Correlação

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

É o grau de associação entre duas ou mais variáveis. Pode ser:

correlacional

ou

experimental.

Numa relação experimental os valores de uma das variáveis são controlados.

No relacionamento correlacional, por outro lado, não se tem nenhum controle sobre as variáveis sendo estudadas.

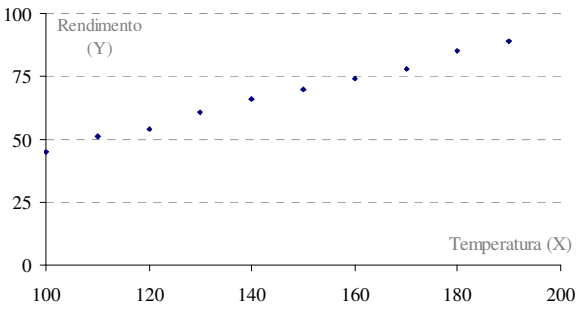
Indicadores de Associação

Um engenheiro químico está investigando o efeito da temperatura de operação do processo no rendimento do produto. O estudo resultou nos dados da tabela seguinte:

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

O primeiro passo para determinar se existe relacionamento entre as duas variáveis é obter o **diagrama de dispersão** (scatter diagram).

Diagrama de Dispersão



O diagrama de dispersão fornece uma idéia do tipo de relacionamento entre as duas variáveis. Neste caso, percebe-se que existe um **relacionamento linear**.

Quando o relacionamento entre duas variáveis quantitativas for do tipo **linear**, ele pode ser medido através do:

Coeficiente de Correlação

Observado um **relacionamento linear** entre as duas variáveis é possível determinar a intensidade deste relacionamento. O coeficiente que mede este relacionamento é denominado de Coeficiente de Correlação (linear).

*Quando se está trabalhando com amostras o coeficiente de correlação é indicado pela letra "**r**" e é uma estimativa do coeficiente de correlação populacional que é representado por "ρ" (rho).*

Determinação do Coeficiente de Correlação

Para determinar o coeficiente de correlação (grau de relacionamento linear entre duas variáveis) vamos determinar inicialmente a variação conjunta entre elas, isto é, a covariância.

*A covariância entre duas variáveis X e Y, é representada por "**Cov(X; Y)**" e calculada por:*

$$Cov(X,Y) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}$$

Mas

$$\sum (X_i - \overline{X})(Y_i - \overline{Y}) =$$

$$= \sum [X_i Y_i - \overline{X} Y_i - X_i \overline{Y} + \overline{X}\overline{Y}] =$$

$$= \sum X_i Y_i - \sum \overline{X} Y_i - \sum X_i \overline{Y} + \sum \overline{X}\overline{Y} =$$

$$= \sum X_i Y_i - \overline{X} \sum Y_i - \overline{Y} \sum X_i + \sum \overline{X}\overline{Y} =$$

$$= \sum X_i Y_i - n\overline{X}\overline{Y} - n\overline{X}\overline{Y} + n\overline{X}\overline{Y} =$$

$$= \sum X_i Y_i - n\overline{X}\overline{Y}$$

Então:

$$Cov(X,Y) = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{n-1} =$$

$$= \frac{\sum X_i Y_i - n\overline{X}\overline{Y}}{n-1}$$

A covariância poderia ser utilizada para medir o ***grau*** *e o* ***sinal*** *do relacionamento entre as duas variáveis, mas ela é difícil de interpretar por variar de -∞ a +∞. Assim vamos utilizar o* ***coeficiente de correlação linear de Pearson.***

O coeficiente de correlação linear (de Pearson) é definido por:

$$r = \frac{Cov(X,Y)}{S_X S_Y}$$

Onde:

$$Cov(X,Y) = \frac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{n-1}$$

$$S_X = \sqrt{\frac{\sum X_i^2 - n\overline{X}^2}{n-1}}$$

$$S_Y = \sqrt{\frac{\sum Y_i^2 - n\overline{Y}^2}{n-1}}$$

Esta expressão não é muito prática para calcular manualmente o coeficiente de correlação. Pode-se obter uma expressão mais conveniente para o cálculo manual e o cálculo de outras medidas necessárias mais tarde.

Tem-se:

$$r = \frac{Cov(X,Y)}{S_X S_Y} = \frac{\dfrac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{n-1}}{\sqrt{\dfrac{\sum X_i^2 - n\overline{X}^2}{n-1}}\sqrt{\dfrac{\sum Y_i^2 - n\overline{Y}^2}{n-1}}} = \frac{\sum X_i Y_i - n\overline{X}\,\overline{Y}}{\sqrt{\left(\sum X_i^2 - n\overline{X}^2\right)\left(\sum Y_i^2 - n\overline{Y}^2\right)}}$$

Fazendo

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

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

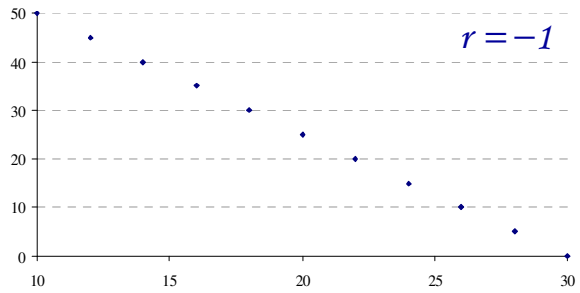
$$S_{YY} = \sum Y_i^2 - n\overline{Y}^2$$

$$Tem-se: \quad r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}}$$

A vantagem do coeficiente de correlação (de Pearson) é ser adimensional e variar de – 1 a + 1, que o torna de fácil interpretação.

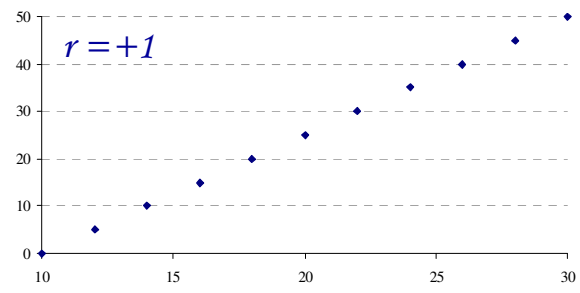
Assim se r = -1, temos uma relacionamento linear negativo perfeito, isto é, os pontos estão todos alinhados e quando X aumenta Y decresce e vice-versa.

Correlação perfeita e negativa



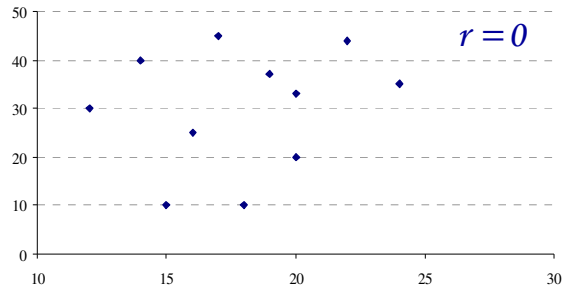
Se r = +1, temos uma relacionamento linear positivo perfeito, isto é, os pontos estão todos alinhados e quando X aumenta Y também aumenta.

Correlação perfeita e positiva



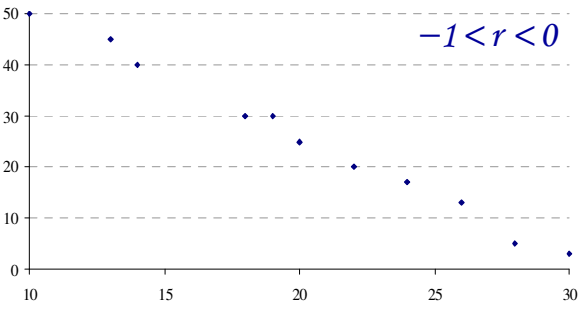
Assim se r = 0, temos uma ausência de relacionamento linear, isto é, os pontos não mostram "alinhamento".

Correlação nula



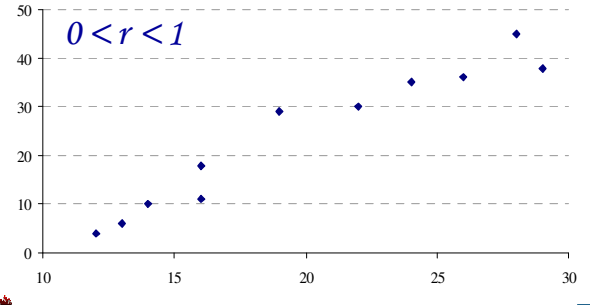
Assim se $-1 < r < 0$, temos uma relacionamento linear negativo, isto é, os pontos estão mais ou menos alinhados e quando X aumenta Y decresce e vice-versa.

Correlação negativa



Assim se $0 < r < 1$, temos uma relacionamento linear positivo, isto é, os pontos estão mais ou menos alinhados e quando X aumenta Y também aumenta.

Correlação positiva



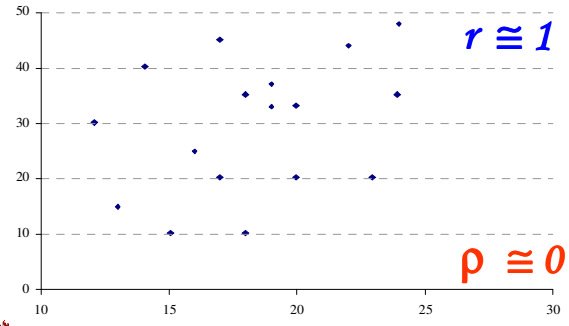
Observação:

Uma correlação amostral não significa necessariamente uma correlação populacional e vice-versa. É necessário testar o coeficiente de correlação para verificar se a correlação amostral é também populacional.

Ilustração

Observada uma amostra de seis pares, pode-se perceber que a correlação é quase um, isto é, $r \cong 1$. *No entanto, observe o que ocorre quando mais pontos são acrescentados, isto é, quando se observa a população!*

Correlação amostral X populacional



Exemplo

Determinar o "grau de relacionamento linear" entre as variáveis X = temperatura de operação do processo versus Y = rendimento do produto, conforme tabela.

X	Y	XY	X	Y
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**

Vamos calcular "r" utilizando a expressão em destaque vista anteriormente, isto é, através das quantidades, S_{XY}, S_{XX} *e* S_{YY}.

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

$= 101570 - 10.145.67,3 =$

$= 3985$

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

$= 218500 - 10.145^2 =$

$= 8250$

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

$= 47225 - 10.67,3^2 =$

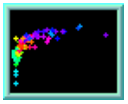
$= 1932,10$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} =$$

$$= \frac{3985}{\sqrt{8250.1932,10}} =$$

$$= 0,9981$$

*Apesar de "**r**" ser um valor adimensional, ele não é uma **taxa**. Assim o resultado não deve ser expresso em percentagem.*



Regressão

Prof. Lori Viali, Dr.
viali@mat.ufrgs.br
http://www.mat.ufrgs.br/~viali/

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**"*

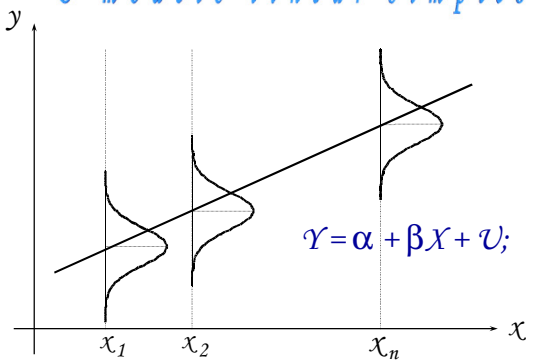
Quando existir apenas uma variável regressora (X) tem-se a **regressão simples**, se Y depender de duas ou mais variáveis regressoras, então tem-se a "**regressão múltipla**".

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: $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ e que através deles queremos estimar o modelo acima.

A equação de regressão

A reta estimada será representada por:

$$\hat{Y} = a + bX \quad ou \quad Y = a + bX + E$$

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:

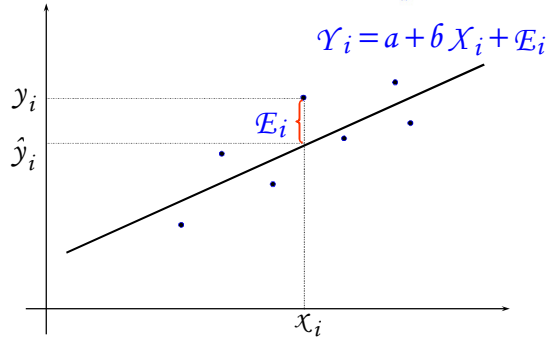
$$E_i = Y_i - (a + bx_i)$$

Deve-se minimizar:

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

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

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:

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

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

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

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

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

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.

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

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

$= 101570 - 10.145.67,3 =$

$= 3985$

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

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

$= 218500 - 10.145^2 =$

$= 8250$

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

$= 47225 - 10.67,3^2 =$

$= 1932,10$

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

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 $\hat{Y} = a + bX$

Desenvolvendo o numerador da expressão, vem:

$$\sum (Y - a - bX)^2 = \sum [Y - (\bar{Y} - b\bar{X}) - bX]^2 =$$

$$= \sum [Y - \bar{Y} + b\bar{X} - bX]^2 = \sum [Y - \bar{Y} - b(X - \bar{X})]^2 =$$

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

$$= S_{YY} - 2bS_{XY} + b^2 S_{XX}$$

Uma vez que:

$$\sum (X - \bar{X})(Y - \bar{Y}) =$$

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

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

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

Deste modo, tem-se:

$$\sum (Y - a - bX)^2 = S_{YY} - 2b S_{XY} + b^2 S_{XX}$$

Mas: $b = \dfrac{S_{XY}}{S_{XX}} \Rightarrow S_{XY} = b\, S_{XX}$

Então:

$$\sum (Y - a - bX)^2 = S_{YY} - 2b S_{XY} + b^2 S_{XX} =$$

$$= S_{YY} - 2b^2 S_{XX} + b^2 S_{XX} = S_{YY} - b^2 S_{XX}$$

Finalmente:

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

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

Exemplo

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

Tem-se:

$$S_{YY} = 1932{,}10$$

$$S_{XX} = 8250$$

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

Então:

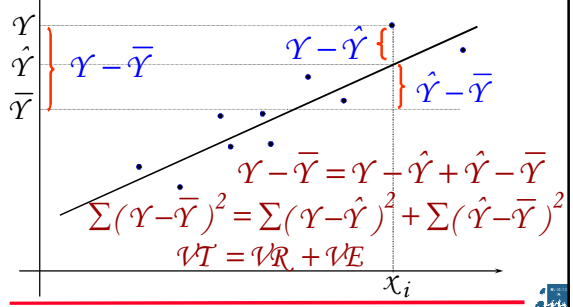
$$s = \sqrt{\frac{S_{YY} - b\, S_{XY}}{n-2}} =$$

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

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

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

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

(c) Variação Explicada: VE

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

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**".

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

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) Teste a existência da regressão;

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

(d) Determine um IC, de 95%, para o valor da pureza, na hipótese do % de hidrocarbono ser 1,20% .