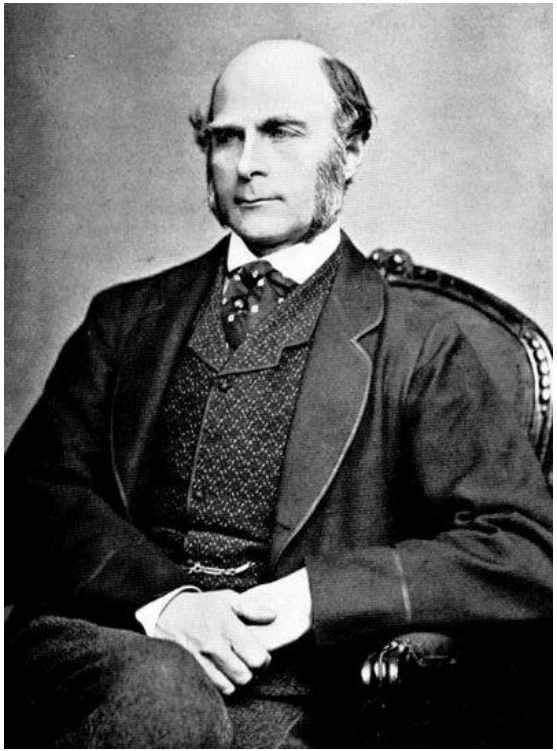
Introdução a Regressão Linear

Reading Assignment

Capítulos 2 & 3 de
Introduction to Statistical Learning
por Gareth James, et al.

História

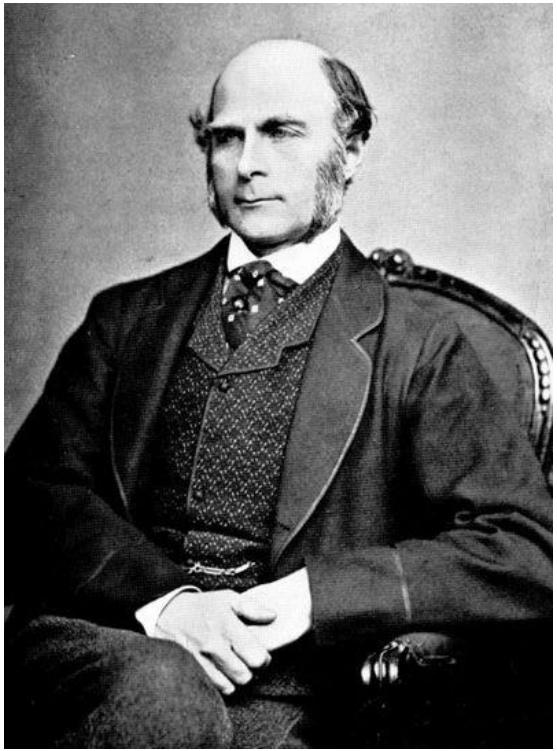
Tudo isso começou em 1800 com um cara chamado Francis Galton. Galton estava estudando a relação entre pais e filhos. Em particular, ele investigou a relação entre as alturas dos pais e seus filhos.



História

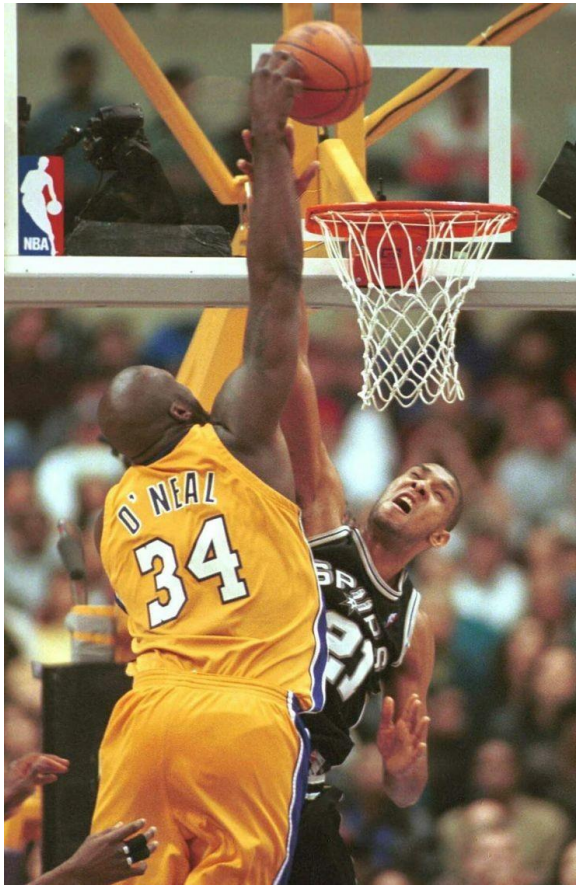
O que ele descobriu foi que o filho de um dado homem tendia a ser, mais ou menos, tão alto quanto seu pai.

No entanto, a grande descoberta de Galton foi que **a altura do filho tendia a ser mais próxima da altura média geral** de todas as pessoas.



Exemplo

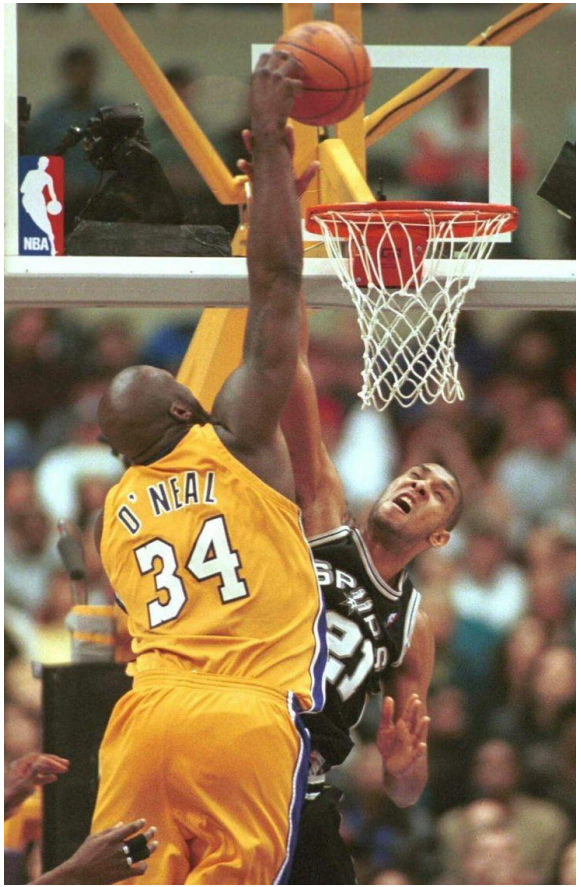
Tomemos Shaquille O'Neal como exemplo. Shaq é realmente alto: 2,2 metros.
Se Shaq tiver um filho, é provável que ele também seja bem alto.
No entanto, Shaq é uma anomalia tão grande que também há uma boa chance de que seu filho **não seja tão alto quanto ele mesmo.**



Exemplo

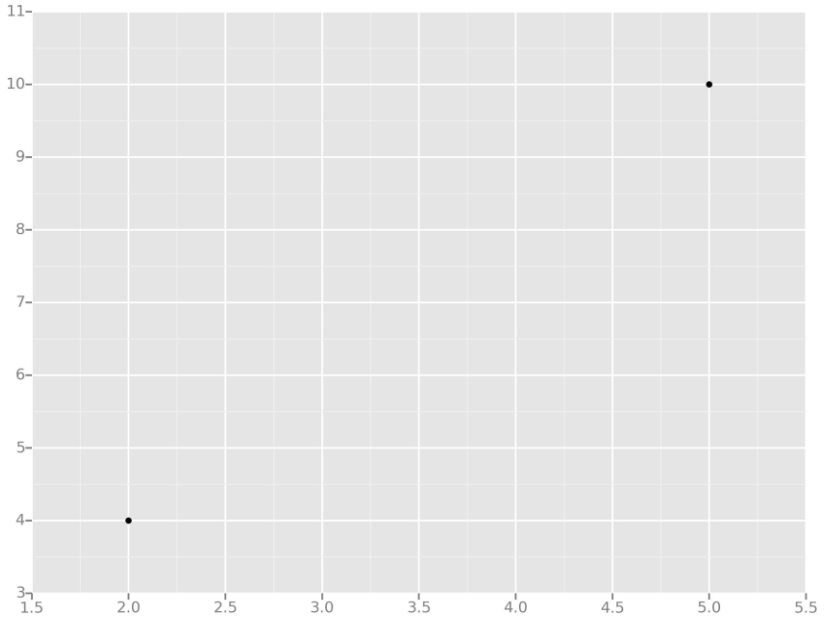
Acontece que este é o caso: o filho de Shaq é bem alto (2,0 m), mas não tão alto quanto seu pai.

Galton chamou esse fenômeno de **regressão**, como em "A altura do filho de um pai tende a regredir (ou se aproximar) da altura média (média)".



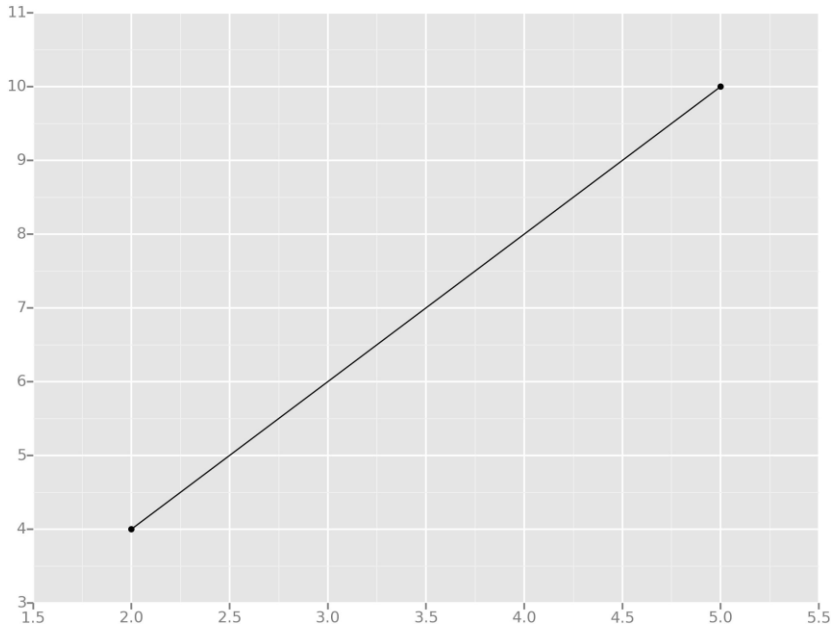
Exemplo

Vamos pegar o exemplo mais simples possível: calcular uma regressão com apenas 2 pontos de dados.



Exemplo

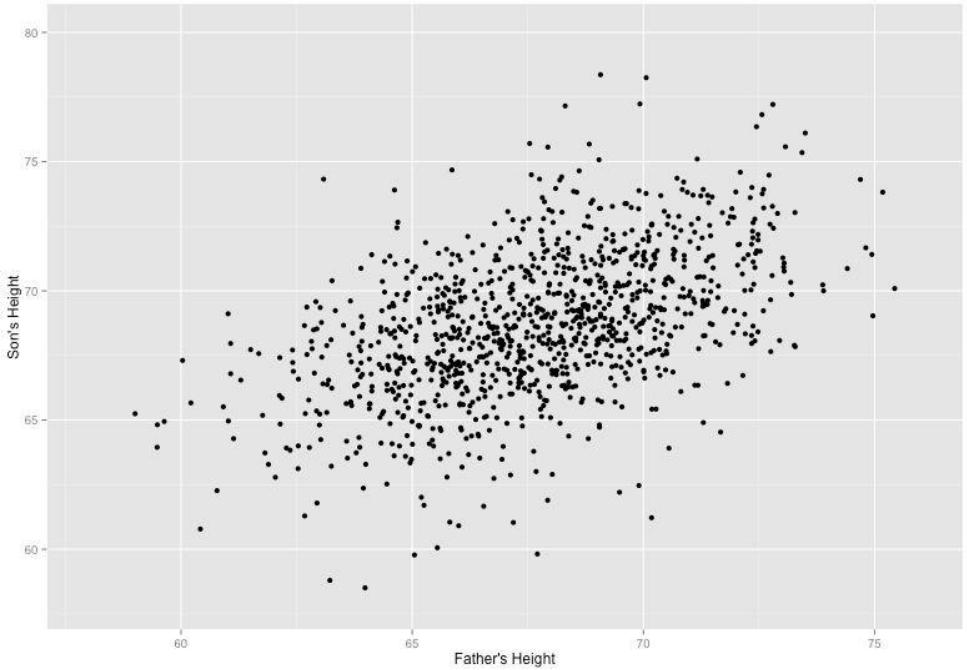
O que estamos tentando fazer quando calculamos a linha de regressão é desenhar uma linha que passe o mais próximo possível de cada ponto.
Para regressão linear clássica, ou "Método dos Mínimos Quadrados", você só mede a proximidade na direção "para cima e para baixo"



Exemplo

Agora, não seria ótimo se pudéssemos aplicar esse mesmo conceito a um gráfico com mais do que apenas dois pontos de dados?

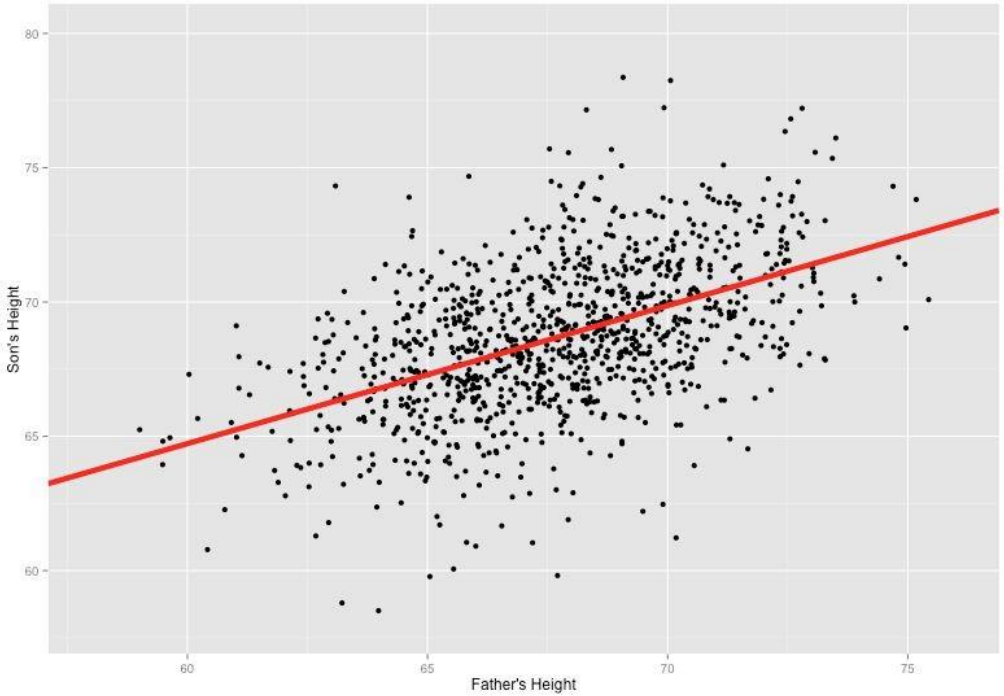
Ao fazer isso, poderíamos pegar vários homens e a altura de seus filhos e fazer coisas como dizer a um homem o quão alto esperamos que seu filho seja... antes mesmo de ele ter um filho!



Exemplo

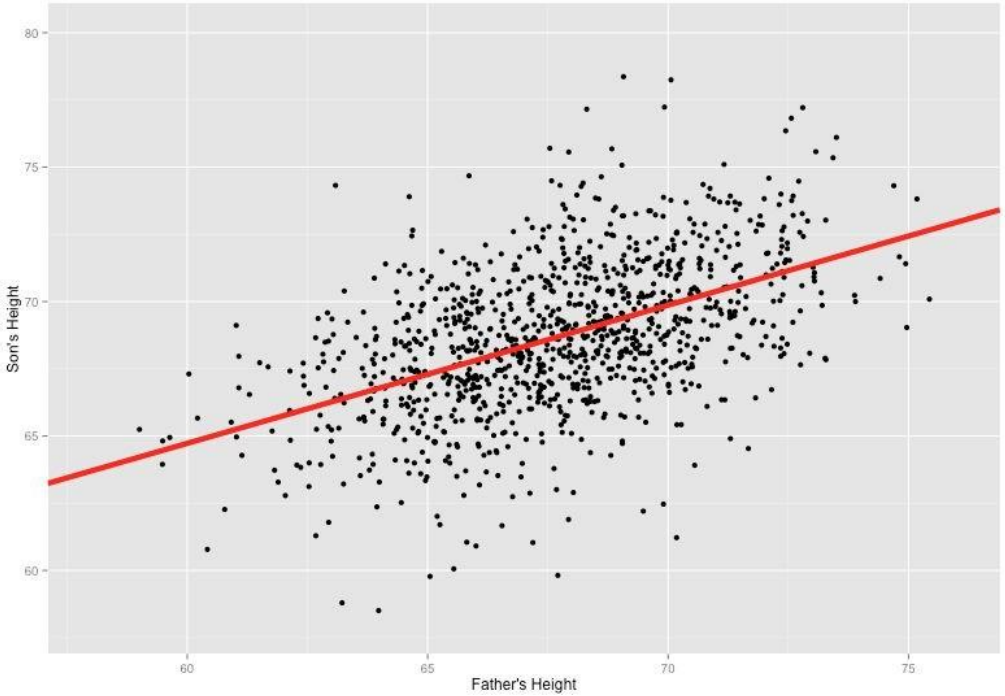
Nosso objetivo com a regressão linear é **minimizar a distância vertical** entre todos os pontos de dados e nossa linha.

Assim, ao determinar a **melhor linha**, estamos tentando minimizar a distância de **todos** os pontos à nossa linha.



Exemplo

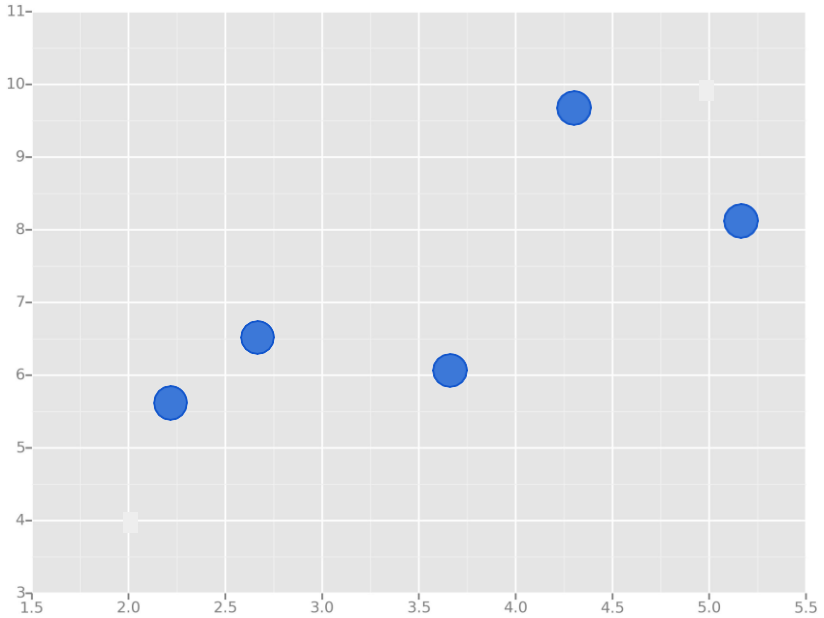
Existem muitas maneiras diferentes de minimizar isso (soma dos erros ao quadrado, soma dos erros absolutos, etc), mas todos esses métodos têm o objetivo geral de minimizar essa distância.



Exemplo

Por exemplo, um dos métodos mais populares é o método dos mínimos quadrados.

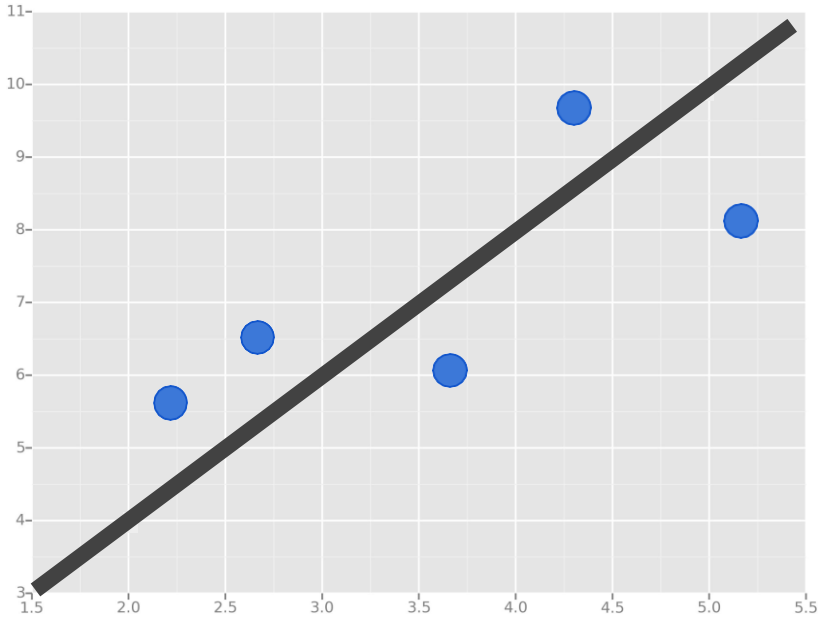
Aqui temos pontos de dados azuis ao longo de um eixo x e y.



Exemplo

Agora queremos ajustar uma linha de regressão linear.

A questão é: como decidimos qual linha é a mais adequada?



Exemplo

Usaremos o Método dos Mínimos Quadrados, que é ajustado minimizando a **soma dos quadrados dos resíduos.**

Os resíduos para uma observação são a diferença entre a observação (o valor de y) e a linha ajustada.