando os títulos dos eixos
plt.xlabel("X")
plt.ylabel("Y")
```

```
Text(0, 0.5, 'Y')
```




```
# Por fim, gostaríamos de saber o quanto o nosso modelo conseguiu prever, para␣
 ↪isso usaremos algo chamado
# coeficiente de determinação ou R^2
coef_det = model.score(x, y)
print(f'Coeficiente de Determinação: {coef_det}')
```

```
Coeficiente de Determinação: 0.03244443008111253
```

Assim, percebe-se que o modelo foi pouco eficiente e explica apenas 3% de Y.

Referências:

Figura 1 - https://sdsclub.com/linear-regression-vs-multiple-regression-know-the-difference/

Figura 2 - https://www.youtube.com/watch?v=zPG4NjIkCjc&t=146s

T01E31 – Regressões Lineares (Lean Six Sigma for Experts) - https://open.spotify.com/episode/5VMI3uVphvbU0DBdp9mKp7?si=af08c896eb4a49c2

Como funciona uma regressão linear? - https://medium.com/data-hackers/como-funciona-uma-regressão-linear-f7208fa6c662