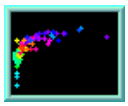


Regressão

Prof. Lorí Viali, Dr.
vialli@mat.ufrgs.br
http://www.mat.ufrgs.br/~vialli/

Em muitas situações duas ou mais variáveis estão relacionadas e surge então a necessidade de determinar a natureza deste relacionamento.

A análise de regressão é uma técnica estatística para modelar e investigar o relacionamento entre duas ou mais variáveis.

De fato a regressão pode ser dividida em dois problemas:

(i) o da especificação e

(ii) o da determinação.

A especificação

O problema da especificação é descobrir dentre os possíveis modelos (linear, quadrático, exponencial, etc.) qual o mais adequado.

A determinação

O problema da determinação é uma vez definido o modelo (linear, quadrático, exponencial, etc.) estimar os parâmetros da equação.

O modelo

Normalmente é suposto que exista uma variável Y (dependente ou resposta), que está relacionada a "k" variáveis (independentes ou regressoras) X_i (i = 1, 2, ..., k).

*A variável resposta **Y** é aleatória, enquanto que as variáveis regressoras X_i são normalmente **controladas**. O relacionamento entre elas é caracterizado por uma equação denominada de "**equação de regressão**"*

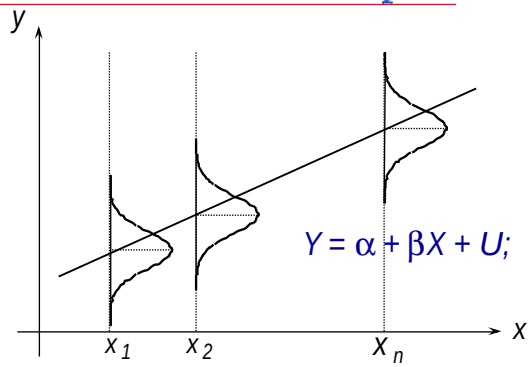
*Quando existir apenas uma variável regressora (X) tem-se a **regressão simples**, se Y depender de duas ou mais variáveis regressoras, então tem-se a "**regressão múltipla**".*

O modelo considerado

*Vamos supor que a regressão é do tipo **simples** e que o o modelo seja **linear**, isto é, vamos supor que a equação de regressão seja do tipo:*

$$Y = \alpha + \beta X + U$$

O modelo linear simples



O termo "U" é o termo erro, isto é, "U" representa outras influências sobre a variável Y, além da exercida pela variável "X". A variação residual (termo U) é suposto de média zero e desvio constante e igual a σ.

Ou ainda pode-se admitir que o modelo fornece o valor médio de Y, para um dado "x", isto é,

$$E(Y/x) = \alpha + \beta X$$

Em resumo, as hipóteses são:

$Y = \alpha + \beta X + U$;

$E(Y/x) = \alpha + \beta X$, isto é, $E(U) = 0$

$V(Y/x) = \sigma^2$;

$Cov(Ui, Uj) = 0$, para $i \neq j$;

A variável X permanece fixa em observações sucessivas e os erros U são normalmente distribuídos.

A equação de regressão

O modelo suposto **$E(Y/x) = \alpha + \beta X$** é populacional.

Vamos supor que se tenha n pares de observações, digamos: **$(x_1, y_1), (x_2, y_2), \ldots, (x_2, y_2)$** e que através deles queremos estimar o modelo acima.

A reta estimada será representada por: $\hat{Y} = a + bX$ ou $Y = a + bX + E$

Onde "**a**" é um estimador de α e "**b**" é um estimador de β, sendo $\hat{Y}$ um estimador de E(Y/x).

O método utilizado

Existem diversos métodos para a determinação da reta desejada. Um deles, denominado de **MMQ** (**M**étodos dos **M**ínimos **Q**uadrados), consiste em minimizar a "**soma dos quadrados das distâncias da reta aos pontos**".

Tem-se:

$$Y_i = a + bx_i + E_i,$$

Então:

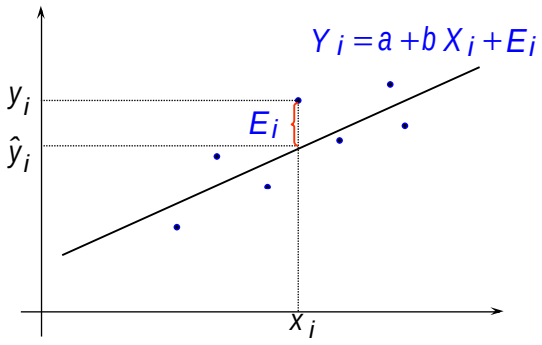
$$E_i = Y_i - (a + bx_i)$$

Deve-se minimizar:

$$\phi = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 =$$

$$= \sum_{i=1}^{n} (Y_i - a - b X_i)^2$$

O método dos mínimos quadrados



Derivando parcialmente tem-se:

$$\frac{\partial \phi}{\partial a} = -2 \sum_{i=1}^{n} (Y_i - a - b X_i)$$

$$\frac{\partial \phi}{\partial b} = -2 \sum_{i=1}^{n} x_i (Y_i - a - b X_i)$$

Igualando as derivadas parciais a zero vem:

$$\sum_{i=1}^{n} (Y_i - a - b X_i) = 0$$

$$\sum_{i=1}^{n} x_i (Y_i - a - b X_i) = 0$$

Isolando as incógnitas, tem-se:

$$\sum Y_i = na + b \sum X_i$$

$$\sum X_i Y_i = n \sum X_i + b \sum X_i^2$$

Resolvendo para "a" e "b", segue:

$$b = \frac{\sum X_i y_i - n \overline{X}\,\overline{Y}}{\sum X_i^2 - n \overline{X}^2} = \frac{S_{XY}}{S_{XX}}$$

$$a = \overline{Y} - b \overline{X}$$

Lembrando que:

$$S_{XY} = \sum X_i Y_i - n\bar{X}\bar{Y}$$

$$S_{XX} = \sum X_i^2 - n\bar{X}^2$$

$$S_{YY} = \sum Y_i^2 - n\bar{Y}^2$$

Exemplo

Um engenheiro químico está investigando o efeito da temperatura de operação do processo no rendimento do produto. O estudo resultou nos dados da tabela, ao lado. Determinar a linha de regressão.

Temperatura, C^0 (X)	Rendimento (Y)
100	45
110	51
120	54
130	61
140	66
150	70
160	74
170	78
180	85
190	89

Da mesma forma que para calcular o coeficiente de correlação é necessário a construção de três novas colunas. Uma para X^2, uma para Y^2 e outra para XY.

X	Y	XY	X^2	Y^2
100	45	4500	10000	2025
110	51	5610	12100	2601
120	54	6480	14400	2916
130	61	7930	16900	3721
140	66	9240	19600	4356
150	70	10500	22500	4900
160	74	11840	25600	5476
170	78	13260	28900	6084
180	85	15300	32400	7225
190	89	16910	36100	7921
1450	**673**	**101570**	**218500**	**47225**

Tem-se: $n=10 \quad \sum X=1450 \quad \sum Y=673$

$\bar{X}=145 \quad \bar{Y}=67,3 \quad \sum XY=101570$

$\sum X^2=218500 \quad \sum Y^2=47225$

Então: $S_{XY}=\sum X_i Y_i - n\bar{X}\bar{Y} =$

$=101570-10.145.67,3=$

$=3985$

$S_{XX}=\sum X_i^2 - n\bar{X}^2 =$

$=218500-10.145^2=$

$=8250$

$S_{YY}=\sum Y_i^2 - n\bar{Y}^2 =$

$=47225-10.67,3^2=$

$=1932,10$

A equação de regressão, será, então:

$$b=\frac{S_{XY}}{S_{XX}}=\frac{3985}{8250}=0,4830 \cong 0,48$$

$$a=\bar{Y}-b\bar{X}=67,30-0,4830.145=$$

$$=-2,7394 \cong -2,74$$

$$\hat{Y}=-2,74+0,48x$$

A pergunta que cabe agora é: este modelo representa bem os pontos dados? A resposta é dada através do erro padrão da regressão.

Variância Residual e Erro Padrão da Regressão

O objetivo do MMQ é minimizar a variação residual em torno da reta de regressão. Uma avaliação desta variação é dada por:

$$S^2=\frac{\sum E^2}{n-2}=\frac{\sum(Y-a-bX)^2}{n-2}$$

O cálculo da variância residual, por esta expressão, é muito trabalhoso, pois é necessário primeiro determinar os valores previstos. Entretanto é possível obter uma expressão que não requeira o cálculo dos valores previstos, isto é, de $\hat{Y} = a + bX$

Erro padrão da regressão

$$s = \sqrt{\frac{\sum E^2}{n-2}} = \sqrt{\frac{\sum (Y - a - bX)^2}{n-2}} =$$

$$= \sqrt{\frac{S_{YY} - b^2 S_{XX}}{n-2}} = \sqrt{\frac{S_{YY} - b S_{XY}}{n-2}}$$

Exemplo

Considerando os valores do exemplo anterior, determinar o erro padrão da regressão.

Tem-se: $S_{YY} = 1932,10 \quad S_{XX} = 8250$

$$b = \frac{S_{XY}}{S_{XX}} = \frac{3985}{8250} = 0,4830$$

Então:

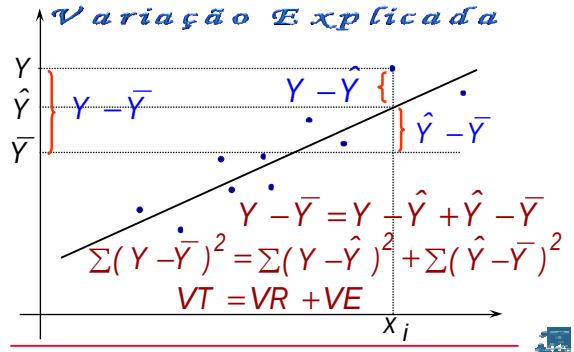
$$s = \sqrt{\frac{S_{YY} - b S_{XY}}{n-2}} =$$

$$= \sqrt{\frac{1932,10 - \frac{3985}{8250}.3985}{10-2}} =$$

$$= 0,9503 \cong 0,95$$

Decomposição da Variação

Variação Total = Variação Não-Explicada + Variação Explicada



$$Y - \bar{Y} = Y - \hat{Y} + \hat{Y} - \bar{Y}$$

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

$$VT = VR + VE$$

(a) Variação Total: VT

$$VT = \sum (Y - \bar{Y})^2 = S_{YY}$$

(b) Variação Residual: VR

$$VR = \sum (Y - \hat{Y})^2 = S_{YY} - b^2 S_{XX} = VT - VE$$

(c) Variação Explicada: VE

$$VE = \sum (\hat{Y} - \bar{Y})^2 = b^2 S_{XX}$$

Uma maneira de medir o grau de aderência (adequação) de um modelo é verificar o quanto da variação total de Y é explicada pela reta de regressão.

Para isto, toma-se o quociente entre a variação explicada, VE, pela variação total ,VT:

$$R^2 = VE / VT$$

*Este resultado é denominado de "**Coeficiente de Determinação**".*

$$R^2 = \frac{VE}{VT} = \frac{b^2 S_{XX}}{S_{YY}} = \frac{b\, S_{XY}}{S_{YY}} = \frac{S_{XY}^2}{S_{XX} S_{YY}}$$

Este resultado mede o quanto as variações de uma das variáveis são explicadas pelas variações da outra variável.

Ou ainda, ele mede a parcela da variação total que é explicada pela reta de regressão, isto é:

$$VE = b^2 S_{XX} = R^2 S_{YY}$$

A variação residual corresponde a:

$$VR = (1 - R^2) S_{YY}$$

Assim $1 - R^2$ é o Coeficiente de Indeterminação.

Exercício

O % de impurezas no gás oxigênio produzido por um processo de destilação supõem-se que esteja relacionado com o % de hidrocarbono no condensador principal do processador. Os dados de um mês de operação produziram a seguinte tabela:

X	Y	X	Y
1,02	86,91	1,46	96,73
1,11	89,85	1,55	99,42
1,43	90,28	1,55	98,66
1,11	86,34	1,55	96,07
1,01	92,58	1,40	93,65
0,95	87,33	1,15	87,31
1,11	86,29	1,01	95,00
0,87	91,86	0,99	96,85
1,43	95,61	0,95	85,20
1,02	89,86	0,98	90,56

(a) *Ajuste um modelo linear aos dados;*

(b) *Determine o valor de R^2 para este modelo;*

Solução

Dados

$n = 20$

$\sum X = 23,65$ $\quad \bar{X} = 1,1825$

$\sum X^2 = 29,0311$ $\quad \bar{Y} = 91,8180$

$\sum Y = 1836,36$ $\quad S_{XX} = 1,064975$

$\sum Y^2 = 168992,0498$ $\quad S_{XY} = 12,5678$

$\sum XY = 2184,0635$ $\quad S_{YY} = 381,147320$

(a) *A equação de regressão, será, então:*

$$b = \frac{S_{XY}}{S_{XX}} = \frac{12,5678}{1,064975} = 11,8010 \cong 11,80$$

$$a = \bar{Y} - b\bar{X} = 91,8180 - 11,8010.1,1825 =$$

$$= 77,8633 \cong 77,86$$

$$\hat{Y} = 77,86 + 11,80\,x$$

(b) *O coeficiente de determinação*

$$R^2 = \frac{S_{XY}^2}{S_{XX}\,S_{YY}} =$$

$$= \frac{12,5678^2}{1,0650.3811473} = 38,91\%$$

Ou seja, 38,91% das variações em Y são explicadas pelas variações em x

Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística