

Regressão múltipla

Quando o passado permite prever o futuro, há um método estatístico denominado Regressão Linear que busca o melhor ajuste para um conjunto de dados tentando minimizar a soma dos quadrados das diferenças entre o valor estimado e os dados observados.

Em estatística, regressão é uma técnica que permite explorar e inferir a relação de uma variável dependente (variável de resposta) com variáveis independentes específicas (variáveis explicatórias). A análise da regressão pode ser usada como um método descritivo da análise de dados (como, por exemplo, o ajustamento de curvas) sem serem necessárias quaisquer suposições acerca dos processos que permitiram gerar os dados. Regressão designa também uma equação matemática que descreva a relação entre duas ou mais variáveis

Quando há apenas 2 variáveis em análise diz-se "regressão simples" e quando há mais do que 2 variáveis envolvidas, diz-se "regressão múltipla".

Na fórmula original tem-se

y = Xb + e

, que deve ser lido como o parametro matricial b é o que ajusta os y e os x observados.

A fórmula de cálculo de b é b = (X'X)⁻¹X'y, onde X' é a matriz transposta de X e X⁻¹ é a matriz inversa de X.

Acompanhe o exemplo

Suponhamos querer estudar o comportamento do consumo de bens em uma determinada família, sujeita a dois parâmetros:

- renda e
- taxa de juros

Dos estudos de econometria, sabe-se que o consumo aumenta em função da renda e diminui em função da taxa de juros.

O trabalho começa pelo levantamento de dados prévios e obteve-se a seguinte tabela:

medida	consumo y	renda x ₁	tx juros x ₂
1	122	139	11,5%
2	114	126	12,0%
3	86	90	10,5%
4	134	144	9,0%
5	146	163	10,0%
6	107	136	12,0%
7	68	61	10,5%
8	117	62	8,0%
9	71	41	10,0%
10	98	120	11,5%

A matriz X do nosso problema é:

1	139	0.115
1	126	0.12
1	90	0.105
1	144	0.09
1	163	0.1
1	136	0.12
1	61	0.105
1	62	0.08
1	41	0.1
1	120	0.115

Note que a primeira coluna sempre é formada de unidades e as demais colunas são as variáveis em estudo (no caso: renda e taxa de juros).

Tomando a matriz transposta fica-se com

1	1	1	1	1	1	1	1	1	1
139	126	90	144	163	136	61	62	41	120
0.11	0.12	0.10	0.09	0.1	0.12	0.10	0.08	0.1	0.11

Multiplicando uma pela outra, tem-se

10	1082	1.05
139	132744	115.4
1.05	115.4	0.1118

E a matriz inversa desta é

7.225436859 0.00095983853 -68.85039418
0.00095983853 0.00007350542 -0.084886905
-68.85039418 -0.08488690598 743.1919753

Agora, multiplicando a inversa por X' fica:

-5	-.91	.08	1.16	.49	-.90	.05	1.7	.3	-.5
.0	.00	-.00	.00	.00	.0	-.00	-.0	-.0	.0
4.8	9.63	1.54	-14.1	-8.36	8.7	4.0	-14.6	1.9	6.4

Agora multiplicando esta matriz pela matriz coluna de y (consumo) que é (devidamente horizontalizada para economizar papel)

122 114 86 134 146 107 68 117 71 98

Fica

148.5220194
0.6135933813
-1034.40784

Interpretação

Neste problema, o consumo é

Consumo =

148,52 + 0.6135 × Renda - 1.034.41 × TxJuros

Ou seja, tirando a parte do Consumo que não é influenciada pela Taxa de Juros, o incremento de \$ 1 na Renda causa um incremento esperado de \$ 0,6136 no Consumo; além disso, o incremento de 1 ponto percentual (0,01) na Taxa de Juros causa um decréscimo esperado de \$ 10,3441 no Consumo.

Para você fazer

1. Considere o mesmo problema com os seguintes valores obtidos

medida	consumo y	renda x ₁	tx juros x ₂
1	92	129	.1200
2	98	121	.1170
3	87	128	.1040
4	96	106	.1170
5	81	116	.1000
6	84	120	.0950
7	107	104	.1140
8	103	128	.1160
9	89	118	.1150
10	109	108	.1050

Ache os parâmetros de renda e taxa de juros para o problema proposto.

t. fixo	renda	tx de juros
---------	-------	-------------

2. Mais um problema, agora envolvendo um outro modelo econométrico (obviamente FICTICIO) no qual estão relacionadas 3 variáveis independentes e uma dependente. Não se esqueça que a matemática lava as mãos sobre essas definições. É coisa do planejador. São elas:

t	dep.	consumo de whiskie importado (em caixas)
x	indep.	venda de propriedades em Caiobá (em um)
y	indep.	número de falências (sentido oposto, em um)
z	indep.	número de BMW licenciadas (em um)

O objetivo então é achar

t = a + b.x + c.y + d.z

No último ano foram levantados os valores de x, y e z e obteve-se

medida	t	x	y	z
1	-1468792	24	856	567
2	1482889	620	837	716
3	2579824	264	200	427
4	-6341432	-6	906	-38
5	5533480	60	122	838
6	-481927	656	389	28
7	-318727	160	831	647
8	2388515	797	797	733
9	3041105	168	597	888
10	2725889	652	797	836

Usando a teoria acima, você deve achar os valores de a, b, c e d.

a	b	c	d

3. O problema agora é que você é o dono do maior restaurante de Santa Felicidade. Precisa prever a quantidade de refeições a serem servidas no próximo Dia das Mães. Depois de muito estudar, você contratou um engenheiro que levantou dados dos últimos 6 anos e elaborou o seguinte modelo econométrico:

a	indep	número de refeições a serem servidas no próximo dia das mães
b	dep	número de casamentos em Curitiba no ano anterior
c	dep	renda média das famílias no Paraná no ano anterior, em reais
d	dep	frangos abatidos na última semana do mês de abril no Paraná
e	dep	veículos emplacadas na ciretran de S. Felicidade no ano anterior

A fórmula do modelo é:

a = α + βb + γc + δd + ε

Os dados:

	a	b	c	d	e
1	2312	13854	12906	104827	4990
2	2774	14196	8856	127368	6497
3	2196	14273	11719	93277	6847
4	3007	16088	7437	134054	8717
5	2598	17696	9114	111180	8509
6	2592	18473	13569	108764	8693

Achar os valores de α, β, γ, δ e ε.

α	β	γ	δ	ε

Observação importante: No Python, ao gerar as matrizes, não esquecer de pedir campos float, porque senão ele pode resolver usando inteiros e vai achar respostas erradas...