Regressão Linear

Introdução a Técnicas de Regressão

Paulo Henrique Ribeiro Gabriel

Faculdade de Computação
Universidade Federal de Uberlândia

2024

Recapitulando

Regressão

- A análise de regressão é um dos campos fundamentais da estatística e do aprendizado de máquina
- Procura relacionamentos entre variáveis

- Por exemplo, podemos observar diversos funcionários de alguma empresa e tentar entender como seus salários dependem de suas características
  - Experiência
  - Escolaridade
  - Cargo, etc.
- Neste caso, os dados relativos a cada funcionário representam uma observação
- Nossa hipótese é que a experiência, a educação e o cargo são características independentes, enquanto o salário depende delas

Recapitulando

Regressão

Outro exemplo:

- Podemos tentar estabelecer a dependência matemática dos preços de imóveis em relação a diferentes atributos:
  - Área
  - Número de quartos
  - Bairro da cidade
  - ... e assim por diante

Recapitulando
Regressão

- Geralmente consideramos algum fenômeno de interesse e temos uma série de observações
- Cada observação possui duas ou mais características (*features*)
- Partindo do pressuposto de que pelo menos uma das características depende das demais, tenta-se estabelecer uma relação entre elas
- Em outras palavras, precisamos encontrar uma função que mapeie algumas variáveis em outras
  - Da melhor forma possível

Recapitulando
Regressão

- Os problemas de regressão geralmente têm uma variável dependente contínua e ilimitada
- As entradas, no entanto, podem ser dados contínuos, discretos ou, ainda, categóricos
  - Gênero
  - Nacionalidade
  - Marca

Recapitulando

Regressão

- É comum denotar os resultados (*output*) com y e as observações (*input*) como x
- Se houver duas ou mais variáveis independentes, elas são representadas como o vetor

$$\mathbf{x} = (x_1, \ldots, x_d)$$

onde d é o número de entradas (*dimensão*)

Análise de Regressão

Aplicações

- Normalmente, precisamos de regressão para responder *se* ou *como* um fenômeno influencia o outro
- Ou, ainda, como diversas variáveis estão relacionadas
- Por exemplo, podemos usar a regressão para determinar em que medida a experiência ou o gênero impactam os salários

Análise de Regressão

Aplicações

- A regressão também é útil quando desejamos prever uma resposta usando um novo conjunto de preditores

- Por exemplo, poderíamos tentar prever o consumo de eletricidade de uma residência para a próxima hora, dados:
  - a temperatura externa
  - a hora do dia
  - o número de residentes

Regressão Linear

- A regressão linear é uma das técnicas de regressão mais importantes e amplamente utilizadas
- Está entre os métodos de regressão mais simples
- Uma de suas principais vantagens é a facilidade de interpretação dos resultados

Regressão Linear

Formulação do Problema

- Seja uma variável dependente y e um conjunto de variáveis independentes $\mathbf{x} = (x_1, \ldots, x_d)$, onde d é o número de observações (preditores)
- A regressão linear assume uma relação linear entre y e **x**, dada por:

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d + \varepsilon$$

- Esta equação é a equação de regressão
- $\beta_0, \beta_1, \ldots, \beta_d$ são os coeficientes de regressão
- ε é o erro aleatório

Regressão Linear

Formulação do Problema

- A regressão linear calcula os estimadores dos coeficientes de regressão ou simplesmente os pesos previstos
- Esses pesos são denotados por $b_0, b_1, \ldots, b_d$
- Esses estimadores definem a função de regressão estimada

$$f(\mathbf{x}) = b_0 + b_1 x_1 + \ldots + b_d x_d$$

- Essa função deve capturar suficientemente bem as dependências entre as entradas e a saída

Regressão Linear

Formulação do Problema

- A resposta estimada ou prevista
$$f(x_i)$$
para cada observação
$$i = 1, \ldots, d$$
deve ser o mais próximo possível da resposta real correspondente y_i
- As diferenças $y_i - f(x_i)$ para todas são chamadas de resíduos
- A regressão trata de determinar os melhores pesos previstos
  - ou seja, os pesos correspondentes aos menores resíduos

- Para obter os melhores pesos, geralmente minimizamos a soma dos quadrados dos resíduos (*sum of squared residuals*) para todas as observações $i = 1, \ldots, d$

$$SQR = \sum_i (y_i - f(x_i))^2$$

Regressão Linear

Desempenho da Regressão

- A variação das respostas reais ocorre em parte devido à dependência dos preditores
- No entanto, há também uma variação inerente adicional da saída
- O coeficiente de determinação , denotado como R^2, informa qual quantidade de variação em y pode ser explicada pela dependência de $\mathbf{x}$
- Um R^2 maior indica um melhor ajuste e significa que o modelo pode explicar melhor a variação da saída com diferentes entradas
- O valor $R^2 = 1$ corresponde a $SQR = 0$
  - Esse é o ajuste perfeito, pois os valores das respostas previstas e reais se ajustam completamente entre si

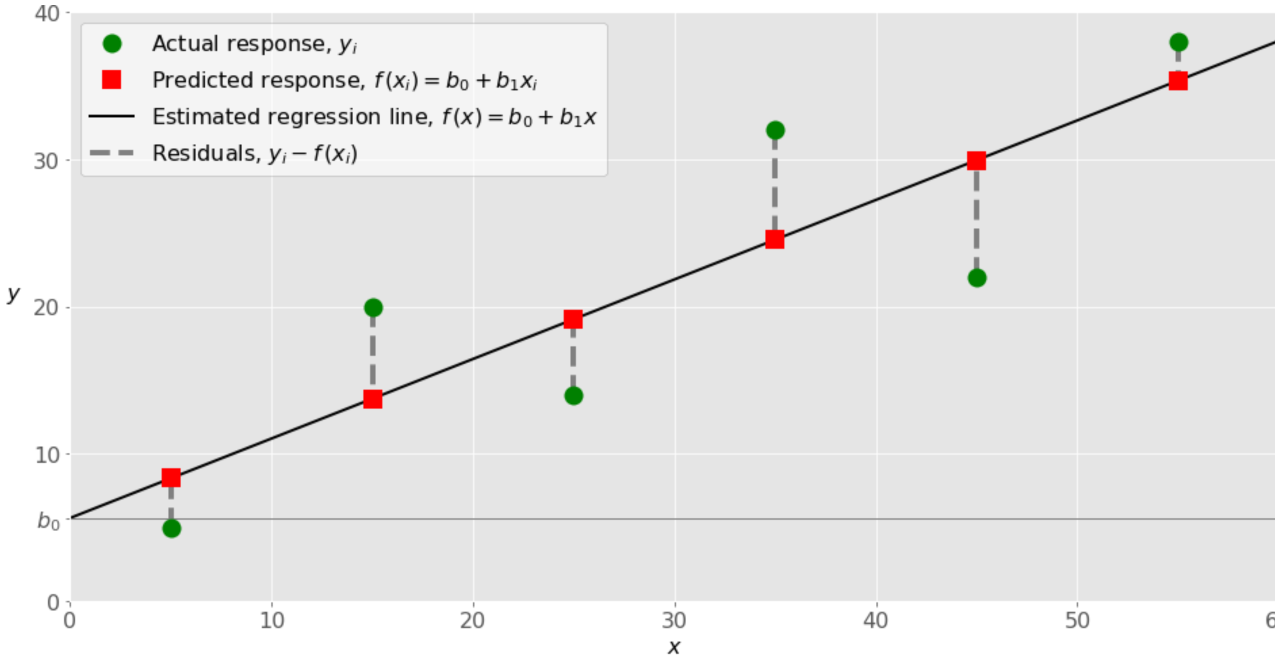
- A regressão linear simples ou univariada é o caso mais simples de regressão linear
- Possui possui uma única variável independente, ou seja, $\mathbf{x} = x$.
- A função de regressão estimada tem a equação

$$f(x) = b_0 + b_1 x$$

- Nosso objetivo é calcular os valores ideais dos pesos previstos b_0 e b_1 que minimizam o SQR e determinar a função de regressão estimada

Regressão Linear

Regressão Linear Simples

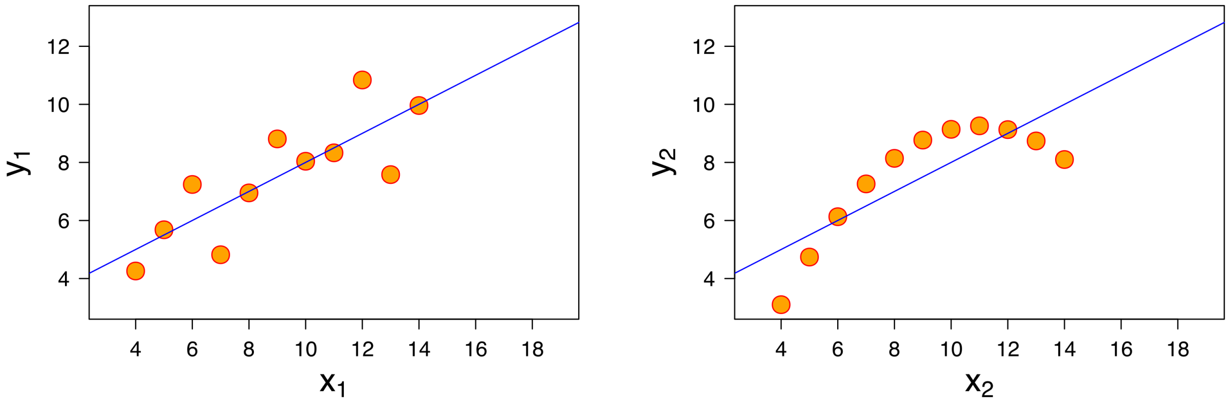


- O valor de b_0, também chamado de interceptação, mostra o ponto onde a reta de regressão estimada cruza o eixo y
- Esse é o valor da resposta estimada $f(x)$ para $x = 0$
- O valor de b_1 determina a inclinação da reta de regressão estimada
- As respostas previstas são os pontos na linha de regressão que correspondem aos valores de entrada

Regressão Linear

Limitações

- Lembre-se: a regressão linear não é um método à prova de falhas
- Por exemplo, conjuntos de dados distintos podem ser “aproximados” pela mesma equação da reta

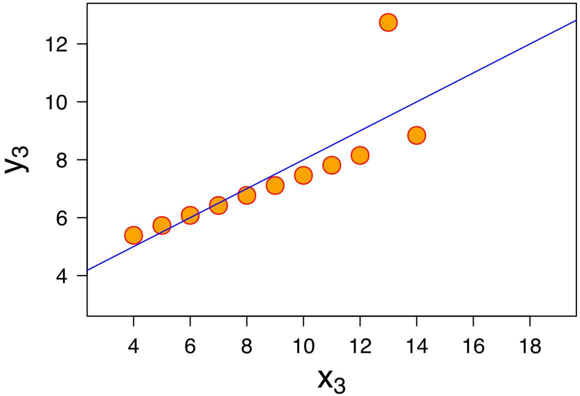


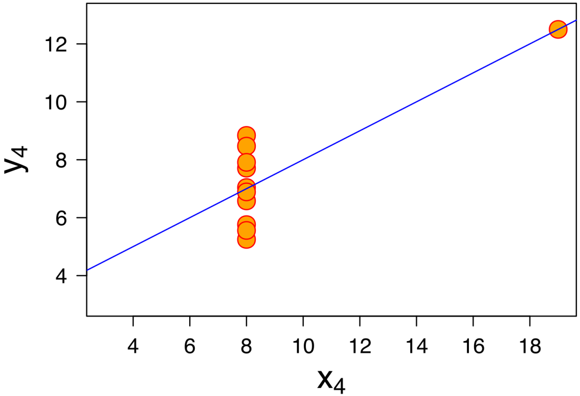
Regressão Linear

Limitações

- Lembre-se: a regressão linear não é um método à prova de falhas

- Por exemplo, conjuntos de dados distintos podem ser “aproximados” pela mesma equação da reta





Regressão Linear Múltipla

- A regressão linear múltipla ou multivariada é um caso de regressão linear com duas ou mais variáveis independentes
- Se houver apenas duas variáveis independentes, então a função de regressão estimada é

$$f(x_1, x_2) = b_0 + b_1x_1 + b_2x_2$$

- Nesse caso, temos um plano de regressão em um espaço tridimensional
- O objetivo da regressão é, portanto, determinar os valores dos pesos b_0, b_1 e b_2 de modo que este plano esteja o mais próximo possível das respostas reais
  - Ao mesmo tempo que produz o SQR mínimo

Regressão Linear Múltipla

- O caso de mais de duas variáveis independentes é mais geral
- A função de regressão estimada é

$$f(x_1, \ldots, x_d) = b_0 + b_1 x_1 + \ldots + b_d x_d$$

- Há, portanto, $d + 1$ pesos a serem determinados