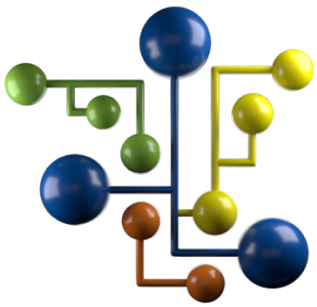


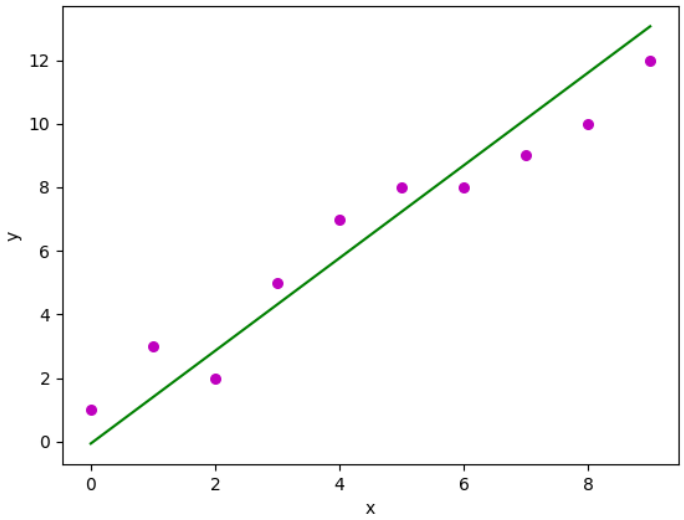
Data Science Academy

www.datascienceacademy.com.br

Business Analytics

Como Funciona a Regressão

Para realizar uma análise de regressão, você reúne os dados nas variáveis em questão. Você usa todos os seus números de vendas mensais nos últimos três anos e, por exemplo, os dados das variáveis independentes nas quais você está interessado. Então, neste caso, digamos que você descubra também a precipitação média mensal nos últimos três anos. Em seguida, você plota todas essas informações em um gráfico semelhante a este:



O eixo y é a quantidade de vendas (a variável dependente, a coisa em que você está interessado, está sempre no eixo y) e o eixo x é a precipitação total. Cada ponto rosa representa os dados de um mês - quanto choveu naquele mês e quantas vendas você fez no mesmo mês.

Olhando para esses dados, você provavelmente percebe que as vendas são maiores nos dias em que chove muito. É interessante saber, mas em quanto? Se chover 3 centímetros, você sabe quanto vai vender? E se chover 10 cm?

Agora imagine desenhar uma linha no gráfico acima, que percorre aproximadamente o meio de todos os pontos de dados. Essa linha o ajudará a responder, com certo grau de certeza, quanto você normalmente vende quando chove uma certa quantidade. É o que nos diz a linha verde no gráfico acima.

Isso é chamado de linha de regressão (a linha verde no gráfico acima) e é desenhado para mostrar a linha que melhor se ajusta aos dados. Em outras palavras: “a linha verde é a melhor explicação do relacionamento entre a variável independente e a variável dependente”.

Além de desenhar a linha, seu programa de estatísticas também gera uma fórmula que explica a inclinação da linha e se parece com isso:

Y = 200 + 5x + erro

Mas vamos desconsiderar o erro por enquanto e focar nesta parte da fórmula:

Y = 200 + 5x

O que essa fórmula está dizendo é que, se não houver um "x", então Y = 200. Então, historicamente, quando não choveu, você fez uma média de 200 vendas e pode esperar fazer o mesmo daqui para frente. assumindo que outras variáveis permaneçam iguais. E no passado, para cada centímetro adicional de chuva, você fazia em média mais cinco vendas. "Para cada incremento que x sobe um, y sobe cinco".

Agora vamos voltar ao termo de erro. Você pode ficar tentado a dizer que a chuva tem um grande impacto nas vendas se, a cada centímetro, você conseguir mais cinco vendas, mas se essa variável vale a pena, sua atenção dependerá do termo de erro. Uma linha de regressão sempre tem um termo de erro porque, na vida real, variáveis independentes nunca são preditores perfeitos das variáveis dependentes. Em vez disso, a linha é uma estimativa com base nos dados disponíveis. Portanto, o termo de erro informa o quanto você pode ter certeza sobre a fórmula. Quanto maior, menor a certeza da linha de regressão.

O exemplo acima usa apenas uma variável para prever o fator de interesse - nesse caso, chuva para prever vendas. Normalmente, você inicia uma análise de regressão que deseja entender o impacto de diversas variáveis independentes. Portanto, você pode incluir não apenas chuva, mas também dados sobre a promoção de um concorrente. "Você continua fazendo isso até que o termo de erro seja muito pequeno". "Você está tentando obter a linha que melhor se ajusta aos seus dados." Embora possa haver perigos em tentar incluir muitas variáveis em uma análise de regressão, Cientistas de Dados qualificados podem minimizar esses riscos.

E considerar o impacto de diversas variáveis ao mesmo tempo é uma das maiores vantagens da regressão. Isso é o que chamamos de análise multivariada.