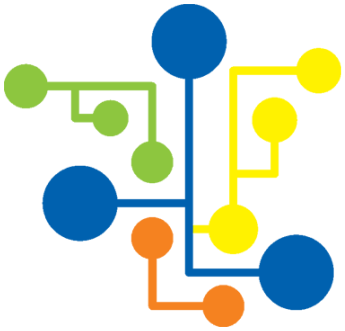


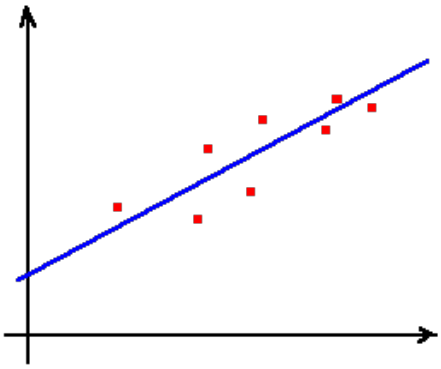
Data Science Academy

www.datascienceacademy.com.br

Microsoft Power BI Para Data Science

Regressão

Uma das preocupações estatísticas ao analisar dados, é a de criar modelos que explicitem estruturas do fenômeno em observação. O **modelo de regressão** é um dos métodos estatísticos mais usados para investigar a relação entre variáveis.



Como estudante, você já deve ter se perguntado quantas horas de estudo por semana seriam necessárias para conseguir **9.5** na sua prova final. O que você estava fazendo com esta pergunta, era buscando o relacionamento entre **horas de estudo** e sua **nota final**. E existem basicamente **2** técnicas que descrevem este relacionamento. Primeiro a **análise de correlação**, que determina a força e direção do relacionamento entre **duas** variáveis. Podemos usar o **teste de hipótese** para determinar a força do **relacionamento** entre **horas** de estudo e **nota final** no exame, por exemplo. Em seguida, podemos explorar a regressão linear simples, que descreve o relacionamento entre duas variáveis usando uma equação. Com esta equação, poderemos prever a nota final no exame, dado um número **x** de horas de estudo. E para montar esta equação, é preciso compreender o que são variáveis dependentes e independentes.

De forma bem objetiva: uma **variável independente x**, explica a variação em outra variável, que é chamada **variável dependente y**. Este relacionamento existe em apenas uma direção:

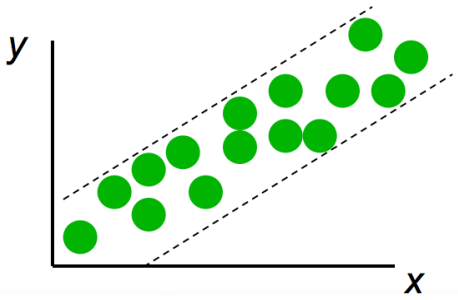
Variável independente (x) → variável dependente (y)

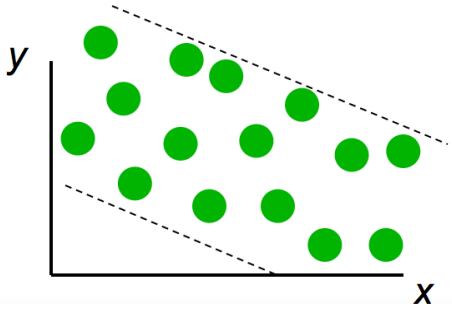
Por exemplo: A quantidade de quilômetros rodados de um carro, seria uma variável independente e o preço do carro seria uma variável dependente. **À medida que a quantidade de quilômetros rodados do carro aumenta, o preço do carro diminui.** Este relacionamento não funciona em modo reverso, ou seja, se alterarmos o preço do carro, a quantidade de quilômetros rodados não será alterada.

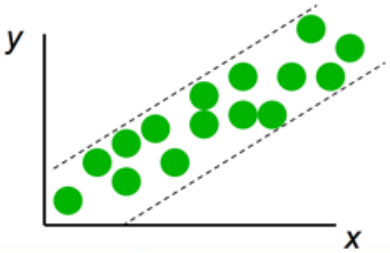
A análise de correlação nos permite medir a **força** e **direção** de um relacionamento linear entre **duas** variáveis.

Correlação

O relacionamento entre duas variáveis **é linear**, se o gráfico de dispersão entre elas tem o **padrão de uma linha reta**. Exemplos de relação linear:

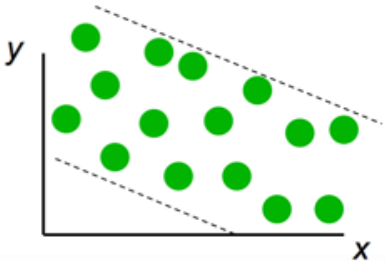






Relacionamento **positivo**, inclinação se move para cima.

Relacionamento **negativo**, inclinação se move para baixo.



O coeficiente de correlação **(r)** indica a força e direção de uma relação linear entre a variável independente e dependente.

Exemplos de valores de r:

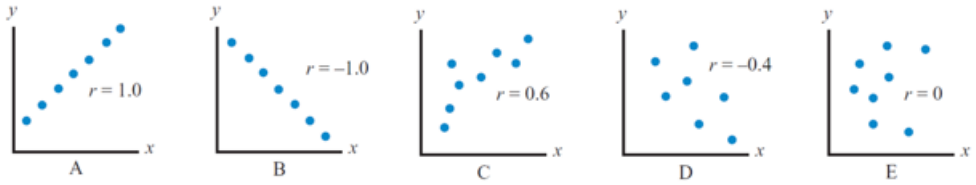


Gráfico A (*r* = 1.0): correlação positiva perfeita entre *x* e *y*

Gráfico B (*r* = -1.0): correlação negativa perfeita entre *x* e *y*

Gráfico C (*r* = 0.6): relação positiva moderada: y tende a aumentar se x aumenta, mas não necessariamente na mesma taxa observada no Gráfico A

Gráfico D (*r* = -0.4): relação negativa fraca: o coeficiente de correlação é próximo de zero ou negativo: y tende a diminuir se x aumenta

Gráfico E (*r* = 0): Sem relação entre x e y

Os valores de **r** variam entre **-1.0** (uma forte relação negativa) até **+1.0**, uma forte relação positiva.

A correlação, isto é, a ligação entre dois eventos, não implica necessariamente uma relação de causalidade, ou seja, que um dos eventos tenha causado a ocorrência do outro. A correlação pode, no entanto, indicar possíveis causas ou áreas para um estudo mais aprofundado, ou em outras palavras, a correlação pode ser uma pista. A ideia oposta, de que correlação prova automaticamente causalidade, é uma falácia lógica. Obviamente, dois eventos que possuam de fato uma relação de causalidade deverão apresentar também uma correlação. **O que constitui a falácia é o salto imediato para a conclusão de causalidade**, sem que esta seja devidamente demonstrada. Só porque (A) acontece juntamente com (B) não significa que (A) causa (B). É necessária investigação adicional em função de diferentes cenários que podem ocorrer:

1. (A) causa realmente (B);
2. (B) pode ser a causa de (A);
3. Um terceiro fator (C) pode ser causa tanto de (A) como de (B);
4. Pode ser uma combinação das três situações anteriores: (A) causa (B) e ao mesmo tempo (B) causa também (A);
5. A correlação pode ser apenas uma coincidência, ou seja, os dois eventos não têm qualquer relação para além do fato de ocorrerem ao mesmo tempo. (Se estivermos falando de um estudo científico, utilizar uma amostra grande ajuda a reduzir a probabilidade de coincidência).

Então como se determina a causalidade? Depende sobretudo da complexidade do problema, mas a verdade é que a causalidade dificilmente poderá ser determinada com certeza absoluta. Daí que em ciência já está subentendido que não existem verdades absolutas e que todas as teorias estão abertas a revisão face a novas evidências. No entanto, muitos erros podem ser evitados se tivermos mais cuidado com as conclusões precipitadas. Utilizando o método científico é possível muitas vezes estabelecer uma relação de causa-efeito com uma segurança confortável. O que acaba por ter mais importância no final é a reprodutibilidade da relação causa-efeito e a possibilidade de fazer previsões corretas sobre eventos futuros. A indústria do tabaco não pode continuar a alegar que a correlação entre o tabaco e o câncer de pulmão não implica necessariamente causalidade porque existe uma grande quantidade de evidências científicas a favor da relação causa-efeito. Já o movimento anti-vacinação não possui quaisquer evidências credíveis que suportem a afirmação de que as vacinas causam autismo. É aí que reside a diferença fundamental.

Regressão

O conjunto de técnicas de regressão é provavelmente um dos mais utilizados em análise de dados e também um dos mais simples. A Regressão compreende uma categoria inteira de algoritmos de Machine Learning para previsões numéricas.

O conjunto de técnicas de Regressão nos ajudam a entender a relação entre o comportamento de determinado fenômeno e o comportamento de uma ou mais variáveis potencialmente preditoras, sem que haja, entretanto, uma obrigatória relação de causa e efeito. Por exemplo, a relação entre a quantidade de horas de estudo de preparação e as notas no vestibular para Medicina é, obviamente, de natureza causal, ou seja, quanto maior a dedicação aos estudos, maiores serão as notas no vestibular, mesmo que também existam outros fatores que possam influenciar as notas no exame, como ansiedade e poder de concentração do candidato. Existem diversos modelos de Regressão:

- Regressão Linear Simples e Múltipla
- Regressão Logística Binária
- Regressão Logística Multinomial
- Regressão Poisson
- Regressão Binomial
- Regressão Ridge
- Regressão Lasso
- Regressão ElasticNet

Com exceção do ElasticNet, todos os demais modelos são estudados nas Formações Cientista de Dados e IA. Os modelos de regressão costumam ser a porta de entrada para quem está aprendendo Machine Learning.

Análise de Regressão é uma metodologia **estatística** que utiliza a relação entre duas ou mais variáveis quantitativas de tal forma que uma variável possa ser predita a partir de outra. É importante enfatizar que todo e qualquer modelo de regressão deve ser definido com base na teoria e na experiência do pesquisador, de modo que seja possível estimar o modelo desejado, analisar os resultados obtidos por meio de testes estatísticos e elaborar previsões.

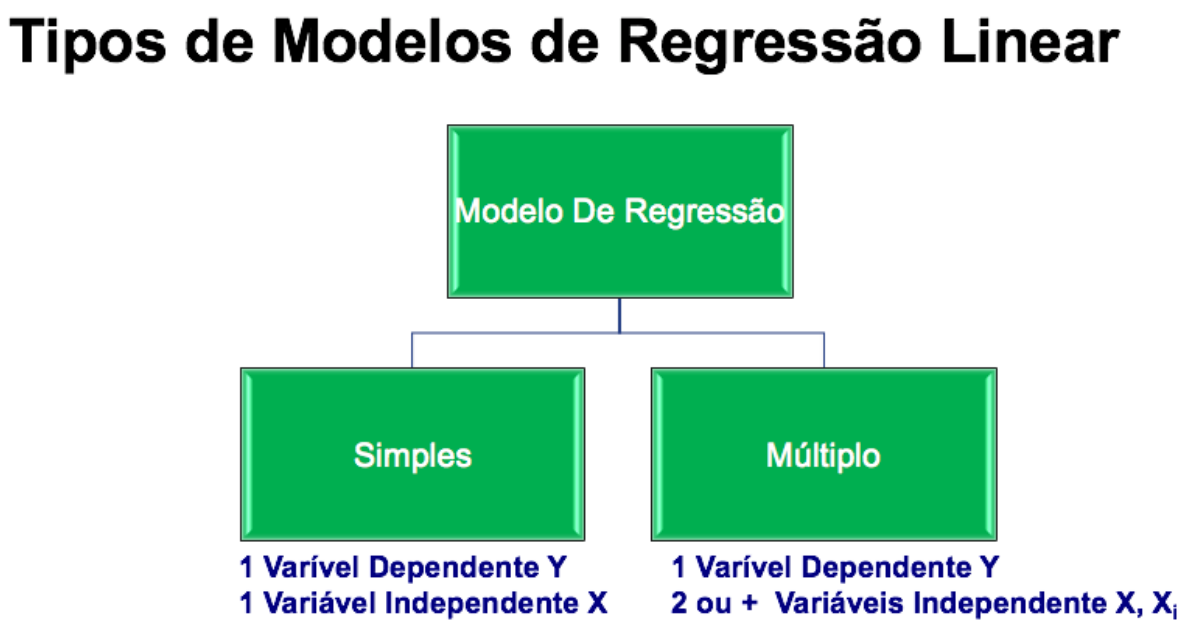
Origem do Modelo Clássico de Análise de Regressão

Tudo começou pelo ano de 1800 quando o termo ”Regressão” foi introduzido por Francis Galton. Durante anos ele estudou se existia uma tendência de pais com estatura alta terem filhos com estatura alta e de pais com estatura baixa terem filhos baixos. Entretanto ele verificou que a altura média dos filhos de pais de uma dada altura tendia a se deslocar ou seja “regredir” até a altura média da população como um todo. Em outras palavras, a altura dos filhos de pais muito altos ou baixos (que aqui poderíamos definir como outliers) tendem a se mover para a altura média da população.

A interpretação moderna da regressão é diferente – ocupa-se do estudo da dependência de uma variável (chamada variável endógena, resposta ou dependente), em relação a uma ou mais variáveis, as variáveis explicativas (ou exógenas), com o objetivo de estimar e/ou prever a média (da população) ou valor médio de dependente em termos dos valores conhecidos ou fixos (em amostragem repetida) das explicativas.

Regressão e Correlação são a mesma coisa? Não. **Análise de regressão** – prevê o valor médio de uma variável com base nos valores estabelecidos de outras variáveis. Já a **Análise de Correlação** permite medir o grau de associação linear entre duas variáveis.

Dentre os modelos de Regressão Linear, os mais comuns são o Linear Simples e Múltiplo:



Vamos compreender como funciona a Regressão Linear Simples.

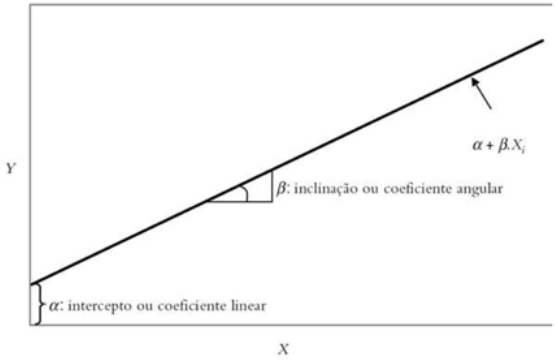
Nós já sabemos que o coeficiente de correlação **r** nos provê uma medida que descreve a força e direção do relacionamento entre duas variáveis. Nosso próximo passo é realizar uma **análise de regressão linear simples**, que nos habilite descrever uma linha reta que melhor representa uma série de pares ordenados (x, y). Ter uma linha reta que descreve o relacionamento entre a variável independente (x) e a variável dependente (y) nos oferece uma série de vantagens sobre o coeficiente de correlação.

Fórmula para a equação que descreve uma linha reta através de um par ordenado:

$$\hat{y} = a + bx$$

Onde:

- $\hat{y}$ = valor previsto de *y* dado um valor para *x*
- ***x*** = variável independente
- ***a*** = ponto onde a linha intercepta o eixo y
- ***b*** = inclinação da linha reta



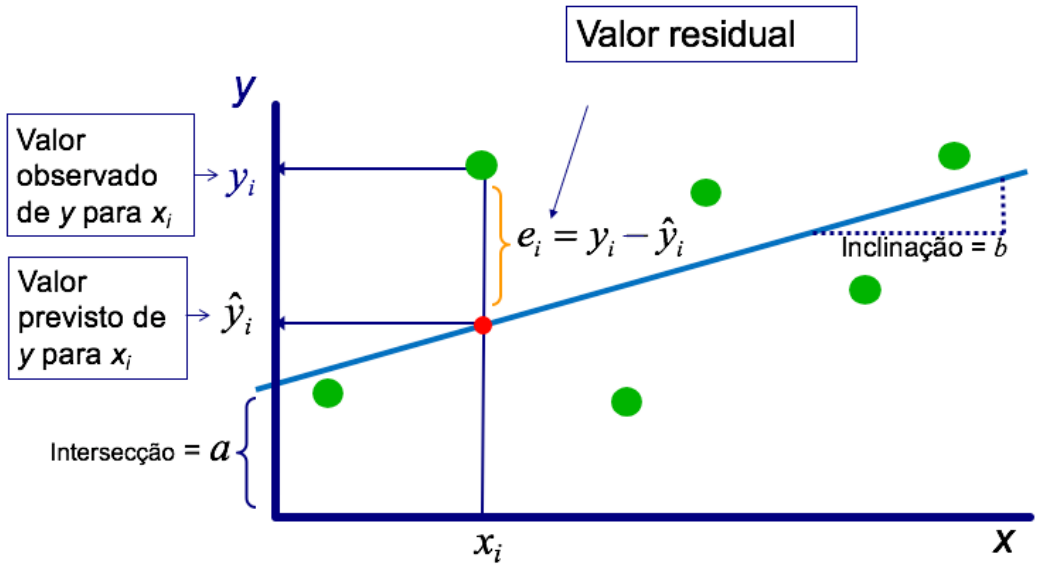
A diferença entre o valor atual e o valor previsto é conhecido como **residual, e_i.**

$$e_i = y_i - \hat{y}_i$$

onde:

- e_i = resíduo da observação *em uma posição específica* na amostra
- y_i = o valor atual da variável dependente no ponto *na posição específica*
- $\hat{y}_i$ = o valor previsto da variável dependente no ponto *na posição específica*

O valor residual é justificado pelo fato de que qualquer relação que seja proposta dificilmente se apresentará de maneira perfeita. No gráfico abaixo podemos ver que o valor residual é a diferença entre o valor atual e o valor previsto.



O ponto vermelho está sobre a linha azul e esta linha é chamada de linha de regressão e estabele os valores previstos pelo modelo. A título de comparação, perceba a diferença entre as equações das regressões lineares simples e múltiplas, de acordo com o número de variáveis envolvidas.

Regressão Linear Simples (2 variáveis)

$$\hat{Y}_i = \alpha + \beta . X_i$$

Regressão Linear Múltipla (Mais de 2 variáveis)

$$Y_i = a + b_1 . X_{1i} + b_2 . X_{2i} + \ldots + b_k . X_{ki} + u_i$$

Esta explicação simples resume como funciona um modelo básico de Machine Learning. Alimentamos nosso modelo com dados (pares ordenados de x e y), executamos a análise de regressão e encontramos uma função matemática que descreve o relacionamento entre x e y. Esta função matemática é encontrada dentro do espaço de hipóteses do algoritmo e o modelo resultante terá um determinado nível de precisão (erros são esperados). Nosso trabalho então é apresentar novos valores de x ao modelo e coletar as previsões y. Todo modelo de Machine Learning segue o mesmo raciocínio, variando a abordagem e o espaço de hipóteses do algoritmo sejam diferentes.

Qual o objetivo da Análise de Regressão?

Frequentemente vislumbramos, de forma racional ou intuitiva, a relação entre comportamentos de variáveis que se apresentam de forma direta ou indireta. Será que se eu frequentar mais a academia aumentarei a minha massa muscular? Será que se eu mudar de emprego terei mais tempo para ficar com meus filhos? Será que se eu poupar maior parcela de meu salário poderei me aposentar mais jovem? Estas questões oferecem nitidamente relações entre uma determinada variável dependente, que representa o fenômeno que se deseja estudar, e, no caso, uma única variável explicativa ou explanatória. O objetivo principal da análise de regressão é, portanto, propiciar ao analista condições de avaliar como se comporta uma variável Y com base no comportamento de uma ou mais variáveis X, sem que, necessariamente, ocorra uma relação de causa e efeito.