Correlação & Regressão

Prof. Lori 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.

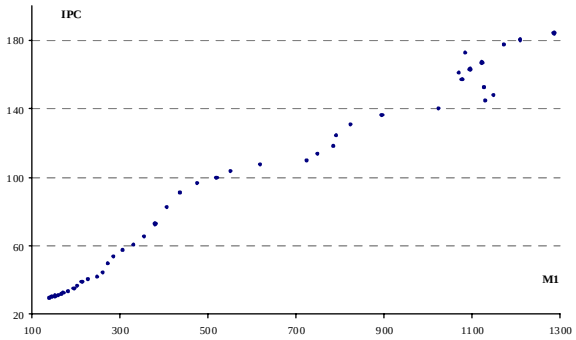
Indicadores de Associação

O Estoque de Moeda (M1) está relacionado com a variação dos preços. Verifique se existe correlação entre o IPC americano com a oferta monetária, considerando dados do período de 1960 a 2003.

Ano	Y = M1	X = IPC
1960	140,7	29,6
1961	145,2	29,9
1962	147,8	30,2
1963	153,3	30,6
1964	160,3	31,5
1965	167,8	32,4
...	...	...
2000	1172,9	177,1
2002	1210,4	179,9
2003	1287,1	184,0

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

Diagrama de Dispersão



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

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

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

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

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

Coeficiente de Correlação

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

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).*

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

*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:*

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

Mas

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) =$$
$$= \sum [X_i Y_i - \bar{X} Y_i - X_i \bar{Y} + \bar{X}\bar{Y}] =$$
$$= \sum X_i Y_i - \sum \bar{X} Y_i - \sum X_i \bar{Y} + \sum \bar{X}\bar{Y} =$$
$$= \sum X_i Y_i - \bar{X} \sum Y_i - \bar{Y} \sum X_i + \sum \bar{X}\bar{Y} =$$
$$= \sum X_i Y_i - n\bar{X}\bar{Y} - n\bar{X}\bar{Y} + n\bar{X}\bar{Y} =$$
$$= \sum X_i Y_i - n\bar{X}\bar{Y}$$

Então:

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

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 $-\infty$ a $+\infty$. Assim é mais conveniente utilizar o **coeficiente de correlação linear de Pearson (momento produto)**.

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

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

Onde:

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

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

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

Esta expressão não é muito prática para calcular 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:

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

$$= \frac{\frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{n-1}}{\sqrt{\frac{\sum X_i^2 - n\bar{X}^2}{n-1}}\sqrt{\frac{\sum Y_i^2 - n\bar{Y}^2}{n-1}}} =$$

$$= \frac{\sum X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\left(\sum X_i^2 - n\bar{X}^2\right)\left(\sum Y_i^2 - n\bar{Y}^2\right)}}$$

Fazendo

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

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

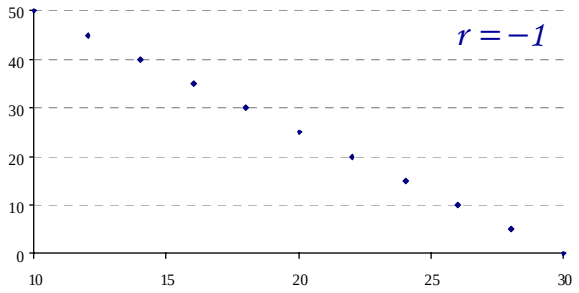
$$S_{YY} = \sum Y_i^2 - n\bar{Y}^2$$

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

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

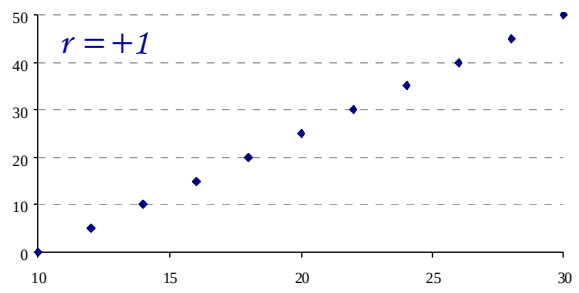
Assim se r = -1, temos uma relacionamento linear negativo perfeito, isto é, os pontos estão todos alinhados e quando X aumenta Y decresce e vice-versa.

Correlação perfeita e negativa



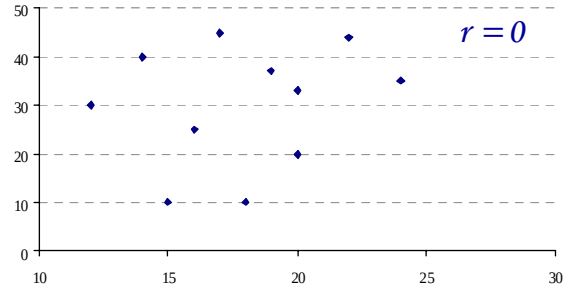
Se r = +1, temos uma relacionamento linear positivo perfeito, isto é, os pontos estão todos alinhados e quando X aumenta Y também aumenta.

Correlação perfeita e positiva



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

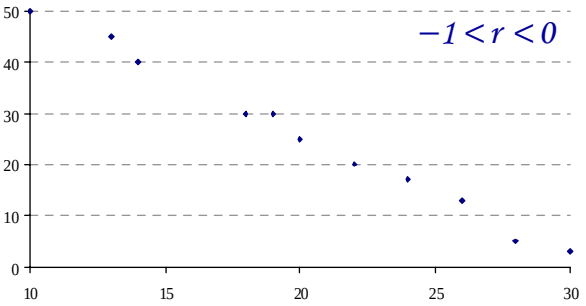
Correlação nula



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.

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

Correlação negativa

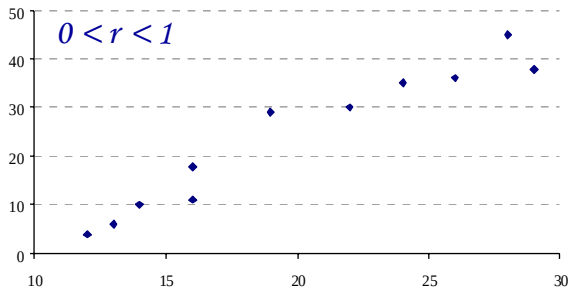


Prof. Lori Viali, Dr. – UFRGS – Instituto de Matemática - Departamento de Estatística

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.

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

Correlação positiva



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

Observação:

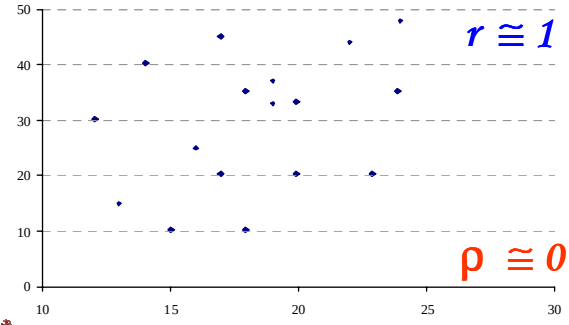
Uma correlação amostral não significa necessariamente uma correlação populacional e vice-versa. É necessário testar o coeficiente de correlação para verificar se a correlação amostral é também populacional.

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

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 = Índice de Preços ao Consumidor versus Y = Estoque de Moeda, para os valores da Economia Americana de 1960 a 2003.

Ano	X	Y	XY	X^2	Y^2
1960	140,7	29,6			
1961	145,2	29,9			
1962	147,8	30,2			
1963	153,3	30,6			
1964	160,3	31,5			
1965	167,8	32,4			
...	...	...			
2000	1172,9	177,1			
2002	1210,4	179,9			
2003	1287,1	184,0			
Total	25894,5	4102,9	3295760,69	21856837,21	503187,97

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

Tem-se:

$n = 44 \quad \sum X = 2589{,}50 \quad \sum Y = 410{,}290$

$\bar{X} = 588{,}5114 \quad \bar{Y} = 93{,}2477 \quad \sum XY = 132957606{,}69$

$\sum X^2 = 2185683{,}721 \quad \sum Y^2 = 50318{,}797$

Então:

$$S_{XY} = \sum X_i Y_i - n\bar{X}\bar{Y} =$$
$$= 881157{,}4161$$

$$S_{XX} = \sum X_i^2 - n\bar{X}^2 =$$
$$= 6617629{,}7043$$

$$S_{YY} = \sum Y_i^2 - n\bar{Y}^2 =$$
$$= 120601{,}8698$$

$$r = \frac{S_{XY}}{\sqrt{S_{XX} \cdot S_{YY}}} =$$
$$= \frac{881157{,}4161}{\sqrt{6617629{,}7043.120601{,}8698}} =$$
$$= 0{,}9863$$

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

Teste para o Coeficiente de Correlação

*O valor de "**r**" é obtido com base em uma amostra. Ele é portanto, uma estimativa do verdadeiro valor da correlação populacional (ρ).*

A teoria dos testes de hipóteses pode ser utilizada para verificar se com base na estimativa "r" é possível concluir se existe ou não correlação populacional, isto é, desejamos testar:

$\mathcal{H}_0: \rho = 0$

$\mathcal{H}_1: \rho > 0$

(teste unilateral/unicaudal à direita)

$\rho < 0$

(teste unilateral/unicaudal à esquerda)

$\rho \neq 0$

(teste bilateral/bicaudal).

O teste para a existência de correlação linear entre duas variáveis é realizado por:

$$t_{n-2} = \frac{r - \mu_r}{\hat{\sigma}_r} = \frac{r - 0}{\sqrt{\frac{1 - r^2}{n - 2}}} =$$

$$= r\sqrt{\frac{n-2}{1-r^2}}$$

Rejeita-se a Hipótese nula se:

$t_{n-2} > t_c$

(teste unilateral/unicaudal à direita)

$t_{n-2} < t_c$

(teste unilateral/unicaudal à esquerda)

$|t_{n-2}| > t_c$

(teste bilateral/bicaudal).

Onde t_c é tal que:

$P(t < t_c) = 1 - \alpha$

(teste unilateral/unicaudal à direita)

$P(t < t_c) = \alpha$

(teste unilateral/unicaudal à esquerda)

$P(t < t_c) = \alpha/2$ *ou* $P(t > t_c) = \alpha/2$

(teste bilateral/bicaudal).

Exemplo

Suponha que uma amostra de ***n = 12****, alunos forneceu um coeficiente de correlação amostral de* ***r = 0,66****, entre X = "nota em cálculo" e Y = "nota em Econometria". Verifique se é possível afirmar que uma nota boa em Cálculo está relacionada com uma nota boa em Econometria a* ***1%*** *de significância.*

Solução:

Hipóteses:

$\mathcal{H}_0$: $\rho = 0$

$\mathcal{H}_1$: $\rho > 0$

Dados:

$n = 12$

$r = 0,66$

$\alpha = 1\%$

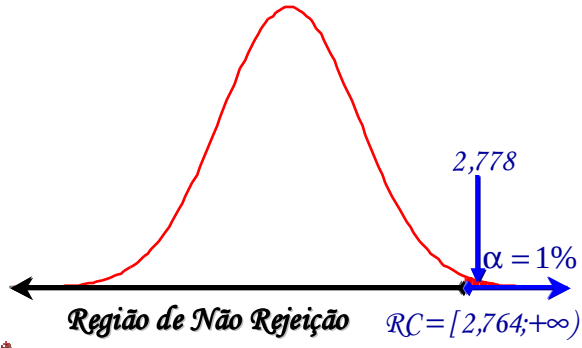
Trata-se de um teste unilateral à direita para o coeficiente de correlação.

A variável teste é:

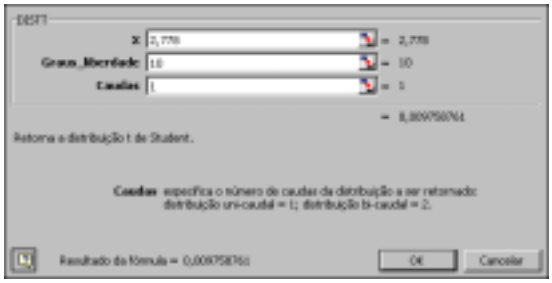
$$t_{n-2} = r\sqrt{\frac{n-2}{1-r^2}}$$

Então:

$$t_{10} = r\sqrt{\frac{n-2}{1-r^2}} = 0,66\sqrt{\frac{12-2}{1-066^2}} = 2,778$$



A significância do resultado obtido (2,778), isto é, o valor-p é dada por $P(T_{10} > 2,778)$. Utilizando o Excel, tem-se:



Como a significância do resultado (0,98%) é menor que a significância do teste (1%) é possível rejeitar a hipótese nula.

A transformada de Fisher

O procedimento realizado para testar o coeficiente de correlação só é válido para testar a hipótese nula de que ***não*** *existe correlação, isto é,* $\rho = 0$. *Outros tipos de testes só podem ser realizados através da transformada "zeta" de Fisher.*

A transformada "ζ*" é dada por:*

$$\zeta = \frac{1}{2} ln\left(\frac{1+r}{1-r}\right)$$

*O que equivale a considerar "****r****" como a tangente hiperbólica de "*ζ*"*

A vantagem desta transformação é que os valores de "ζ*" estão distribuídos aproximadamente de acordo com uma normal de média:* $\mu_\zeta = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right)$

E desvio: $\sigma_\zeta = \sqrt{\frac{1}{n-3}}$

Esta transformação permite, realizar, testes de hipóteses e construir intervalos de confiança para o coeficiente de correlação, através de ζ *e da* ***distribuição normal****.*

$\mathcal{H}_0$: $\rho = \rho_0$

$\mathcal{H}_1$: $\rho > \rho_0$

(teste unilateral/unicaudal à direita)

$\rho < \rho_0$

(teste unilateral/unicaudal à esquerda)

$\rho \neq \rho_0$

(teste bilateral/bicaudal).

O teste para a existência de correlação linear populacional entre duas variáveis X e Y é realizado por:

$$z = \frac{\zeta - \mu_\zeta}{\sigma\zeta} = \frac{\zeta - \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right)}{\sqrt{\frac{1}{n-3}}}$$

Rejeita-se a Hipótese nula se:

$z > z_c$

(teste unilateral/unicaudal à direita)

$z < z_c$

(teste unilateral/unicaudal à esquerda)

$|z| > z_c$

(teste bilateral/bicaudal).

Onde z_c é tal que:

$\Phi(z_c) = 1 - \alpha$

(teste unilateral/unicaudal à direita)

$\Phi(z_c) = \alpha$

(teste unilateral/unicaudal à esquerda)

$\Phi(z_c) = \alpha/2$ *ou* $\Phi(z_c) = 1 - \alpha/2$

(teste bilateral/bicaudal).

Exemplo

Suponha que uma amostra de ***n = 35****, alunos forneceu um coeficiente de correlação amostral de* ***r = 0,75****, entre X = "número de horas de estudo" e Y = "nota em Econometria". Verifique se é possível afirmar que o "o número de horas de estudo" apresenta uma correlação de pelo menos 0,5 na população com a "Econometria", a* ***1%*** *de significância.*

Solução:

Hipóteses:

$\mathcal{H}_0$: $\rho = 0,5$

$\mathcal{H}_1$: $\rho > 0,5$

Dados:

$n = 35$

$r = 0,75$

$\alpha = 1\%$

Trata-se de um teste unilateral à direita para o coeficiente de correlação.

A variável teste é:

$$z = \frac{\zeta - \mu_\zeta}{\sigma\zeta} = \frac{\zeta - \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right)}{\sqrt{\frac{1}{n-3}}}$$

Então:

$$\zeta = \frac{1}{2} ln\left(\frac{1+0,75}{1-0,75}\right) = 0,9730$$

A média vale:

$$\mu_\zeta = \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right) = \frac{1}{2} ln\left(\frac{1+0,5}{1-0,5}\right) = 0,5493$$

E o desvio padrão vale:

$$\sigma\zeta = \sqrt{\frac{1}{n-3}} = \sqrt{\frac{1}{35-3}} = \sqrt{\frac{1}{32}} = 0,1768$$

Padronizando, tem-se:

$$z = \frac{\zeta - \mu_\zeta}{\sigma\zeta} = \frac{\zeta - \frac{1}{2} ln\left(\frac{1+\rho}{1-\rho}\right)}{\sqrt{\frac{1}{n-3}}} =$$

$$= \frac{0,9730 - 0,5493}{0,1768} = 2,40$$

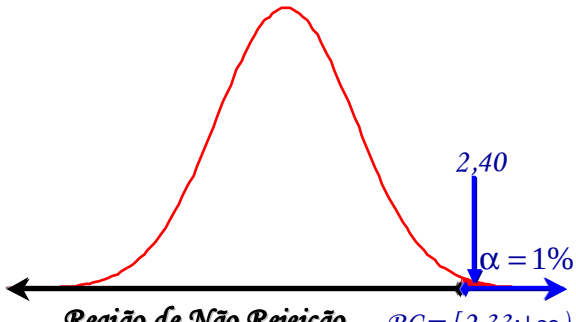
O valor crítico z_c é tal que:

$P(Z > z_c) = \alpha = 1\%.$

Ou $\Phi(z_c) = 99\%.$

Então $z_c = 2,33.$

Assim $RC = [2,33; \infty)$



A significância do resultado obtido (2,40), isto é, o valor-p. Para isto, deve-se calcular $P(Z > 2,40)$, isto é, $\Phi(-2,40) = 0,82\%$.

Como $p = 0,82\% < \alpha = 1\%$. Rejeito H_0.

Regressão Linear Simples

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**"*

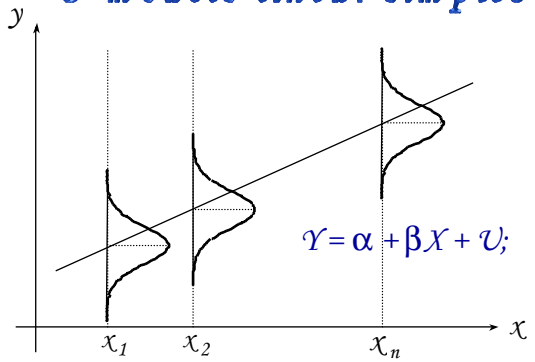
*Quando existir apenas uma variável regressora (X) tem-se a **regressão simples**, se Y depender de duas ou mais variáveis regressoras, então tem-se a "**regressão múltipla**".*

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** *(Métodos dos Mínimos Quadrados), consiste em minimizar a "***soma dos quadrados das distâncias da reta aos pontos***".*

Tem-se:

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

Então:

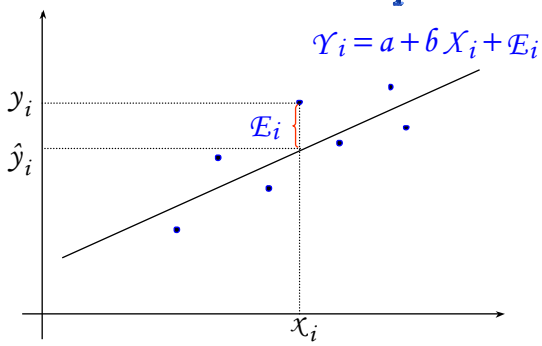
$$E_i = Y_i - (a + bx_i)$$

Deve-se minimizar:

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

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

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:

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

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

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

Exemplo

Considerando os valores das variáveis "Oferta Monetária" e "Índice de Preços ao Consumidor", consideradas anteriormente, determinar uma equação de regressão linear para prever o IPC dado um determinado nível de Oferta Monetária.

Ano	Y = IPC	X = M1
1960	29,6	140,7
1961	29,9	145,2
1962	30,2	147,8
1963	30,6	153,3
1964	31,5	160,3
1965	32,4	167,8
...	...	...
2000	177,1	1172,9
2002	179,9	1210,4
2003	184,0	1287,1

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.

Ano	X	Y	XY	X^2	Y^2
1960	140,7	29,6			
1961	145,2	29,9			
1962	147,8	30,2			
1963	153,3	30,6			
1964	160,3	31,5			
1965	167,8	32,4			
...	...	...			
2000	1172,9	177,1			
2002	1210,4	179,9			
2003	1287,1	184,0			
Total	25894,5	4102,9	3295760,69	21856837,21	503187,97

Tem-se:

$n = 44 \quad \sum X = 2589{,}50 \quad \sum Y = 4102{,}90$

$\bar{X} = 588{,}5114 \quad \bar{Y} = 93{,}2477 \quad \sum XY = 1329576{,}69$

$\sum X^2 = 2185683{,}21 \quad \sum Y^2 = 503187{,}97$

Então:

$$S_{XY} = \sum X_i Y_i - n\bar{X}\bar{Y} =$$
$$= 881157{,}4161$$

$$S_{XX} = \sum X_i^2 - n\bar{X}^2 =$$
$$= 6617629{,}7043$$

$$S_{YY} = \sum Y_i^2 - n\bar{Y}^2 =$$
$$= 120601{,}8698$$

A equação de regressão, será, então:

$$b = \frac{S_{XY}}{S_{XX}} = \frac{881157{,}4161}{6617629{,}7043} = 0{,}1332 \cong 0{,}13$$

$$a = \bar{Y} - b\bar{X} = 93{,}2477 - 0{,}1332 \cdot 588{,}5114 =$$
$$= 14{,}8857 \cong 14{,}89$$

$$\hat{Y} = 14{,}89 + 0{,}13X$$

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:

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

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

Então:

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

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

Assim:

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

Será, finalmente:

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

Exemplo

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

Tem-se: $S_{YY} = 120601,8698$

$S_{XX} = 6617629,7043$

$$b = \frac{S_{XY}}{S_{XX}} = \frac{881157,4161}{6617629,7043} = 0,1332$$

Então:

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

$$= \sqrt{\frac{120601,8698 - 0,1332.881157,4161}{44-2}} =$$

$$= 8,8278 \cong 8,83$$

A pergunta, agora, é: este erro é razoável?, quer dizer, ele não é muito grande?

A resposta envolve o cálculo do erro relativo, isto é, devemos comparar este resultado com a variável de interesse.

A variável envolvida aqui é a Y, isto é, a base monetária, então, o erro relativo, será:

$$g_s = \frac{s}{\overline{Y}} = \frac{8,8278}{93,2477} = 9,47\%$$

Distribuições das Estimativas

*Os valores de "**a**" e "**b**" são estimadores de "α" e "β". As propriedades estatísticas destes estimadores são úteis para testar a adequação do modelo. Eles são variáveis aleatórias uma vez que são combinações lineares dos Y_i que são, por sua vez, variáveis aleatórias.*

As principais propriedades de interesse são a média (expectância), a variabilidade (erro padrão) e a distribuição de probabilidade de cada um dos estimadores.

Comportamento de "a"

(i) Expectância

$$E(a) = E\left(\bar{Y} - b\bar{X}\right) = \ldots = \alpha$$

(ii) Variância

$$V(a) = V\left(\bar{Y} - b\bar{X}\right) = \ldots = \sigma^2\left(\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}\right)$$

Portanto a distribuição da estatística "a", será:

$$a \sim N\left(\alpha,\ \sigma\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}\right)$$

Como o valor "σ" não é conhecido e precisa ser estimado por "s", então, de fato, utiliza-se a distribuição t_{n-2}.

Comportamento de "b"

(i) Expectância

$$E(b) = E\left(\frac{S_{XY}}{S_{XX}}\right) = \ldots = \beta$$

(ii) Variância

$$V(b) = V\left(\frac{S_{XY}}{S_{XX}}\right) = \ldots = \frac{\sigma^2}{S_{XX}}$$

Portanto a distribuição da estatística "b", será:

$$b \sim N\left(\beta,\ \frac{\sigma}{\sqrt{S_{XX}}}\right)$$

Como o valor "σ" não é conhecido e precisa ser estimado por "s", então, de fato, utiliza-se a distribuição t_{n-2}.

Covariância entre "a" e "b"

Por definição:

$$Cov(a, b) = E(ab) - E(a).E(b) = E(ab) - \alpha\beta.$$

Mas

$$E(ab) = E[(\bar{Y} - b\bar{X}).b] = E(\bar{Y}b) - E(\bar{X}b^2) =$$
$$= \bar{Y}E(b) - \bar{X}E(b^2) = \beta\bar{Y} - \bar{X}(\beta^2 - \sigma_b^2) =$$
$$= \beta(\bar{Y} - \beta\bar{X}) - \bar{X}\sigma_b^2 = \alpha\beta - \bar{X}\sigma_b^2$$

Então:

$$Cov(a,b) = E(ab) - \alpha\beta =$$
$$= \alpha\beta - \bar{X}\sigma_b^2 - \alpha\beta = -\bar{X}\sigma_b^2$$

Assim:

$$Cov(a,b) = -\bar{X}V(b) = -\bar{X}V\left(\frac{S_{XY}}{S_{XX}}\right) =$$
$$= \frac{-\bar{X}\sigma^2}{S_{XX}}$$

Intervalos de Confiança para os parâmetros da regressão

Da mesma forma que foram obtidos IC para a média, a proporção e a variância de uma população, pode-se determinar intervalos para os parâmetros "α" e "β" da regressão.

IC para o coeficiente linear "α"

O IC de "1 – α" de confiança para o coeficiente linear "α" é dado por:

$$P\left(a - t_{n-2}S\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}} \le \alpha \le a + t_{n-2}S\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}\right) =$$
$$= 1 - \alpha$$

IC para o coeficiente angular "β"

O IC de "1 – α" de confiança para o coeficiente da regressão "β" é dado por:

$$P\left(b - t_{n-2}\frac{S}{\sqrt{S_{XX}}} \le \beta \le b + t_{n-2}\frac{S}{\sqrt{S_{XX}}}\right) = 1 - \alpha$$

Exemplo

Determinar intervalos de confiança de 95% para os parâmetros da equação de regressão, utilizando os dados do exercício anterior.

$$\hat{Y} = 14{,}89 + 0{,}13x$$

Dados

$S_{YY} = 120601{,}8698$ $\quad a = 14{,}8857$

$S_{XX} = 6617629{,}7043$ $\quad b = 0{,}1332$

$S_{XY} = 881157{,}4161$ $\quad s = 8{,}8278$

$\bar{X} = 588{,}5114$ $\quad n = 44$

$\bar{Y} = 93{,}2477$ $\quad 1 - \alpha = 95\%$

O IC de "1- α" para o Coef. Linear "α" é dado por:

$$a \pm t_{n-2} S \sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}$$

Então:

$$14{,}8857 \pm 2{,}0181.8{,}8278 \sqrt{\frac{1}{44} + \frac{588{,}5114^2}{6617629{,}7043}}$$

$14{,}8857 \pm 4{,}9161$

$[9{,}97;\ 19{,}80]$

O IC de "1- α" para o Coef. Angular "β" é dado por:

$$b \pm t_{n-2} \frac{S}{\sqrt{S_{XX}}}$$

Então:

$$0{,}1332 \pm 2{,}0181. \frac{8{,}8278}{\sqrt{6617629{,}7043}}$$

$0{,}1332 \pm 0{,}0069$

$[0{,}1262;\ 0{,}1401]$

$[0{,}13;\ 0{,}14]$

Intervalos de Confiança para o valor médio e para um valor individual de Y

Da mesma forma que foram obtidos IC para os parâmetros da regressão, pode-se obter IC para os valores estimados de Y para um dado x. Vamos considerar dois casos:

(a) Considerando somente a incerteza da linha de regressão;

(b) Considerando a incerteza da linha mais a variação da variável Y.

IC para um valor médio de Y, dado x

Para construir o IC de "$1-\alpha$" para o valor médio de Y, dado x, é necessário conhecer sua distribuição. Tem-se:

$$\hat{Y} \sim \mathcal{N}(\alpha+\beta x;\ \sigma\sqrt{\frac{1}{n}+\frac{(X-\bar{X})^2}{S_{XX}}})$$

Então IC de "$1-\alpha$" de confiança para o um valor médio de Y, dado x, é:

$$\hat{Y} \pm t_{n-2} S\sqrt{\frac{1}{n}+\frac{(X-\bar{X})^2}{S_{XX}}}$$

IC para um valor médio individual de Y, dado x

Uma estimativa do valor individual de Y é dado por "$a + bx$" e a distribuição desta estimativa será dada por:

$$\hat{Y} \sim \mathcal{N}(0;\ \sigma\sqrt{1+\frac{1}{n}+\frac{(X-\bar{X})^2}{S_{XX}}})$$

Então IC de "$1-\alpha$" de confiança para o um valor individual de Y, dado x, será:

$$\hat{Y} \pm t_{n-2} S\sqrt{1+\frac{1}{n}+\frac{(X-\bar{X})^2}{S_{XX}}}$$

Exemplo

Determinar intervalos de confiança de 95% para os valores médio e individual de Y, na hipótese de $x = 200$.

Dados

$S_{YY} = 1932,10$

$S_{XX} = 8250$

$S_{XY} = 3985$

$\overline{X} = 145$

$\overline{Y} = 67,30$

$x = 200$

$a = -2,7394$

$b = 0,4830$

$s = 0,9503$

$n = 10$

$1 - \alpha = 95\%$

O IC de "1- α" para o valor médio de Y, dado "x" é:

$$\hat{Y} \pm t_{n-2} S \sqrt{\frac{1}{n} + \frac{(X-\overline{X})^2}{S_{XX}}}$$

Então:

$$\hat{y} = -2,7394 + 0,4830.200 = 93,8606$$

$$93,8606 \pm 2,306.0,9503 \sqrt{\frac{1}{10} + \frac{(200-145)^2}{8250}}$$

$$93,8606 \pm 1,4970$$

[92,36; 95,36]

O IC de "1- α" para o valor individual de Y, dado "x" é:

$$\hat{Y} \pm t_{n-2} S \sqrt{1 + \frac{1}{n} + \frac{(X-\overline{X})^2}{S_{XX}}}$$

Então:

$$93,8606 \pm 2,306.0,9503 \sqrt{1 + \frac{1}{10} + \frac{(200-145)^2}{8250}}$$

$$93,8606 \pm 2,6539$$

[91,21; 96,51]

Testes de Hipóteses para os parâmetros da regressão

Da mesma forma que foram testados todos os parâmetros até então pode-se testar os parâmetros "α" e "β" da regressão.

Teste para o coeficiente linear "α"

A variável teste para testar o coeficiente linear é dado por:

$$t_{n-2} = \frac{a - \alpha}{S \sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}}}$$

Teste para o coeficiente angular "β"

A variável teste para testar o coeficiente da regressão "β" é dada por:

$$t_{n-2} = \frac{b - \beta}{\frac{S}{\sqrt{S_{XX}}}}$$

Exemplo

(a) Testar, a 1% de significância, se é possível afirmar que a linha de regressão, do exemplo dado, não passa pela origem.

(b) Testar se é possível, a 1% de significância, afirmar que existe regressão positiva entre as duas variáveis.

Dados

$a = -2,7394$

$b = 0,4830$

$s = 0,9503$

$n = 10$

$1 - \alpha = 1\%$

$S_{YY} = 1932,10$

$S_{XX} = 8250$

$S_{XY} = 3985$

Solução:

Hipóteses:

H_0: $\alpha = 0$

H_1: $\alpha \neq 0$

(A)

Dados:

$n = 10$

$a = -2,739$

$\alpha = 1\%$

Trata-se de um teste bilateral para o coeficiente linear da regressão.

A variável teste é:

$$t_{n-2} = \frac{a - \alpha}{S\sqrt{\frac{1}{n} + \frac{\bar{X}^2}{S_{XX}}}}$$

Então:

$$t_8 = \frac{-2,739 - 0}{0,9503\sqrt{\frac{1}{10} + \frac{145^2}{8250}}} = -1,771$$

O valor crítico t_c é tal que: $P(|T| > t_c) = \alpha$
Então $t_c = -3{,}355$. Assim $RC = [-3{,}355; \infty)$

DECISÃO e CONCLUSÃO:

Como $t_8 = -1{,}771 \in RC$ ou $-1{,}771 > -3{,}355$. Aceito H_0, isto é, a 1% de significância, **não** se pode afirmar que a linha de regressão não passe pela origem.

Solução:

Hipóteses:

$H_0: \beta = 0$

$H_1: \beta > 0$

(B)

Dados:

$n = 10$

$b = 0{,}4830$

$\alpha = 1\%$

Trata-se de um teste unilateral para o coeficiente angular da regressão.

A variável teste é:

$$t_{n-2} = \frac{b - \beta}{\frac{S}{\sqrt{S_{XX}}}}$$

Então:

$$t_8 = \frac{0{,}4830 - 0}{0{,}9503 / \sqrt{8250}} = 46{,}165$$

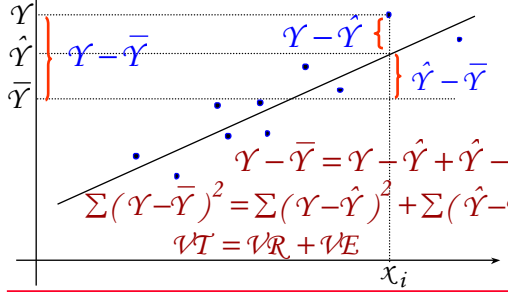
O valor crítico t_c é tal que: $P(T > t_c) = \alpha$
Então $t_c = 2{,}896$. Assim $RC = [2{,}896; \infty)$

DECISÃO e CONCLUSÃO:

Como $t_8 = 46{,}165 \in RC$ ou $46{,}165 > 2{,}896$. Rejeito H_0, isto é, a 1% de significância, pode-se afirmar que existe regressão entre as duas variáveis.

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.