Regressão Linear

24 de julho de 2023

1 O que é uma regressão linear?

A regressão linear é uma técnica estatística usada para entender a relação entre duas variáveis contínuas. Ela assume uma relação linear entre essas variáveis, ou seja, assume que uma variável pode ser expressa como uma combinação linear das outras.

O modelo mais simples de regressão linear é o modelo de regressão linear simples, que se relaciona com duas variáveis: uma variável independente x e uma variável dependente y. É formulado como:

$$y = \beta_0 + \beta_1 x$$

onde: - y é a variável dependente (resposta),
- x é a variável independente (preditor),
- β_0 é o intercepto (o valor esperado de y quando $x = 0$),
- β_1 é a inclinação (o quanto esperamos que y mude, em média, com um aumento de uma unidade em x),

Se tivermos mais de uma variável independente, o modelo é chamado de regressão linear múltipla e é formulado como:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n$$

Aqui, $x_1, x_2, ..., x_n$ são as variáveis independentes e $\beta_1, \beta_2, ..., \beta_n$ são os coeficientes que quantificam a relação entre cada variável independente e a variável dependente.

Os coeficientes do modelo são geralmente estimados usando o método dos mínimos quadrados, que minimiza a soma dos quadrados das diferenças entre os valores observados e previstos de y.

A regressão linear é amplamente utilizada em muitos campos, incluindo ciências naturais, ciências sociais, negócios e engenharia, devido à sua simplicidade e flexibilidade. No entanto, é baseada em várias suposições (incluindo a linearidade e a independência dos erros) que, se violadas, podem levar a estimativas imprecisas ou enganosas.

2 Como funciona o método dos mínimos quadrados?

O Método dos Mínimos Quadrados (Least Squares Method) é um procedimento de otimização que busca encontrar a melhor linha de ajuste em uma regressão linear. Ele faz isso minimizando a soma dos quadrados dos resíduos (as diferenças entre os valores observados e previstos da variável dependente).

Suponha que temos um conjunto de dados com n observações e queremos ajustar um modelo de regressão linear simples. Vamos denotar a variável dependente como y_i e a variável independente como x_i para a i-ésima observação ($i = 1, ..., n$). O modelo é dado por:

$$y_i = \beta_0 + \beta_1 x_i$$

O objetivo do método dos mínimos quadrados é encontrar os valores de β_0 e β_1 que minimizam a soma dos quadrados dos resíduos, denotada como RSS (Residual Sum of Squares):

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

Onde $\hat{y}_i = \beta_0 + \beta_1 x_i$ é o valor previsto de y_i.

O problema de otimização pode ser formulado como:

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)^2$$

As soluções para β_0 e β_1 que minimizam RSS podem ser encontradas calculando as derivadas parciais de RSS em relação a β_0 e β_1, igualando-as a zero e resolvendo as equações resultantes. As soluções são dadas por:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Onde $\bar{x}$ e $\bar{y}$ são as médias dos valores observados de x e y, respectivamente, e $\hat{\beta}_0$ e $\hat{\beta}_1$ são os estimadores de mínimos quadrados de β_0 e β_1.

Esse método fornece uma maneira prática de estimar os coeficientes de um modelo de regressão linear. No entanto, como mencionei antes, é baseado em várias suposições, incluindo a linearidade e a independência dos erros. Se essas suposições forem violadas, as estimativas dos mínimos quadrados podem ser imprecisas ou enganosas.

2.1 Forma matricial

O método dos mínimos quadrados também pode ser formulado em termos de álgebra matricial. Esse formato é particularmente útil quando se trata de regressão linear múltipla, onde temos mais de uma variável independente.

Suponha que temos um conjunto de dados com n observações e p variáveis independentes. Denotamos a matriz de design como X, que é uma matriz $n \times (p+1)$, onde a primeira coluna é toda composta por uns (para o termo de intercepto) e as colunas restantes contêm os valores das variáveis independentes. Denotamos o vetor de respostas como y, que é um vetor $n \times 1$ que contém os valores da variável dependente. Finalmente, denotamos o vetor de coeficientes como $\boldsymbol{\beta}$, que é um vetor $(p+1) \times 1$ que contém os coeficientes do modelo.

O modelo de regressão linear múltipla pode ser escrito na forma matricial como:

$$y = X\boldsymbol{\beta}$$

O objetivo do método dos mínimos quadrados é encontrar o vetor de coeficientes $\boldsymbol{\beta}$ que minimiza a soma dos quadrados dos resíduos, que é dada por:

$$RSS = (y - X\boldsymbol{\beta})^T (y - X\boldsymbol{\beta})$$

Onde T denota a transposição.

O problema de otimização pode ser formulado como:

$$\min_{\boldsymbol{\beta}} (y - X\boldsymbol{\beta})^T (y - X\boldsymbol{\beta})$$

A solução para $\boldsymbol{\beta}$ que minimiza RSS pode ser encontrada calculando a derivada de RSS em relação a $\boldsymbol{\beta}$, igualando-a a zero e resolvendo a equação resultante. A solução é dada por:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$$

Onde $\hat{\boldsymbol{\beta}}$ é o estimador de mínimos quadrados de $\boldsymbol{\beta}$, e $^{-1}$ denota a inversa de uma matriz.

Observe que a solução só existe se a matriz $X^T X$ for invertível. Se as colunas de X forem linearmente dependentes, a matriz $X^T X$ não será invertível e a solução não será única. Isso é conhecido como o problema de multicolinearidade na regressão linear.

2.2 Como calcular a derivada?

Podemos chegar à expressão para os coeficientes de mínimos quadrados diferenciando a função da Soma dos Quadrados dos Resíduos (RSS) e igualando a derivada a zero. Vamos passar por isso passo a passo.

Primeiro, definimos a função da Soma dos Quadrados dos Resíduos (RSS) em termos matriciais:

$$RSS = (y - X\boldsymbol{\beta})^T (y - X\boldsymbol{\beta})$$

Isso é equivalente a:

$$RSS = y^T y - \boldsymbol{\beta}^T X^T y - y^T X\boldsymbol{\beta} + \boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$$

Agora, diferenciamos essa expressão em relação a $\boldsymbol{\beta}$. Isso requer um pouco de cálculo matricial. A regra que usamos é que a derivada de um escalar a com relação a um vetor $\boldsymbol{x}$ é dada por $\frac{da}{d\boldsymbol{x}} = X^T \frac{da}{dX}$. Também precisamos das seguintes regras de diferenciação para matrizes: $\frac{d(Ax)}{dx} = A$ e $\frac{d(x^T Ax)}{dx} = (A + A^T)x$.

Vamos começar diferenciando cada termo separadamente.

1. O primeiro termo $y^T y$ não contém $\boldsymbol{\beta}$, portanto, sua derivada em relação a $\boldsymbol{\beta}$ é zero.
2. O segundo termo $-\boldsymbol{\beta}^T X^T y$ é um produto de um vetor e um escalar, e sua derivada em relação a $\boldsymbol{\beta}$ é $-X^T y$.
3. O terceiro termo $-y^T X\boldsymbol{\beta}$ é um produto de um escalar e um vetor, e sua derivada em relação a $\boldsymbol{\beta}$ é também $-X^T y$.
4. O quarto termo $\boldsymbol{\beta}^T X^T X\boldsymbol{\beta}$ é um produto quadrático de um vetor, e sua derivada em relação a $\boldsymbol{\beta}$ é $2X^T X\boldsymbol{\beta}$.

Agora, somamos essas derivadas para obter a derivada total de RSS em relação a $\boldsymbol{\beta}$:

$$\frac{dRSS}{d\boldsymbol{\beta}} = 0 - X^T y - X^T y + 2X^T X\boldsymbol{\beta} = -2X^T y + 2X^T X\boldsymbol{\beta}$$

Agora, igualamos essa derivada a zero e resolvemos para $\boldsymbol{\beta}$:

$$-2X^T y + 2X^T X\boldsymbol{\beta} = 0$$

Isso simplifica para:

$$(X^T X)\boldsymbol{\beta} = X^T y$$

E finalmente, resolvemos para $\boldsymbol{\beta}$ para obter a solução de mínimos quadrados:

$$\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T y$$

Essa é a derivação completa do estimador de mínimos quadrados para os coeficientes em uma regressão linear. Isso nos permite encontrar a linha (ou hiperplano, em casos de múltiplas variáveis independentes) que melhor se ajusta aos dados, no sentido de que minimiza a soma dos quadrados das diferenças entre os valores observados e previstos da variável dependente.

3 Exemplo

Claro, vou começar com um exemplo simples de regressão linear univariada. Suponha que temos os seguintes dados:

- Valores X (variável independente): 1, 2, 3, 4, 5
- Valores y (variável dependente): 2, 3, 5, 4, 6

Vamos usar esses dados para ajustar um modelo de regressão linear usando o método dos mínimos quadrados. A equação do nosso modelo será $y = \beta_0 + \beta_1 X$.

Primeiro, precisamos construir a matriz de design X. Como estamos lidando com regressão univariada, a matriz de design será uma matriz 5×2 em que a primeira coluna é toda composta por uns (para o termo de intercepto) e a segunda coluna contém os valores de X. O vetor de respostas y é um vetor de 5 elementos com os valores de y. Então temos:

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}, \quad y = \begin{bmatrix} 2 \\ 3 \\ 5 \\ 4 \\ 6 \end{bmatrix}$$

Agora, podemos calcular os coeficientes do modelo usando a fórmula do método dos mínimos quadrados:

$$\boldsymbol{\beta} = (X^T X)^{-1} X^T y$$

Primeiro, precisamos calcular a transposição de X e multiplicar por X:

$$X^T X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix} = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix}$$

Agora, calculamos a inversa desta matriz:

$$(X^T X)^{-1} = \frac{1}{5 * 55 - 15 * 15} \begin{bmatrix} 55 & -15 \\ -15 & 5 \end{bmatrix} = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & -0.02 \end{bmatrix}$$

Agora, calculamos $X^T y$:

$$X^T y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 1 & 2 & 3 & 4 & 5 \end{bmatrix} \begin{bmatrix} 2 \\ 3 \\ 5 \\ 4 \\ 6 \end{bmatrix} = \begin{bmatrix} 20 \\ 70 \end{bmatrix}$$

Finalmente, podemos calcular os coeficientes $\boldsymbol{\beta}$:

$$\boldsymbol{\beta} = \begin{bmatrix} 0.2 & 0.1 \\ 0.1 & -0.02 \end{bmatrix} \begin{bmatrix} 20 \\ 70 \end{bmatrix} = \begin{bmatrix} 1 \\ 0.8 \end{bmatrix}$$

Portanto, a equação do nosso modelo é $y = 1 + 0.8X$.

Este é um exemplo muito simples de como usar o método dos mínimos quadrados para ajustar um modelo de regressão linear. Em uma situação real, você provavelmente usaria uma biblioteca de computação científica ou estatística para fazer esses cálculos, mas é útil entender o que está acontecendo por baixo do capô.

4 É possível usar regressão linear para problemas não-lineares?

A regressão linear é um método de modelagem que assume uma relação linear entre a variável dependente e as variáveis independentes. No entanto, é possível usar a regressão linear para modelar relações não lineares, através de transformações de variáveis ou da criação de variáveis polinomiais.

Existem várias maneiras de criar um modelo não linear a partir da regressão linear:

1. **Transformações de variáveis:** Às vezes, uma transformação simples da variável independente ou dependente pode linearizar a relação. Por exemplo, se a relação é exponencial (por exemplo, $y = ae^{bx}$), podemos tirar o logaritmo natural de ambos os lados para obter uma equação linear (por exemplo, $\ln y = \ln a + bx$).

2. **Variáveis polinomiais:** Se a relação entre a variável dependente e independente é polinomial (por exemplo, $y = ax^2 + bx + c$), podemos criar uma nova variável que é o quadrado da variável original (por exemplo, $z = x^2$) e então ajustar um modelo linear às variáveis transformadas (por exemplo, $y = az + bx + c$).

3. **Interações entre variáveis:** Às vezes, a relação entre a variável dependente e várias variáveis independentes não é aditiva, o que significa que a efeito combinado de duas variáveis independentes sobre a variável dependente não é simplesmente a soma de seus efeitos individuais. Nesses casos, podemos criar uma nova variável que é o produto das variáveis originais e incluir essa variável de interação no modelo.

4. **Funções de base:** Uma abordagem mais geral para modelar relações não lineares é usar funções de base, que são funções não lineares das variáveis independentes. Por exemplo, em um modelo linear com base em splines, ajustamos um modelo linear à variável dependente e a várias funções spline das variáveis independentes.

Essas são todas formas de criar modelos lineares generalizados, que são modelos que mantêm a natureza linear das equações de regressão linear, mas permitem uma maior flexibilidade na modelagem de relações entre a variável dependente e as variáveis independentes. Isso nos permite capturar relações não lineares e interações entre variáveis de uma maneira que ainda é computacionalmente eficiente e relativamente fácil de interpretar.

5 Exercícios

1. **Exercício 1: Modelo Linear Simples** Dados os seguintes pontos (1, 3), (2, 5), (3, 7), (4, 9), (5, 11), encontre a linha que melhor se ajusta a esses pontos usando regressão linear. Encontre os coeficientes β_0 (interceptação) e β_1 (inclinação) do modelo.

2. **Exercício 2: Modelo Linear Simples com Transformação** Suponha que você tenha os seguintes dados: (1, 1), (2, 4), (3, 9), (4, 16), (5, 25). Estes pontos parecem formar uma curva parabólica em vez de uma linha. Como você pode transformar esses dados para que eles possam ser ajustados usando um modelo linear?

3. **Exercício 3: Modelo Linear Multivariado** Dados os seguintes pontos (1, 2, 2), (2, 3, 4), (3, 4, 6), (4, 5, 8), onde a primeira entrada de cada ponto é a primeira variável independente x_1, a segunda entrada é a segunda variável independente x_2, e a terceira entrada é a variável dependente y, encontre o plano que melhor se ajusta a esses pontos usando regressão linear multivariada.

4. **Exercício 4: Modelo Linear com Interação** Suponha que você tenha os seguintes dados: (1, 2, 2), (2, 3, 8), (3, 4, 18), (4, 5, 32), onde a primeira entrada de cada ponto é a primeira variável independente x_1, a segunda entrada é a segunda variável independente x_2, e a terceira entrada é a variável dependente y. Os dados parecem seguir um modelo linear com interação. Como você pode transformar esses dados para que eles possam ser ajustados usando um modelo linear?

5. **Exercício 5: Método dos Mínimos Quadrados** Derive a solução de mínimos quadrados para o modelo de regressão linear simples $y = \beta_0 + \beta_1 x$.

6 Respostas

1. **Exercício 1: Modelo Linear Simples**

   Os pontos formam uma linha perfeita $y = 2x + 1$. Portanto, os coeficientes da regressão são $\beta_0 = 1$ e $\beta_1 = 2$.

2. **Exercício 2: Modelo Linear Simples com Transformação**

   Observando os dados, vemos que eles seguem uma relação quadrática $y = x^2$. Podemos transformar a variável dependente y tomando a raiz quadrada de y, para obter novos pontos: (1, 1), (2, 2), (3, 3), (4, 4), (5, 5). Agora, a relação entre x e $\sqrt{y}$ é linear e pode ser modelada como uma regressão linear simples com coeficientes $\beta_0 = 0$ e $\beta_1 = 1$.

3. **Exercício 3: Modelo Linear Multivariado**

   Os pontos seguem a relação $y = x_1 + x_2$. Portanto, os coeficientes da regressão são $\beta_0 = 0$, $\beta_1 = 1$ e $\beta_2 = 1$.

4. **Exercício 4: Modelo Linear com Interação**

   Os dados parecem seguir a relação $y = 2x_1x_2$. Podemos criar uma nova variável de interação $x_3 = x_1 * x_2$ e realizar uma regressão linear simples nos novos pontos: (1, 2, 4), (2, 3, 12), (3, 4, 24), (4, 5, 40). Os coeficientes da regressão são $\beta_0 = 0$ e $\beta_1 = 2$.

5. **Exercício 5: Método dos Mínimos Quadrados**

   Dada a relação $y = \beta_0 + \beta_1 x$ a solução dos mínimos quadrados minimiza a soma dos quadrados dos resíduos, que é:

$$RSS = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

   Tomando a derivada de RSS em relação a β_0 e β_1 e igualando a zero, obtemos o sistema de equações:

$$\frac{\partial RSS}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial RSS}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

   Resolvendo o sistema, obtemos as equações dos mínimos quadrados:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

   onde $\bar{x}$ e $\bar{y}$ são as médias de x e y, respectivamente.

7 Complete derivation

The equation of the simple linear regression is

$$y_i = \beta_0 + \beta_1 x_i$$

The least squares method minimizes the sum of squared residuals:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

The first order conditions are obtained by taking the derivative of L with respect to β_0 and β_1 and setting them equal to zero.

For β_0:

$$\frac{\partial L}{\partial \beta_0} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

Rearranging gives:

$$n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

Dividing through by n, we have

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

For β_1:

$$\frac{\partial L}{\partial \beta_1} = -2 \sum_{i=1}^{n} x_i (y_i - \beta_0 - \beta_1 x_i) = 0$$

Rearranging and substituting β_0 from the previous equation, we obtain:

$$\beta_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

So, the least squares solution is:

$$\beta_0 = \bar{y} - \beta_1 \bar{x}$$

$$\beta_1 = \frac{n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2}$$

This can be rewritten in terms of deviations from the mean.
The numerator is:

$$n \sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i$$

This can be rewritten as:

$$n \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x})(y_i - \bar{y} + \bar{y}) - n\bar{x}\bar{y} = n \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) + n\bar{x}\bar{y} - n\bar{x}\bar{y} = n \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$$

Similarly, the denominator can be rewritten. We have:

$$n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2$$

which can be rewritten as:

$$n \sum_{i=1}^{n} (x_i - \bar{x} + \bar{x})^2 - n\bar{x}^2 = n \sum_{i=1}^{n} (x_i - \bar{x})^2 + n\bar{x}^2 - n\bar{x}^2 = n \sum_{i=1}^{n} (x_i - \bar{x})^2$$

So, the expression for β_1 becomes:

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

which is the formula you asked for.