

Método Dos Mínimos Quadrados

O método dos mínimos quadrados é uma técnica estatística utilizada para encontrar a melhor linha reta que descreve a relação entre duas variáveis. Ele é comumente utilizado para realizar uma regressão linear, que é o processo de encontrar a relação linear entre uma variável independente (x) e uma variável dependente (y).

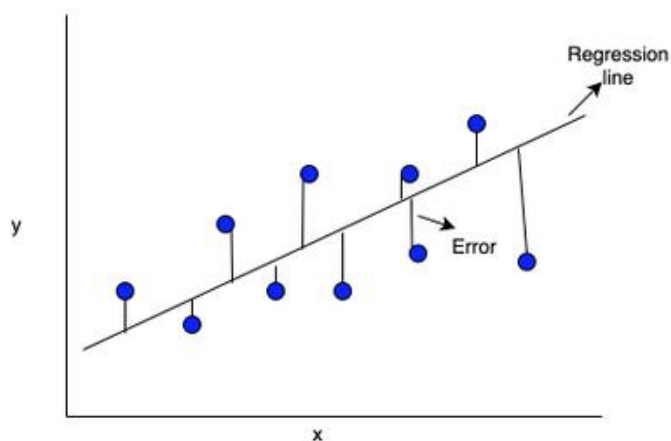
O objetivo do método dos mínimos quadrados é minimizar a soma dos quadrados das diferenças entre os valores observados de y e os valores estimados pela linha de regressão. Em outras palavras, ele busca encontrar a linha que melhor se ajusta aos dados observados, minimizando a soma dos quadrados dos erros de previsão.

O método dos mínimos quadrados é amplamente utilizado em áreas como estatística, engenharia, física, economia, finanças e outras áreas onde é necessário modelar a relação entre duas variáveis.

Interpretação Gráfica do Método dos Mínimos Quadrados

A interpretação gráfica do método dos mínimos quadrados envolve a análise do gráfico de dispersão dos dados e da reta de regressão ajustada.

O gráfico de dispersão é um gráfico bidimensional em que cada ponto representa uma observação de duas variáveis, geralmente a variável independente (x) no eixo horizontal e a variável dependente (y) no eixo vertical. A reta de regressão é a linha que melhor se ajusta aos dados, minimizando a soma dos quadrados dos erros de previsão.



Além disso, a reta de regressão passa pelo ponto médio dos dados (\bar{x} , \bar{y}), o que significa que a reta é um bom estimador da média da variável dependente para qualquer valor de x . A distância entre cada ponto e a reta de regressão representa o erro de previsão, que é a diferença entre o valor observado e o valor previsto pela reta de regressão. Portanto, a interpretação gráfica do método dos mínimos quadrados nos

permite visualizar a relação entre as variáveis e entender como a reta de regressão é construída para melhor descrever essa relação.

Método dos Mínimos Quadrados e a Elasticidade de Preços

O método dos mínimos quadrados também pode ser utilizado para calcular a elasticidade de preços, que é uma medida da sensibilidade da quantidade demandada de um produto em relação a uma mudança no preço.

A elasticidade de preços pode ser calculada usando a seguinte fórmula:

Elasticidade de preços = (Variação percentual na quantidade demandada) / (Variação percentual no preço)

Para calcular a elasticidade de preços usando o método dos mínimos quadrados, primeiro é necessário realizar uma regressão linear entre as variáveis preço e quantidade demandada. A equação da reta de regressão será da forma:

$$Q = a + bP$$

Onde Q é a quantidade demandada, P é o preço, a é o intercepto ou coeficiente linear e b é o coeficiente angular da reta de regressão, portanto descreve a inclinação da reta.

Como já discutimos em aulas anteriores vamos aproximar a reta de regressão a reta tangente descrita pela derivada que aparece na equação que determina a elasticidade de preços.

$$e = \frac{dQ}{dP} \frac{P_0}{Q_0}$$

Onde $\frac{dQ}{dP}$ é a derivada da equação da demanda em relação ao preço.

Determinação da Função de Regressão

A função de regressão é a equação matemática que descreve a relação entre a variável dependente (y) e a variável independente (x) de acordo com os dados observados.

Para encontrar a função de regressão usando o método dos mínimos quadrados, siga os seguintes passos:

1. Observe os dados: Primeiro, colete os dados e observe a relação entre as variáveis. Crie um gráfico de dispersão para visualizar a relação entre as variáveis.

2. Em seguida, determine a equação da reta de regressão usando o método dos mínimos quadrados. A equação da reta de regressão é da forma:

$$y = a + bx$$

Onde y é a variável dependente, x é a variável independente, a é o intercepto e b é o coeficiente angular.

Cálculo da Função de Regressão

Usaremos os dados da tabela um para entender como é calculado a função da reta de regressão.

X	Y
1	4
2	4
0	1
3	7

Tabela 1

A partir da tabela podemos criar um gráfico de dispersão e traçar a reta de regressão que melhor descreve o comportamento do fenômeno analisado.

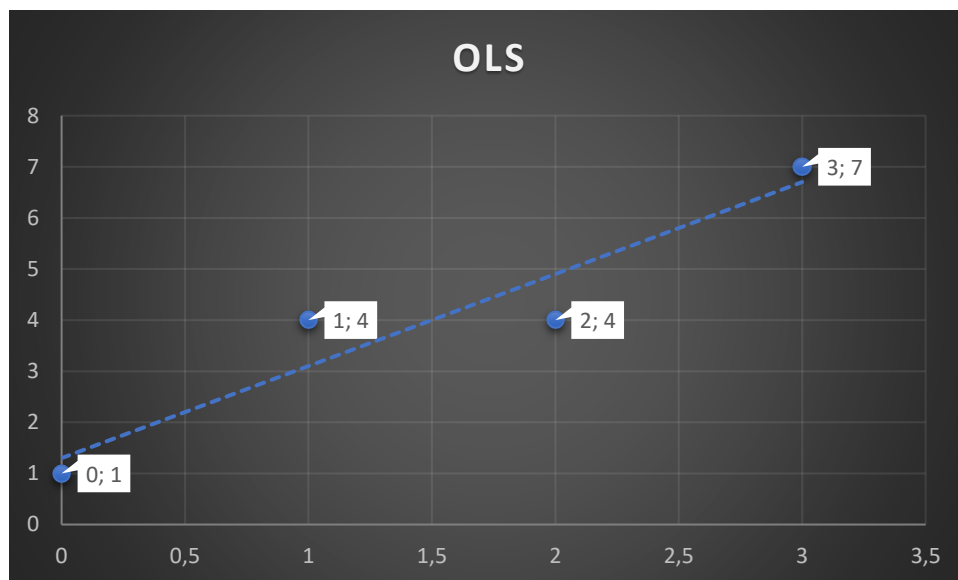


Gráfico 1

O método dos mínimos quadrados tenta traçar a reta minimizando o erro ao máximo, sendo o erro a distância entre o ponto e a reta, portanto o método tem que ter a menor distância possível entre os pontos e a reta. Para escrever a equação precisamos conceituar essa distância bidimensional entre dois pontos, dada por:

$$d = \sqrt{(\Delta x)^2 + (\Delta y)^2}$$

Sendo d a distância entre os pontos, Δx a variação da distância do ponto em x e Δy a variação em y . Observando o gráfico é possível perceber que a variação em x é sempre negativa, podemos usar então:

$$d = \Delta y$$

Sabendo agora calcular a distância dos pontos a reta ou o erro da regressão, podemos passar a discutir como calcular a função da reta. Queremos o valor mínimo da função que defina a e b através da soma dos quadrados das distâncias, definida por:

$$f(a, b) = d_1^2 + d_2^2 + d_3^2 + d_4^2$$

Temos 4 termos elevado ao quadrado, pois o gráfico apresenta 4 pontos, podemos ter equação com mais ou menos termos dependendo do fenômeno analisado.

Antes de começarmos a calcular ainda precisamos definir como calcular Δy , como já discutido a variação de x é zero pois assume apenas um valor, mas y assume valores diferentes definidos por $y = a + bx$, portanto:

$$d = a - y + bx$$

Queremos definir a e b , portanto são as nossas incógnitas e sabemos os valores de x e y da tabela 1.

$$f(a, b) = (a - y_1 + bx_1)^2 + (a - y_2 + bx_2)^2 + (a - y_3 + bx_3)^2 + (a - y_4 + bx_4)^2$$

Usando os valores conhecidos teremos:

$$f(a, b) = (a - 4 + b1)^2 + (a - 4 + b2)^2 + (a - 1 + b0)^2 + (a - 7 + b3)^2$$

Reescrevemos a função como:

$$f(a, b) = (a - 4 + b)^2 + (a - 4 + 2b)^2 + (a - 1)^2 + (a - 7 + 3b)^2$$

Por termos duas variáveis, a e b , para determinarmos os valores mínimos das variáveis precisamos encontrar o ponto crítico e determinar se ele é um mínimo local, para isso usamos a derivada segunda em relação a a e em relação a b e igualamos a 0.

$$\frac{\partial f(a, b)}{\partial a} = 0 \quad \frac{\partial f(a, b)}{\partial b} = 0$$

Teremos então:

$$\begin{aligned}\frac{\partial f(a,b)}{\partial a} &= 2.(a-4+b).1 + 2.(a-4+2b).1 + 2.(a-1).1 \\ &\quad + 2.(a-7+3b).1 \\ 2a-8+2b+2a-8+4b+2a-2+2a-14+6b &= 0 \\ 8a+12b-32 &= 0\end{aligned}$$

$$\begin{aligned}\frac{\partial f(a,b)}{\partial b} &= 2.(a-4+b).1 + 2.(a-4+2b).2 + 2.(a-1).0 \\ &\quad + 2.(a-7+3b).3 \\ 2a-8+2b+4a-16+8b+0+6a-42+18b &= 0 \\ 12a+28b-66 &= 0\end{aligned}$$

A partir dos resultados obtidos construímos um sistema de duas equações

$$8a + 12b - 32 = 0$$

$$12a + 28b - 66 = 0 \times (-1)$$

Obtemos a seguinte equação como resposta:

$$-4a - 16b + 34 = 0 \rightarrow a = \frac{-16b+34}{4}$$

Voltamos a uma das equações resultado das derivadas parciais e substituímos o valor de a:

$$\begin{aligned}8 \frac{-16b+34}{4} + 12b - 32 &= 0 \rightarrow -32b + 68 + 12b - 32 = 0 \rightarrow -20b + 36 = 0 \\ b &= 1,8\end{aligned}$$

Reutilizamos a equação que tem o valor de a isolado

$$a = \frac{-16.(1,8) + 34}{4} \rightarrow a = 1,3$$

Enfim encontramos a equação que descreve a reta do gráfico 1

$$y = 1,3 + 1,8x$$

Cálculo da Função de Regressão Generalista

O método generalista usa as duas equações a seguir para determinar o coeficiente linear a e o coeficiente angular b .

$$a = \frac{\Sigma y - b \Sigma x}{n}$$

$$b = \frac{n \Sigma x_i y_i - \Sigma x_i \Sigma y_i}{n \Sigma x_i^2 - (\Sigma x_i)^2}$$

Usaremos os mesmos dados do método anterior, mas apresentados da seguinte forma:

X	Y	X ²	X.Y
1	4	1	4
2	4	4	8
0	1	0	0
3	7	9	21
Somatório			
6	16	14	33

Resolvendo as equações para a e b :

$$b = \frac{4.33 - 6.16}{4.14 - (6)^2} \rightarrow b = \frac{132 - 96}{56 - 36} \rightarrow b = \frac{36}{20} \rightarrow b = 1,8$$

$$a = \frac{16 - 1,8.6}{4} \rightarrow a = 1,3$$

Como esperado encontramos a equação que descreve a reta do gráfico 1, da seguinte forma:

$$y = 1,3 + 1,8x$$

Previsão Realizada usando o Método dos Mínimos Quadrados

A tabela abaixo apresenta os valores reais representados por Y e os valores previstos pela reta de regressão representados por \bar{Y} .

Y	\bar{Y}
4	3,1
4	4,9
1	1,3
7	6,7

Coeficiente De Determinação (R^2)

O coeficiente de determinação (r^2) é uma medida estatística utilizada para avaliar a qualidade da regressão linear por mínimos quadrados. Ele indica o quão bem os pontos de dados se ajustam à linha de regressão.

O r^2 varia de 0 a 1, sendo 0 indicando que nenhuma variação nos dados é explicada pela linha de regressão e 1 indicando que toda a variação é explicada pela linha de regressão. Portanto, quanto maior o r^2 , melhor é o ajuste da linha de regressão aos dados.

O r^2 pode ser calculado a partir da fórmula:

$$r^2 = 1 - \left(\frac{SSE}{SST} \right)$$

Onde SSE (soma dos quadrados dos erros) é a soma das diferenças entre os valores reais e os valores previstos pela linha de regressão, ao quadrado. Matematicamente, podemos escrever:

$$SSE = \sum (Y - \bar{Y})^2$$

E SST (soma total dos quadrados) é a soma dos quadrados das diferenças entre os valores reais e a média dos valores reais. Matematicamente, podemos escrever:

$$SST = \sum \left(Y - \frac{\sum Y}{n} \right)^2$$

O r^2 é uma medida importante para avaliar a qualidade da regressão linear, mas não deve ser usada como a única medida de avaliação. É importante também avaliar a significância estatística da linha de regressão e considerar outros fatores relevantes ao contexto da análise.

Cálculo de r^2

Para calcular o coeficiente de determinação r^2 , primeiro precisamos calcular a soma dos quadrados dos erros da regressão (SSE) e a soma total dos quadrados (SST).

Podemos calcular a SSE somando os quadrados das diferenças entre os valores observados Y e os valores previstos \bar{Y} :

$$SSE = (4 - 3,1)^2 + (4 - 4,9)^2 + (1 - 1,3)^2 + (7 - 6,7)^2$$

$$SSE = 0,81 + 0,81 + 0,09 + 0,09$$

$$SSE = 1,8$$

Podemos calcular a SST somando os quadrados das diferenças entre os valores observados Y e a média dos valores observados \bar{Y} :

$$Y_{mean} = \frac{4 + 4 + 1 + 7}{4} = 4$$

$$SST = (4 - 4)^2 + (4 - 4)^2 + (1 - 4)^2 + (7 - 4)^2$$

$$SST = 0 + 0 + 9 + 9$$

$$SST = 18$$

Agora que temos SSE e SST, podemos calcular o coeficiente de determinação r^2 :

$$r^2 = 1 - \left(\frac{1,8}{18}\right)$$

$$r^2 = 0.9$$

Portanto, o coeficiente de determinação r^2 para os dados fornecidos é de aproximadamente 0.9, indicando que a linha de regressão se ajusta muito bem aos dados observados. Esse valor indica que aproximadamente 90% da variabilidade nos dados pode ser explicada pela relação linear entre as variáveis independentes e dependentes.

Em outras palavras, a linha de regressão prevê com precisão 90% da variação na variável dependente a partir da variação na variável independente.

Significância Estatística Da Reta De Regressão

O método OLS (Ordinary Least Squares) do Statsmodels compara a variabilidade explicada pelo modelo com a variabilidade não explicada, ou seja, a diferença entre a variância total dos dados e a variância explicada pelo modelo. É um valor de teste estatístico que indica a significância geral do modelo de regressão, esse valor é calculado a partir de um teste F de Fisher-Snedeco com resultado acessível pelo atributo `f_pvalue` que retorna o p-valor do teste.

Em termos simples, o `f_pvalue` indica se o modelo de regressão como um todo é estatisticamente significativo ou não. Ele é um valor de probabilidade que varia entre 0 e 1, e quanto menor for o valor, mais significativo é o modelo. Um `f_pvalue` baixo indica que a probabilidade de o modelo ter sido ajustado aos dados por acaso é baixa, o que aumenta a confiança em seus resultados.

$$F = \frac{\left(\frac{SSR}{k}\right)}{\left(\frac{SSE}{(n - k - 1)}\right)}$$

Sendo SSR é a soma dos quadrados da regressão, que é a variância explicada pelo modelo de regressão.

$$SSR = SST - SSE$$

k é o número de coeficientes independentes no modelo. SSE Já definimos e calculamos acima e n é o número total de observações.

Portanto:





$$SSR = 18 - 1,8 = 16,2$$

Como possuímos apenas uma variável independente k=1 e o temos 4 observações então n=4

$$F = \frac{\left(\frac{16,2}{1}\right)}{\left(\frac{1,8}{(4 - 1 - 1)}\right)} \rightarrow \frac{16,2}{0,9} = 18$$

Com esse valor é necessário buscar nas tabelas para definir qual o p-valor dessa análise, se tivermos o p-valor menor 0,05 podemos rejeitar a hipótese nula de que todos os coeficientes de regressão são iguais a zero ou dizer que o teste tem significância estatística, com esse valor de f-statistic e com 1 grau de liberdade no numerador e 2 graus de liberdade no denominador encontramos um valor crítico de 9,55, como esse valor é menor que 18 o p-valor é 0,05 ou menor.

Relação entre r^2 e o p-valor

R^2 x p-valor	
alto	 "Modelo explica muita variação nos dados e é significativo (melhor cenário)"
R^2	 "Modelo explica muita variação nos dados mas não é significativo (modelo é inútil)"
baixo	 "Modelo não explica muita variação nos dados mas é significativo (melhor que não ter um modelo)"
	 "Modelo não explica muita variação nos dados e não é significativo (pior cenário)"
	baixo $\leq 0,05$ p-valor alto $> 0,05$

@proffernandamaciel

Conclusão

Podemos dizer então que a nossa reta de regressão e consequentemente a equação que define essa reta possuem forte significância estatística, essa afirmação se ancoram nos resultados de $r^2 = 0.9$ e p-valor $< 0,05$.

A seguir faremos essa mesma análise para o nosso conjunto de dados mas usando método OLS (Ordinary Least Squares) da biblioteca Statsmodels. Definiremos a reta de regressão para os dados de preço e demanda, através das análises aprendidas nessa aula poderemos ter segurança estatística para determinarmos se a reta realmente representa o fenômeno em questão.