Estatística

Definição

A Estatística é uma ciência bastante abrangente que pode ser aplicada a diversas áreas de atividade, tais como em estudos demográficos, na indústria para melhorar a qualidade dos processos, na área de recursos humanos para analisar o comportamento e perfil dos funcionários de uma empresa, na saúde para a descoberta de novos medicamentos, nas pesquisas de mercado para prever o comportamento do consumidor e de opinião para prever a possível eleição de determinado candidato, na área financeira para prever riscos na oferta de crédito, em marketing para identificar consumidores propensos a responder determinada comunicação, entre outras. A Estatística tem por objetivo fornecer métodos e técnicas para que se possa, racionalmente, lidar com situações de incerteza oferecendo resultados precisos para tomadas de decisões mais confiáveis. Podemos dizer que entre tantos outros motivos, é necessário o estudo da Estatística para:

- Apresentar corretamente as informações;
- Extrair conclusões sobre as populações baseadas na informação da amostra;
- Saber melhorar os processos de sua empresa;
- Saber obter previsões de confiança e melhor planejar o futuro de sua empresa;

A palavra *estatística* surge da expressão em Latim *statisticum collegium* palestra sobre os assuntos do Estado, de onde surgiu a palavra em língua italiana *statista*, que significa "homem de estado", ou político, e a palavra alemã *Statistik*, designando a análise de dados sobre o Estado. A palavra adquiriu um significado de coleta e classificação de dados, no início do século 19.

Nós descrevemos o nosso conhecimento (e ignorância) de forma matemática e tentamos aprender mais sobre aquilo que podemos observar. Isto requer:

- O planejamento das observações por forma a controlar a sua variabilidade (concepção do experimento)
- Sumarização da coleção de observações
- Inferência estatística - obter um consenso sobre o que as observações nos dizem sobre o mundo que observamos

Em algumas formas de estatística descritiva, nomeadamente em mineração de dados (*data mining*), os segundo e terceiro passos tornam-se normalmente mais importantes que o primeiro.

A probabilidade de um evento é freqüentemente definida como um número entre zero e um. Na realidade, porém, nunca há situações que tenham probabilidades 0 ou 1. Você pode dizer que o sol irá certamente nascer na manhã, mas e se acontecer um evento extremamente difícil de ocorrer que o destrua? E se ocorrer uma guerra nuclear e o céu ficar coberto de cinzas e fumaças?

Normalmente aproximamos a probabilidade de alguma coisa para cima ou para baixo porque elas são tão prováveis ou improváveis de ocorrer, que é fácil de reconhecê-las como probabilidade de um ou zero.

Entretanto, isto normalmente leva a desentendimentos e comportamentos perigosos, porque as pessoas não conseguem distinguir entre, e.g., uma probabilidade de 10^{-4} e uma probabilidade de 10^{-9}, a despeito da grande diferença prática entre elas. Se você espera cruzar a estrada cerca de 10^{5} ou 10^{6} vezes na sua vida, então reduzindo a risco de cruzar a estrada em 10^{-9} irá fazer você seguro pelo resto da sua vida, enquanto um risco de cruzá-la em 10^{-4} irá fazer ser bem provável que você tenha um acidente, mesmo com o sentimento intuitivo que 0,01% é um risco muito baixo.

Algumas ciências usam a estatística aplicada tão extensivamente que elas tem uma terminologia especializada. Estas disciplinas incluem:

- Bioestatística
- Estatística Comercial
- Estatística Econômica
- Estatística Engenharia

- Estatística Física
- Estatística Populacional
- Estatística Psicológica
- Estatística Social (para todas as ciências *sociais*)
- Análise de Processo e Quimiometria (para análise de dados da química analítica e da engenharia química)

Estatística forma uma ferramenta chave nos negócios e na industrialização como um todo. É utilizada a fim de entender sistemas variáveis, controle de processos (chamado de "controle estatístico de processo" ou CEP), para sumarização de dados, e para tomada de decisão baseada em dados. Em nessas funções ela é uma ferramenta chave, e é a única ferramenta segura.

Ligações para estatística observacional fenômeno são coletados pelos Fenômenos Estatísticos

- Estatística Inferencial é o conjunto de técnicas utilizadas para identificar relações entre variáveis que representem ou não relações de causa e efeito.

- Estatística Robusta é o conjunto de técnicas utilizadas para atenuar o efeito de outliers e preservar a forma de uma distribuição tão aderente quanto possível aos dados empíricos.

A base da estatística e sua definição

A Estatística é uma ferramenta matemática que nos informa sobre o quanto de erro nossas observações apresenta sobre a realidade pesquisada. A estatística baseia-se na medição do erro que existe entre a estimativa de quanto uma amostra representa adequadamente a população da qual foi extraída. Assim o conhecimento de teoria de conjuntos, análise combinatória e cálculo são indispensáveis para compreender como o erro se comporta e a magnitude do mesmo. É o erro (erro amostral) que define a qualidade da observação e do delineamento experimental. A faceta dessa ferramenta mais palpável é a Estatística Descritiva. A descrição dos dados coletados é comumente apresentado em gráficos ou relatórios e serve tanto a prospecção de uma ou mais variáveis para posterior aplicação ou não de testes estatísticos bem como a apresentação de resultados de delineamentos experimentais.

Alguns órgãos estatísticos nacionais

- Alemanha: Statistisches Bundesamt
- Austrália: Australian Bureau of Statistics
- Bélgica: Statistics Belgium
- Brasil: Fundação Instituto Brasileiro de Geografia e Estatística (IBGE)
- Canadá: Statistics Canada, Statistique Canada
- Colômbia: Departamento Administrativo Nacional de Estadistica (DANE)
- Espanha: Instituto Nacional de Estadística (INE)
- Estados Unidos: FedStats
- França: Institut National de la Statistique et des Études Économiques (INSEE)
- Grécia: National Statistical Service of Greece
- Holanda: Centraal Bureau Statistiek
- Índia: Indian Statistical Institute
- Itália: Istituto Nazionale di Statistica (ISTAT)
- Irlanda: Central Statistics Office of Ireland
- Nova Zelândia: Statistics New Zealand
- Portugal: Instituto Nacional de Estatistica
- Reino Unido: Office for National Statistics (ONS)
- Suíça: Swiss Federal Statistical Office

Regressão Linear Simples

Regressão linear simples => ***Método de análise da relação entre uma variável independente e uma variável dependente.***

r^2 – ***Medida de como a variável independente em uma análise de regressão linear simples pode explicar variações na variável dependente, seu valor situa-se em 0 (fraco ajuste) e 1(ajuste perfeito). Usa-se o símbolo*** **r^2 porque é o quadrado do coeficiente de correlação amostral entre duas variáveis.**

Reta de Regressão => ***Reta calculada na análise de regressão, usada para estimar a relação entre duas grandezas (a variável independente e a variável dependente)***

De maneira geral, estaremos diante de um modelo de regressão linear simples quando a relação linear entre duas variáveis , X e Y, pode ser satisfatoriamente definida pela seguinte equação matemática:

Y = A+ B ×X + U;

Cujos parâmetros são:

Y= Variável dependente;

X= Variável independente;

A= Coeficiente linear ou intercepto da reta;

B = Coeficiente angular ou declividade da reta;

U= Erro aleatório na população

Também podemos ver a equação acima da seguinte forma:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$i = 1,2,\ldots, \quad n$$

Onde:

Y_i é o i-ésimo valor da variável resposta;

β_0 e β_1 são os parâmetros (coeficientes de regressão);

X_i é o i-ésimo valor da variável preditora (é uma constante conhecida, fixo).

ε_i é o termo do erro aleatório com $E(\varepsilon_i)=0$ e $\sigma^2(\varepsilon_i)=\sigma^2$;

ε_i e ε_j não são correlacionados $\Rightarrow \sigma(\varepsilon_i, \varepsilon_j)=0$ para todo i,j; $i \neq j$; (covariância é nula).

i=1,2,...,n.

Os dados são usados para estimar β_0 e β_1, isto é, ajustar o modelo aos dados, para:

• quantificar a relação entre Y e X;

• usar a relação para predizer uma nova resposta Y_0 para um dado valor de X_0 (não incluído no estudo);

• calibração – ou capacidade de predição de novas observações, pode ser feita usando uma nova amostra e comparando os valores estimados com os observados.

- dado um valor de Y_0, para o qual o correspondente valor de X_0 é desconhecido, estimar o valor de X_0.

Características do modelo:

$$1.\ \mathrm{Y_i}\ \text{é uma v.a.}(\mathrm{Y_i} = \overbrace{\beta_0 + \beta_1 X_i}^{\text{constante}} + \overbrace{\varepsilon_i}^{\text{aleatório}})$$

$$2.\ \mathrm{E(Y_i)} = E(\beta_0 + \beta_1 X_i + \varepsilon_i) = \mu_i = \beta_0 + \beta_1 X_i$$

$$3.\ \sigma^2(Y_i) = \sigma^2(\beta_0 + \beta_1 X_i + \varepsilon_i) = \sigma^2(\varepsilon_i) = \sigma^2\ (\text{variância constante})$$

$$4.\ \mathrm{Y_i}\ \text{e}\ \mathrm{Y_j}\ \text{não são correlacionados}$$

O modelo de regressão (2) mostra que as respostas Y_i são oriundas de uma distribuição de probabilidades com média $E(Y_i) = \beta_0 + \beta_1 X_i$ e cujas variâncias são σ^2, a mesma para todos os valores de X. Além disso, quaisquer duas respostas Y_i *e* Y_j não são correlacionadas.

$\hat{Y} = a + b \cdot x$

Onde :

$\hat{Y}$ = Estimativa da variável dependente y;

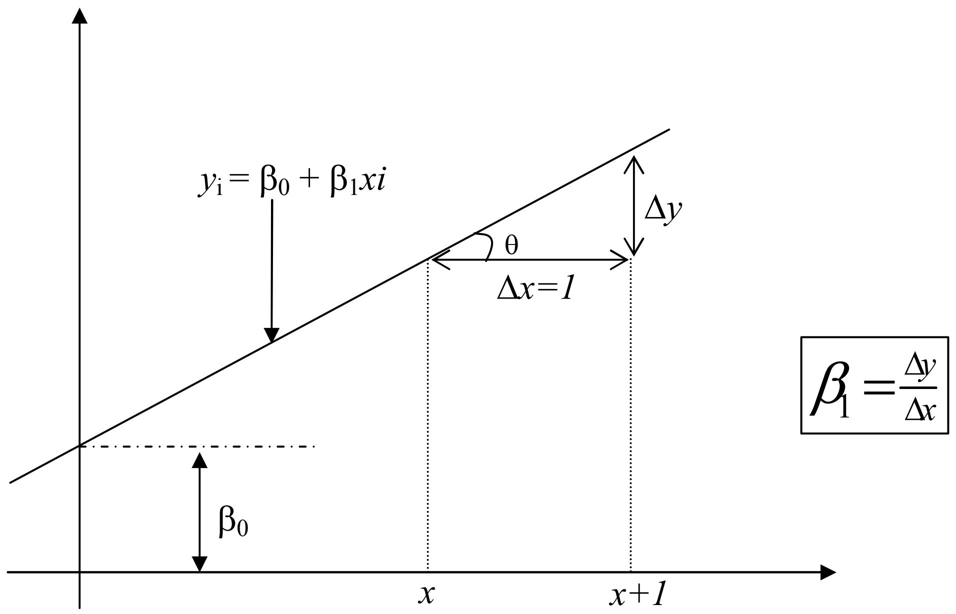
a= Estimativa do coeficiente linear A;

b= Estimativa do coeficiente angular B;

x= Valores amostrais da variável explicativa X.

Obs: O símbolo ^, é utilizado para diferenciar os valores estimados dos amostrais ou observados.

Significado dos parâmetros do modelo de regressão linear simples



β_0 (intercepto); quando a região experimental inclui *X=0,* β_0 é o valor da média da distribuição de Y em X=0, cc, não tem significado prático como um termo separado (isolado) no modelo; β_1 (inclinação) expressa a *taxa de mudança* em *Y,* isto é, é a mudança em *Y* quando ocorre a mudança de uma unidade em *X.* Ele indica a mudança na média da distribuição de probabilidade de *Y* por unidade de acréscimo em *X.*

MÉTODO DOS MÍNIMOS QUADRADOS

.

MODELO MATEMÁTICO, FUNCIONAL E ESTOCÁSTICO

Modelo matemático

Chama-se modelo matemático as sistema teórico ou conceito abstracto através do qual se descreve uma situação física ou um conjunto de acontecimentos. Esta descrição não é necessariamente completa ou exaustiva. Como um modelo serve um propósito específico, a sua formação pode variar largamente de um ponto de vista para outro. Deste modo, o mesmo sistema físico pode ser descrito por mais do que um modelo. Um modelo matemático pode ser dividido em duas partes conceituais: o modelo funcional e o modelo estocástico:

Modelo funcional

O modelo funcional é composto por relações que descrevem a geometria ou características físicas do problema em questão.

Exemplo 1: Para a determinação da área de um terreno rectangular com lados a e b o

modelo funcional é A = ab.

Exemplo 2: Para a determinação da forma de um triângulo com ângulos α, β e γ o modelo funcional será α + β + γ = 180°.

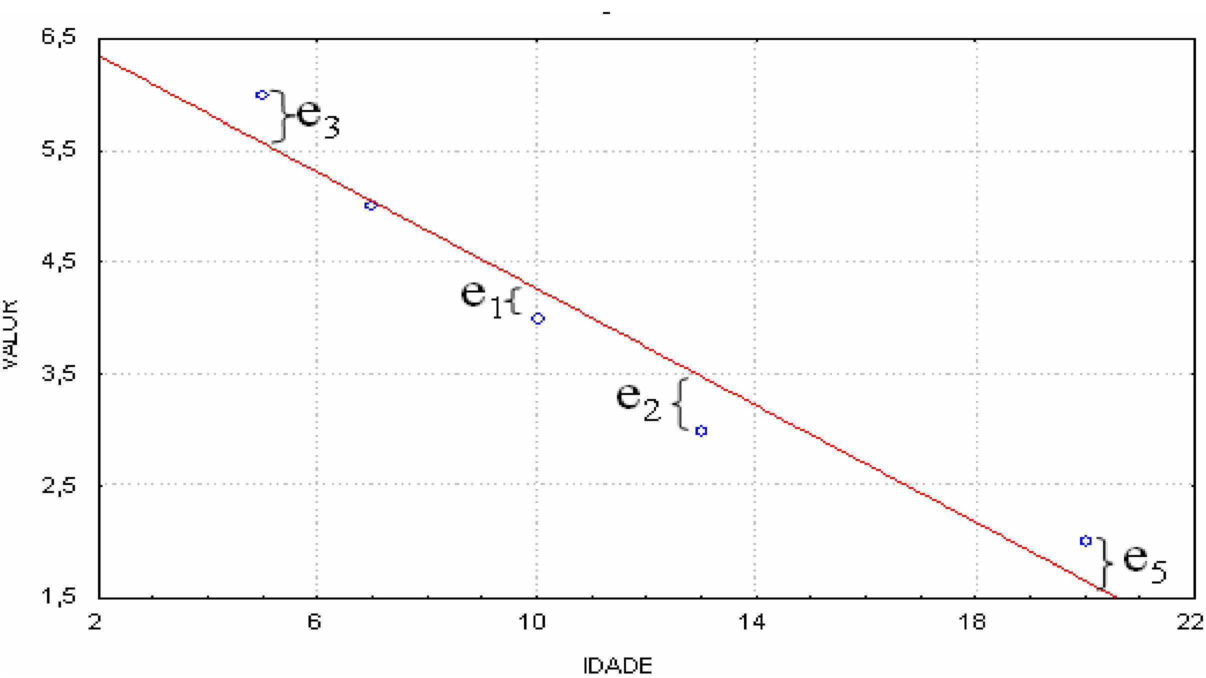
Modelo estocástico

O modelo estocástico é composto pelo conjunto de relações que descrevem as propriedades estatísticas dos elementos envolvidos no modelo funcional. O modelo estocástico indica por exemplo a qualidade das observações feitas (as suas precisões, relativas ou absolutas), indica se as observações estão correlacionadas ou não, indica ainda as variáveis que são consideradas constantes durante o ajustamento e as que se pretendem determinar.

Elevando-se ao quadrado esses desvios e aplicando-se o somatório, temos o critério Q

$$Q = \sum_{i=1}^{n} \left(Y_i - \beta_0 - \beta_1 X_i\right)^2 \qquad (10)$$

De acordo com o método de mínimos quadrados, os estimadores de β0 e β1 são os valores b0 e b1, respectivamente, que minimizam o critério Q para a amostra (X1,Y1),..,(Xn,Yn).



Estimadores de mínimos quadrados

Os valores de β0 e β1 que minimizam o critério Q podem ser obtidos diferenciando-se (10) em relação a β0 e β1 , portanto, obtemos:

$$\frac{\partial Q}{\partial \beta_0} = -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial Q}{\partial \beta_1} = -2\sum_{i=1}^{n}X_i(Y_i - \beta_0 - \beta_1 X_i)$$

Fazendo-se as derivadas parciais de segunda ordem, indicará que um mínimo foi encontrado com os estimadores b0 e b1.

$$-2\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$-2\sum_{i=1}^{n}X_i(Y_i - b_0 - b_1 X_i) = 0$$

simplificando e expandindo, obtemos:

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n}X_i(Y_i - b_0 - b_1 X_i) = 0$$

$$\sum_{i=1}^{n}Y_i - nb_0 - b_1\sum_{i=1}^{n}X_i = 0$$

$$\sum_{i=1}^{n}X_iY_i - b_0\sum_{i=1}^{n}X_i - b_1\sum_{i=1}^{n}X_i^2 = 0$$

Daí, obtemos o *sistema de equações normais,* dado por:

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2$$

As equações normais podem ser resolvidas simultaneamente para b0 e b1(estimadores pontuais):

$$\boxed{\begin{aligned} b_1 &= \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2} \\ b_0 &= \tfrac{1}{n}\left(\sum Y_i - b_1 \sum X_i\right) = \bar{Y} - b_1 \bar{X} \end{aligned}}$$

Outra forma de escrevermos:

$$b_1 = \frac{\sum XY - \frac{\sum X \sum Y}{n}}{\sum X^2 - \frac{\left(\sum X\right)^2}{n}}$$

Ou

$$\mathbf{b = \frac{n\,(\sum xy) - (\sum x)\,(\sum y)}{n\,(\sum x^2) - (\sum X)^2}}$$

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

Propriedades dos estimadores de mínimos quadrados

Teorema de Gauss-Markov: Se as pressuposições do modelo de regressão linear (2) forem atendidas, os estimadores de mínimos quadrados b0 e b1 são não tendenciosos (unbised) e com variância mínima, entre todos os estimadores lineares não tendenciosos. Primeiro, o teorema diz que:

E(b0)=β0 e E(b1)=β1.

Segundo, o teorema diz que os estimadores b0 e b1 são mais precisos (isto é, as suas distribuições amostrais tem menor variabilidade) do que quaisquer outros estimadores pertencentes a classe dos estimadores não tendenciosos que são funções lineares das observações *Y1, Y2,...,Yn.* Os estimadores b0 e b1 são tais funções lineares das observações. Considere, por exemplo, b1,

$$b_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum (X_i - \overline{X}) Y_i}{\sum (X_i - \overline{X})^2} = \sum k_i Y_i$$

$$k_i = \frac{(X_i - \overline{X})}{\sum (X_i - \overline{X})^2}$$

Como *ki* são constantes (pois *Xi* são constantes conhecidas), b1 é uma combinação linear de *Yi* e, assim, é um *estimador linear*. Da mesma forma, *b0* também é um estimador linear. Entre todos os estimadores lineares não tendenciosos, *b0 e b1* tem *menor variabilidade* (demonstração adiante) em repetidas amostras nas quais os níveis de *X* são constante.

<u>Estimação da função de regressão</u>

A média do modelo de regressão linear é dada por:

$$E(Y) = \beta_0 + \beta_1 X$$

Estima-se a função de regressão por:

$$\hat{Y} = b_0 + b_1 X$$

Onde *Y (chapéu)* é o valor estimado da função no nível *X* da variável preditora.

A resposta média (*E(Y)*), corresponde a média da distribuição de probabilidade de *Y* no nível *X* da variável preditora. Pode-se demonstrar, como uma extensão do teorema de Gauss-Markov que *Y (chapéu)* é um estimador não tendencioso de *E(Y)*, com variância mínima dentro da classe dos estimadores lineares não tendenciosos. Temos:

$$\hat{Y}_i = b_0 + b_1 X_i \quad \mathrm{i} = 1,2,\ldots,\mathrm{n}$$

como sendo o valor ajustado para o *i*-ésimo caso

Exemplo: para os dados de uma porcentagem de acerto, os valores estimados da função de regressão são dados por:

$$\hat{Y} = 27{,}83633 - 0{,}0000642\,X$$

Suponha que estejamos interessados na porcentagem média de acerto na cache para *X=300.000 bytes* (muitas amostras com 300.000 bytes sob as mesmas condições que a equação foi estimada); a estimativa pontual vale:

$$\hat{y} = 27{,}83633 + 0{,}0000642(300000) = 47{,}10$$

Valores ajustados dos dados da amostra são obtidos substituindo-se os correspondentes valores da variável preditora *X* na função de regressão.

Exemplo:

ano	Quantidade y	Preço X
1	2	4
2	1	6
3	3	3
4	1	5
5	4	1
6	3	2
Totais	14	21
média	2,33	3,50

Agora vamos construir a planilha para podermos calcular a reta de regressão

ano	Quantidade y	Preço X	y^2	x^2	Y × X
1	2	4	4	16	8
2	1	6	1	36	6
3	3	3	9	9	9
4	1	5	1	25	5
5	4	1	16	1	4
6	3	2	9	4	6
Totais	14	21	40	91	38

$$\mathbf{b = \frac{n\,(\sum xy) - (\sum x)\,(\sum y)}{n\,(\sum x^2) - (\sum X)^2}}$$

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

$$b = \frac{6 \times (38) - (21)(14)}{6 \times (91) - (21)^2}$$

b= -0,6286 => este é o valor do coeficiente angular

$$a = \frac{14 - (-0{,}6286)\ (21)}{6}$$

a= 4,5334 => este é o intercepto

y = 4,533 - -0,6286X

Considerações Relevantes Sobre a Reta de Regressão

- A previsão da variável dependente resultará sempre em um valor médio, pois, a relação entre x e y é média.Assim, no caso acima, não obteremos, para um determinado preço, necessariamente um valor exato de quantidade a serem vendidas;
- Para fazermos previsões acerca da variável dependente y, não devemos utilizar valores da variável independente X que extrapolem o intervalo de valores utilizados no modelo de regressão .

Testes De Hipóteses e Intervalos de Confiança .

Os pares de valores (x , y) estão dispersos em relação a reta estimada. Isso ocorre, entre outras razões, porque existem inúmeras outras variáveis externas, não consideradas no modelo que influenciam y. Por exemplo, no caso acima a quantidade média de venda pode ser influenciada por gastos em propagandas, aumento de renda da população, etc...

Assim, não basta apenas calcularmos os coeficientes a e b da reta de regressão pelo MQ. Precisamos verificar até que ponto tais estimativas são suficientes para explicar o relacionamento entre as variáveis x e y.

Erro Padrão Das Estimativas .

Intuitivamente sabemos que quanto maior é a dispersão entre uma série de números ou população, maior será a dificuldade de se ajustar uma reta aos pontos. A dispersão pode ser estimada pela dispersão dos dados amostrais em relação `a reta de regressão. O erro padrão da estimativa (Se) é uma medida que avalia o grau de precisão da reta de regressão. A fórmula pra cálculo do Se é a seguinte:

$$Se = \frac{\sqrt{\sum (y - y^\wedge)}^{\,2}}{n - 2}$$

onde: y = valor observado da variável dependente;

y^ = Valor estimado da variável dependente;

n = número de observações.

Substituindo y^= a+b .x na equação acima completando os quadrados.

$$Se = \frac{\sqrt{\sum y2 - a\sum y - b\sum xy}}{n - 2}$$

$$Se = \frac{\sqrt{40 - 4{,}533.(14) - (-0{,}6286).38}}{6 - 2}$$

Se = 0,3236

Este valor deve ser interpretado como um desvio padrão; portanto, é medido na mesma unidade da variável dependente y. No nosso caso supondo que as peças sejam dadas em quantidade de mil teríamos: 0,3236 x 1.000 = 323,6 unidades.

Erro Padrão Do Coeficiente Linear (Sa)

O erro padrão Sa indica aproximadamente a distância entre os coeficientes a e o coeficiente linear populacional A, devido à dispersão dos dados amostrais. Assim, quanto menor for o erro S_a melhor será a precisão da estimativa. Veja a equação:

S_a.= Se $\sqrt[Se]{(1/n + médiaX / Sxx)}$

Onde :

Se = Erro padrão da estimativa ;

n = Número de observações

X- = média da variável independente (x)

$$Sxx = \sum x^2 - \frac{(\sum x)^2}{n}$$ => Tamanho da amostra

Agora vamos iniciar os cálculos

Média de X= 21/6 = 3,5

$$Sxx = 91 - \frac{(21)^2}{6}$$

Sxx = 17,5

Sa = 0,3236 x $\sqrt{(1/6 + (3{,}5)\hat{}2/17{,}5)}$

Sa= 0,301

Erro Padrão do Coeficiente Angular (S_b)

Analogamente ao As, o erro do coeficiente b (Sb) é a medida aproximada da distância entre a estimativa b e o coeficiente angular populacional B.

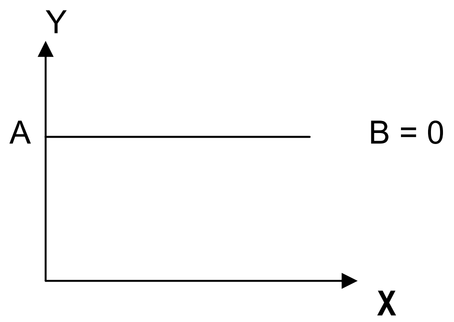
$$S_b = \frac{Se}{\sqrt{Sxx}}$$

$$S_b = \frac{0{,}3236}{\sqrt{17{,}5}}$$

$S_b =$ 0,077

Inferências Sobre o Coeficiente Angular (B)

Em algumas situações, mesmo não havendo relacionamento entre as variáveis na população, os dados amostrais podem sugerir a existência de relação. Isso ocorre quando, devido a fatores aleatórios, os dados extraídos da população dispõem-se de forma que seja possível traçar uma reta que se ajustam esses pontos. Por essa razão, sempre é preciso verificar se o modelo linear obtido é realmente significativo.



O gráfico acima mostra o não relacionamento entre X e Y.

Portanto, para que possamos verificar se as variáveis na população são mesmo relacionadas, devemos testar as seguintes hipóteses:

$H_0 : \beta_1 = 0$ Não há associação entre X e Y.

$H_a : \beta_1 \neq 0$

Diferentes tipos de testes podem ser empregados para esse propósito. Uma das maneiras consiste em elaborar um ***intervalo de confiança*** para o coeficiente angular B. Esse intervalo é definido como:

$b - t \cdot Sb \leq B \leq b + t \cdot Sb$

onde: b= Estimativa do coeficiente angular da reta.

t =valor critico, Distribuição t de probabilidades;

Sb= Erro padrão do coeficiente angular B.

Se o intervalo de confiança para B incluir o zero, não podemos rejeitar a hipótese nula. Caso o intervalo definido não inclua o zero, rejeitamos a hipóteses nula, admitindo, com risco conhecido de erro, que há relação significativa entre as variáveis.

Retomamos ao nosso exemplo, estabelecendo um grau de confiança de 95%. O valor do t critico será obtido usando a tabela na tabela 1, distribuição t, de student, constante no anexo A.

- O valor do nível de confiança α, que é de 5%;
- O número de graus de liberdade, que é obtido pela expressão n-k (número de observações – número de variáveis). Em nosso caso, 6-2= 4

A partir desses parâmetros, obtemos na tabela 1 um t crítico de 2,776. O intervalo de confiança, ao nível de 95%, será o seguinte:

$-0{,}6286 - (2{,}776) \cdot (0{,}0774) \leq B \leq -0{,}6286 + (2{,}776) \times (0{,}0774)$

$-0{,}8435 \leq B \leq -0{,}4137$

Estimamos que o valor do coeficiente B pertence ao intervalo [-0,8435 ; -0,4137] com um nível de confiança de 95%. Como o intervalo não inclui o zero, podemos rejeitar a hipótese nula, concluindo, com risco de erro de 5%, que há relação significativa entre as variáveis na população.

Outra maneira de testarmos a hipótese nula (B=0) é analisar a significância do coeficiente de regressão. Para isso, comparamos, em termos relativos, o valor da estimativa b com seu desvio padrão Sb.

$$t = \frac{b - B}{Sb}$$

Onde: b = Estimativa do coeficiente angular B;

B= Coeficiente angular da população;

Sb= Erro padrão e b.

Como a hipótese nula B= 0, a equação acima resulta em:

$$t = \frac{b}{Sb}$$

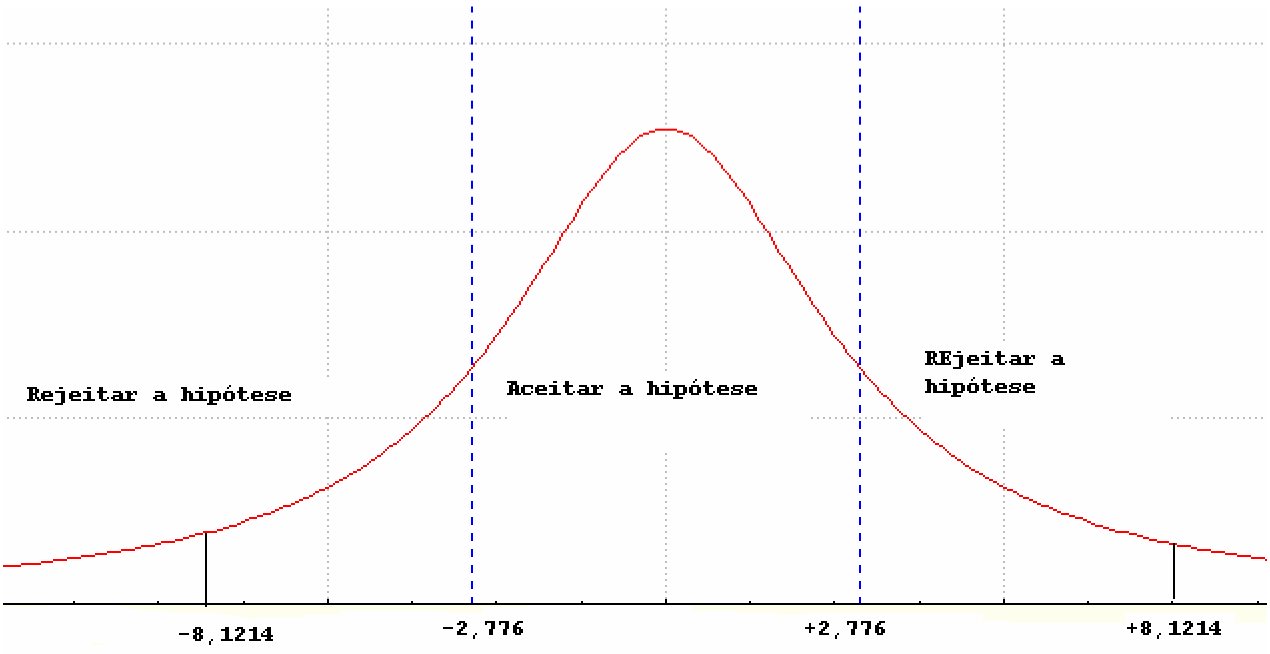
Nota: O Excel dispõe da função estatística INVT, por meio da qual pode ser obtido o valor do t crítico.

O valor t (student) pode ser interpretado como o número de desvios padrões que o estimador b dista do ponto zero. Quanto maior for essa distância , maior será a chance de B ≠ 0.

$$t = \frac{-0,6286}{0,0774}$$

t = -8,1214.

O valor de t crítico calculado no item anterior é de 2,776. Assim concluímos que devemos rejeitar H=0, pois I t teste= -8,1214 < -t critico= -2,776.



Para calcular a probabilidade P de obtemos uma estatística t igual ou superior a este valor, vamos utilizar a função estatística do excel (DISTT).

Valor P= Distt (tteste;graus de liberdade; caudas).

- - Caudas: deve ser sempre igual a 2, pois o teste do valor P é bicaudal
- - t teste = calculado como indicado. E inserior o mesmo na função com valor postivo.

- - Graus_de_liberdade= corresponde a n-2. sendo n o tamanho da amostra .

Valor P = Distt (8,1214 ; 4;2)

- Se o Valor de P for ≤ α, reijeitamos a hipótese nula
- Se o valor de P for > α aceitamos a hipótese nula.

O gráfico acima nos mostra muito bem, os parâmetros de rejeição e aceitação.

Covariância e Coeficiente de Correlação

A covariância mede a força do relacionamento entre duas variáveis em termos absolutos através da seguinte equação:

$$\text{Cov}(X, Y) = \frac{\sum (\text{x- média x}) \times (\text{y- média y})}{n-1}$$

ano	Quantidade y	Preço X	X - média x	y - média Y	x . Y
1	2	4	0,50	(0,33)	-0,16667
2	1	6	2,50	(1,33)	-3,33333
3	3	3	(0,50)	0,67	-0,33333
4	1	5	1,50	(1,33)	-2
5	4	1	(2,50)	1,67	-4,16667
6	3	2	(1,50)	0,67	-1
Totais	14	21	-		(11,00)

$$\text{Cov}(X, Y) = \frac{-11}{6-1}$$

Cov (X, Y) = -2,2

O coeficiente de correlação mede o grau de relação entre duas variáveis . A correlação (representada por r) está sempre ente -1 e 1.

$$r = \frac{\text{Cov}(x,y)}{\sigma x \cdot \sigma y}$$

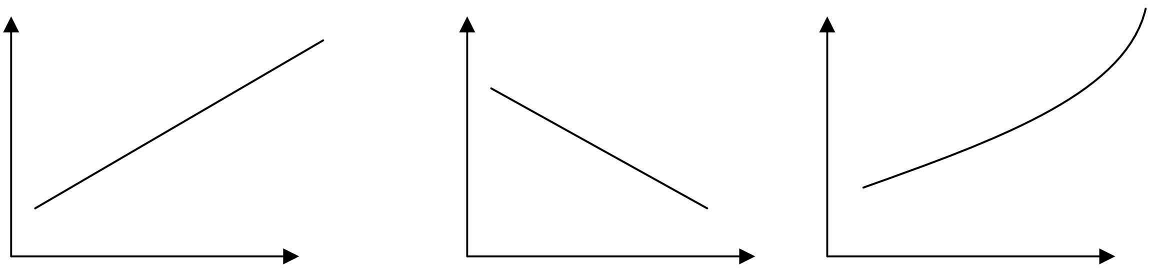
ano	Quantidade y	Preço X
1	2	4
2	1	6
3	3	3
4	1	5
5	4	1
6	3	2
Totais	14	21
Desvio padrão	1,21	1,87

$$r = \frac{-11}{1{,}2111 \times 1{,}871}$$

$r = -0{,}97099$

o valor -1 corresponde à correlação negativa perfeita e o valor de +1 corresponde a correlação positiva perfeita. O coeficiente de correlação (zero) indica que as duas variáveis não estão correlacionadas linearmente. Nesse caso podem estar relacionadas por uma outra função matemática, como por exemplo, $y = X^2$

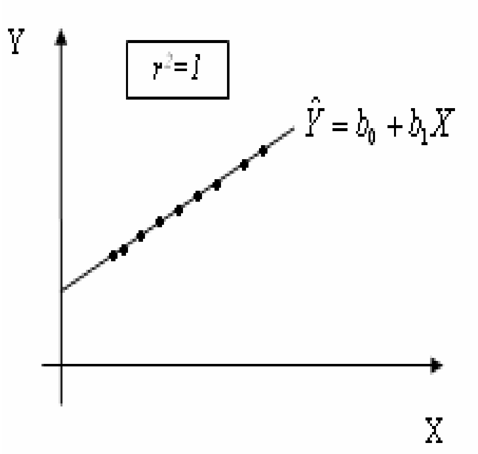
O gráficos abaixo vão ilustrar claramente o que foi comentado acima:

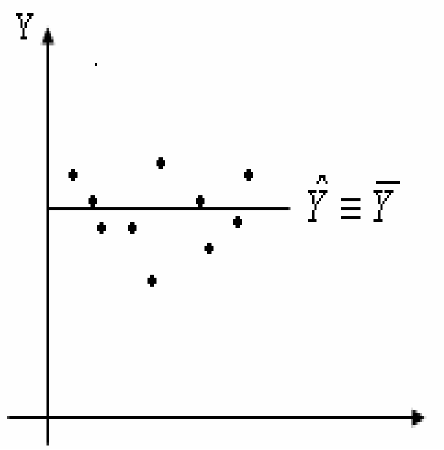


r = +1 r = -1 r =0

Coeficiente de Determinação

O coeficiente de determinação mede o grau de ajustamento da reta de regressão aos dados observados. Indica a proporção da variação total da variável dependente que é explicada pela variação da variável independente.





A equação que permite calcular o coeficiente de determinação é a seguinte

$$R^2 = \frac{\sum (Y\text{^}-\text{média } y)^2}{\sum (y-\text{média } y)^2} = \frac{\text{Variação explicada}}{\text{Variação Total}}$$

ano	Quantidade y	Preço X	y^	y^-médiay	(y^-y)^2	y-médiay	(y-médiay)^2
1	2	4	2,019	(0,3143)	0,09881	(0,3333)	0,1111
2	1	6	0,7619	(1,5714)	2,4694	(1,3333)	1,7778
3	3	3	2,6476	0,3143	0,09876	0,6667	0,4444
4	1	5	1,3905	(0,9428)	0,88893	(1,3333)	1,7778
5	4	1	3,9048	1,5715	2,46951	1,6667	2,7778
6	3	2	3,2762	0,9429	0,889	0,6667	0,4444
Totais	14	21			6,91441		7,33333333

$$R^2 = \frac{6,9141}{7,3333} = 0,942828 \text{ ou } 94,2828\%$$

Este resultado nos mostra que 94,2828% da variações da variável dependente y podem ser explicadas por variações da variável independente x.

Usando O Excel

Agora vamos praticar um pouco a análise usando o Excel, primeiro vamos construir o gráfico com a linha de tendência é a devida equação da reta de regressão.

	A	B	C
1	ano	Preço X	Qtde Y
2	1	4	2
3	2	6	1
4	3	3	3
5	4	5	1
6	5	1	4
7	6	2	3

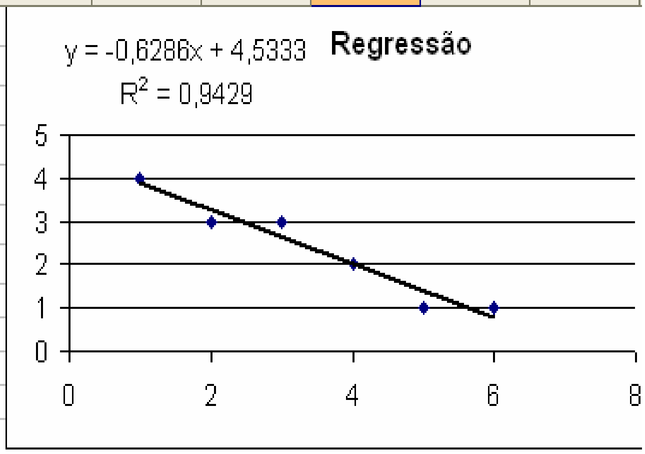


Fig1: Excel 1

Primeiramente, devemos obter o diagrama de dispersão. Os passos são os seguintes:

- Digitar os dados do problema em uma planilha: primeiro, a coluna da variável independente (X) ; depois, a coluna da variável depende (Y);
- Selecionar o menu inserir e clicar em gráfico;
- Selecionar o tipo de gráfico, dispersão e clicar em avançar;
- Selecionar as células que contêm os dados do problema e inseri-los no compo intervalo de dados; selecionar a seqüência colunas e clicar em avançar;

Para obter a reta de regressão e o coeficiente de determinação, devem ser seguintes as etapas:

- Selecionar qualquer dos pontos do gráfico de dispersão com o botão esquero do mouse;
- Clicar como o botão direito do mouse sobre o gráfico e selecionar adicionar linha de tendência;
- Verificar se na aba tipo a opção linear está ativada;
- Clicar na aba opções, selecionar as alternativas exibir equação no gráfico e exibir o valor do R –quadrado no gráfico e clicar em ok. Essa opção é útil na avaliação de que modelo matemático se ajusta melhor ao conjunto de dados (linear, potência etc..).

Para gerar os relatórios sobre a regressão os passos são os seguintes:

- Selecionar o menu ferramentas e clicar em análise de dados, Caso não esteja habilitada está função clique em ferramentas e depois em suplementos e marque a opção análise de dados;
- Selecionar Regressão na caixa diálogo ferramentas análise de dados e clicar em OK. É mostrado na caixa de diálogo Regressão, como indica a Figura abaixo;
- Inserir as células que contêm os valores das variáveis X e Y nos campos intervalo Y de entrada e intervalo X de entrada;
- Selecionar as seguintes caixas de verificação: Nível de confiança de 95%, Resíduos, Resíduos padronizados e plotar resíduos, plotar ajuste de linha, plotagem de probabilidade normal;
- Clicar em OK para que o Excel gere os relatórios de análise.

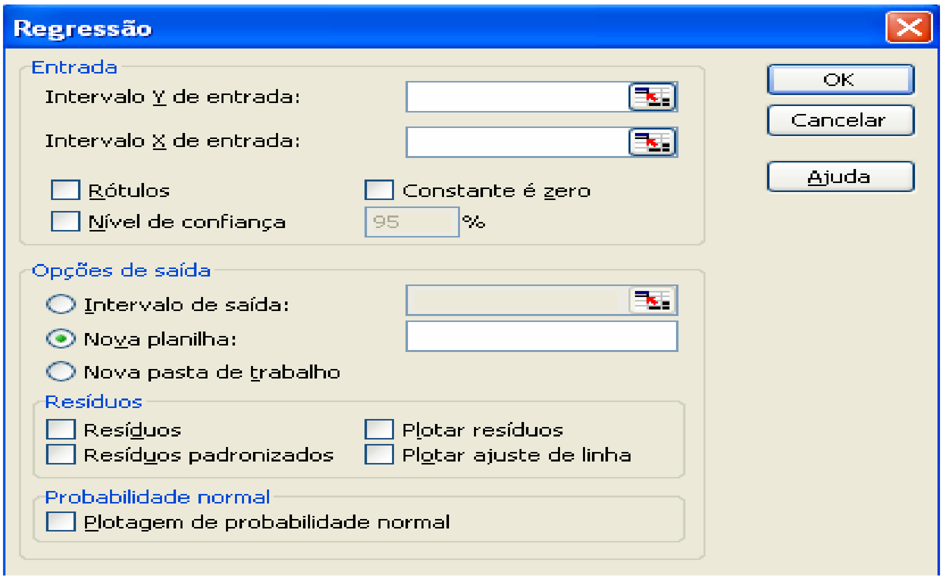


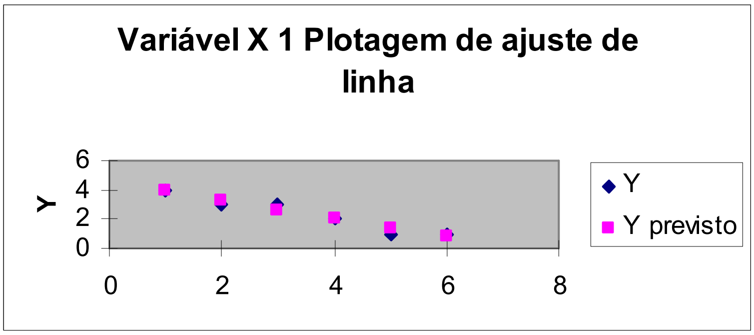
Fig2: Caixa de diálogo- Regressão.

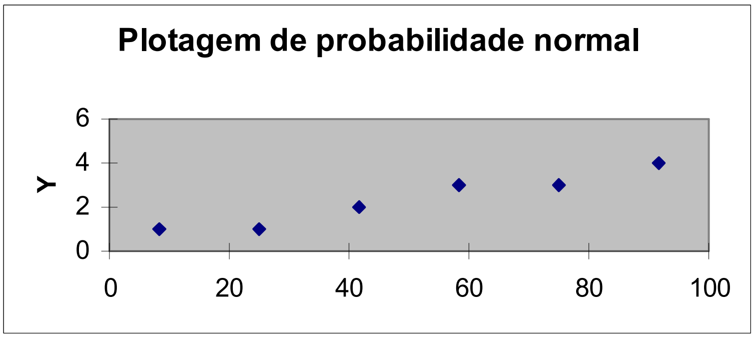
	A	B	C	D	E	F	G	H	I
	RESUMO DOS RESULTADOS								
	Estatística de regressão								
	R múltiplo	0,971008312							
	R-Quadrado	0,942857143							
	R-quadrado ajustac	0,928571429							
	Erro padrão	0,323669437							
	Observações	6							
	ANOVA								
		gl	*SQ*	*MQ*	*F*	*F de significação*			
	Regressão	1	6,914285714	6,91429	66	0,001248593			
	Resíduo	4	0,419047619	0,10476					
	Total	5	7,333333333						
		Coeficientes	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*95% superiores*	*nferior 95,0%*	*Superior 95,0%*
	Interseção	**4,533333333**	0,301319848	15,0449	0,000113738	3,696735316	5,36993135	3,6967353	5,36993135
	Variável X 1	**-0,628571429**	0,077371794	-8,12404	0,001248593	-0,843389968	-0,413752889	-0,84339	-0,413752889

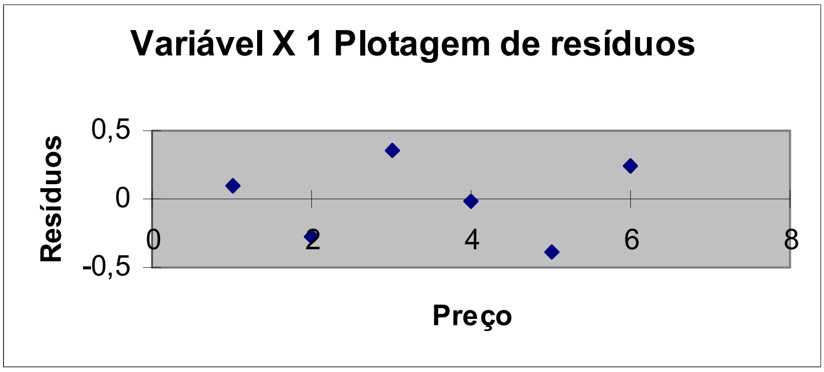
Fig3: Relatórios da análise de Regressão

22	RESULTADOS DE RESÍDUOS				RESULTADOS DE PROBABILIDADE	
23						
24	*Observação*	*Y previsto*	*Resíduos*	*Resíduos padrão*	*Percentil*	*Y*
25	1	2,019047619	-0,019047619	-0,065795169	8,333333333	1
26	2	0,761904762	0,238095238	0,822439619	25	1
27	3	2,647619048	0,352380952	1,217210636	41,66666667	2
28	4	1,39047619	-0,39047619	-1,348800975	58,33333333	3
29	5	3,904761905	0,095238095	0,328975847	75	3
30	6	3,276190476	-0,276190476	-0,954029958	91,66666667	4

Fig 4: Relatórios da análise de Regressão.







Estes são os resumos da análise de regressão.

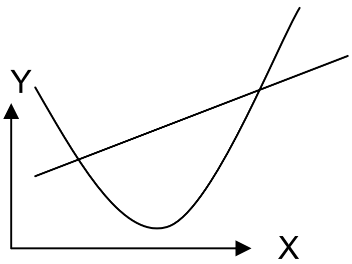
Repare que os relatórios vão nos fornecer os mesmos valores já calculados anteriormente em negrito, a equação da reta de regressão.

Suposições básicas de validação de um modelo de Regressão.

Para que todos os testes estatísticos vistos anteriormente tenham plena validade é preciso que o modelo de regressão sob análise siga os pressupostos básicos referentes à regressão.

- Independência de erros ou autocorrelação residual: os resíduos (u) devem ser distribuídos aleatoriamente em torno da reta de regressão. Portanto, os resíduos não devem ser correlação entre si;
- Homocedasticidade: a variância dos resíduos (u) é constante para todos os valores de x. para verificação desse pressuposto, utilizaremos o teste de pesaran-pesaran;
- Normalidade: o comportamento dos resíduos (u) segue uma distribuição normal de probabilidades. Para verificação desse pressuposto, utilizaremos o teste de Kolmogorov – Smirnov;
- Linearidade: pelo diagrama de dispersão, obteremos uma boa idéia sobre se as variáveis x e y são lineares, ou seja, se podemos ajustar uma reta aos pontos observados.

As causas de autocorrelação, pode acontecer quando utilizamos uma relação linear para definir o comportamento de x e y, em vez de usar uma relação exponencial como mostra gráfico abaixo:



Um outro motivo que pode causar autocorrelação é a omissão da variável independente importante para o modelo de regressão. Se uma variável explicativa de grande relevância for omitida do modelo, o comportamento dos resíduos refletirá a tendência dessa variável porque, ao não ser considerada, ela passa a "pertencer" ao resíduo.

Tipos de autocorrelação Residual

- Autocorrelação positiva: há relação linear,"direta" entre os resíduos ao longo do tempo, ou seja, um resíduo positivo é geralmente seguido de um resíduo positivo e vice-versa;
- Autocorrelação negativa: há relação linear "inversa" entre os resíduos ao longo do tempo, ou seja, um resíduo positivo é geralmente seguido de um resíduo negativo é vice-versa.

Teste de Durbin –Watson

Esse teste serve para detectar se há presença significativa de autocorrelação entre os resíduos em um modelo de regressão. O coeficiente de Durbin- Watson mede a correlação entre cada resíduo e o resíduo da observação imediatamente anterior.

Equação é a seguinte:

$$DW = \frac{\sum (ut-ut-1)^2}{\sum u_t^2}$$

Agora vamos construir a planilha de cálculo.

yt	y^t	ut=yt-y^t	ut^2	Ut-1	Ut-Ut-1	(Ut-Ut-1)^2
2	2,019	-0,019	0,00035721		-0,0189000	0,0003572
1	0,762	0,238	0,05678689	-0,018900	0,2572000	0,0661518
3	2,648	0,353	0,12425625	0,238300	0,1142000	0,0130416
1	1,390	-0,390	0,15233409	0,352500	-0,7428000	0,5517518
4	3,905	0,095	0,00908209	-0,390300	0,4856000	0,2358074
3	3,276	-0,276	0,07623121	0,095300	-0,3714000	0,1379380
Total			**0,41904774**			**1,0050479**

$$DW = \frac{1{,}0050479}{0{,}41904774}$$

DW =2,398

A análise da estatística de Durbin-Watson é feita a partir de uma tabela de valores críticos (ver tabela 2 anexo). As conclusões sobre o coeficiente de Dw devem considerar o limite crítico inferior (dL) e o limite crítico superior (du), obtidos nessa tabela, que depende no nível de significância (α), do número de variáveis independentes (K-1), sendo K o nº de coeficiente da regressão, e do tamanho da amostra (n) do modelo.

Análise Dos Resíduos

Outra maneira de obtermos informações preciosas sobre um modelo de regressão consiste em fazer um gráfico dos resíduos. Para cada ponto dos dados (Xi e Yi), calculamos o resíduo.

Resíduo (u) = yi-Y^xi = y – (m^xi + b^).

A equação da reta de regressão é a seguinte

y = -0,6286x + 4,5333

Agora vamos encontrar o quadro dos resíduos

ano	Preço X	Qtde Y	-0,6286x + 4,5333 (y^)	Resíduo (Yi-y^xi)
1	4	2	2,019	-0,019
2	6	1	0,762	0,238
3	3	3	2,648	0,353
4	5	1	1,390	-0,390
5	1	4	3,905	0,095
6	2	3	3,276	-0,276
Total	21	14	13,9992	0,0008

Agora vamos construir um gráfico de dispersão, com os valores de X dispostos no eixo horizontal e os resíduos ao longo do eixo vertical. Temos:

Preço X	Resíduo
4	-0,019
6	0,238
3	0,353
5	-0,390
1	0,095
2	-0,276

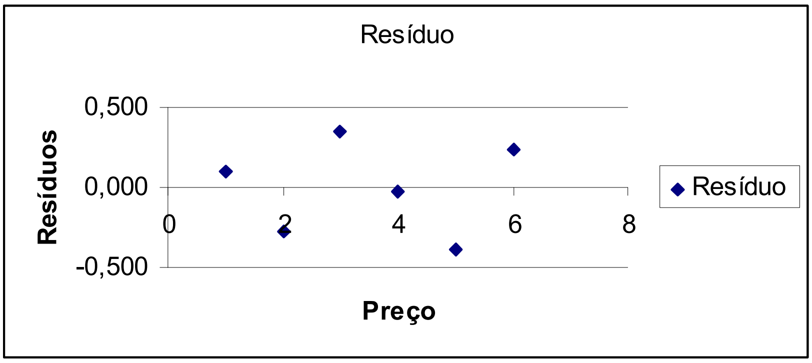


Gráfico dos Resíduos.

Como podemos observar, os resíduos não se concentram acima ou abaixo da linha horizontal. Eles se distribuem aleatoriamente ao longo da reta, pressuponde-se portando que não há haja autocorrelação. Para verificar se de fato não há correlação, utilizaremos o teste Durbin- Watson.

- Se DW < dL, concluímos que há autocorrelação positiva;
- dL < DW < du, podemos dizer que o este é não conclusivo;
- du < DW < 4- du, concluímos pela ausência de autocorrelação;
- 4-du <DW < 4-dL, podemos dizer que o teste é não conclusivo;
- Se DW > 4-dL, concluímos que há autocorrelação negativa.

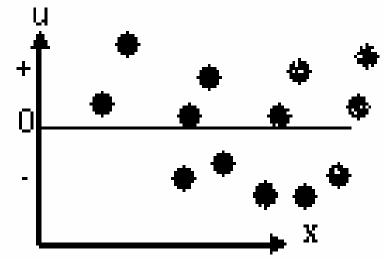
O valor de DW já foi calculado anteriormente DW= 2,398.
Agora vamos obter o valor de dL e du correspondente na tabela Durbin-Watson.

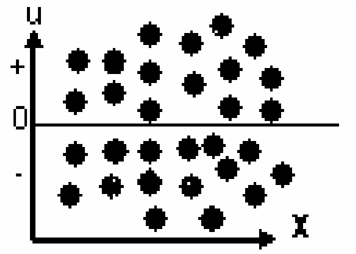
Vamos obter os valores de dL e du correspondente na tabela de Durbin-Watson (tabela 2 do anexo). Os parâmetros a serem considerados são (α=0,05), (K-1=1), pois K=2 (no caso, os coeficientes são 2, a e b) e (n=6). Encontramos os valores de dL=0,610 e de du= 1,4. Portando os limites da área de aceitação são: 2 e 3,6 (que corresponde a (4-du= 4- 1,4). Como DW, de 2,398, situa-se dentro deste intervalo, podemos aceitar a hipótese da ausência de autocorrelação residual e, consequentemente, utilizar o método dos mínimos quadrados.

Avaliando a Homocedasticidade

Quando os resíduos se distribuem aleatoriamente em torno da reta de regressão e de forma constante, ou seja, a variância dos resíduos é igual a uma constante para todos os valores de X, temos que o pressuposto da homocedasticidade está satisfeito.

A violação do pressuposto da homocedasticidade compromete a eficiência das estimativas do modelo de regressão. A seguir encontram-se dois gráficos que vão demonstrar a diferença entre homocedasticidade e hetorocedasticidade.

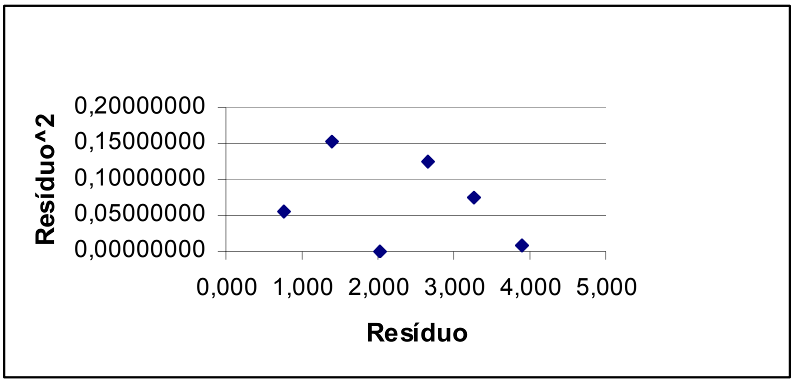




O gráficos acima representam a homocedasticidade e a hetorocedasticidade. No primeiro gráfico observamos uma homogeniedade na variância dos resíduos à medida que x cresce. Já no outro gráfico observamos que a variabilidade dos resíduos cresce à medida que a variável X cresce, demonstrando a violação da hipótese, caracterizando o que chamamos de homocedasticidade.

O teste de Pesaran- Pesaran, consiste em detectar a presença de heterocedasticiade com base nos resultados da regressão em que a variável dependente representa os valores dos quadrados dos resíduos (u^2) e a variável independente é constituída pelos valores estimados da variável dependente (y^).

y^ x	$Ut^{2\ y}$
2,019	0,00035721
0,762	0,05678689
2,648	0,12425625
1,390	0,15233409
3,905	0,00908209
3,276	0,07623121



Graficamente podemos observar que os valores estão bem dispersos, podemos concluir, que há homocedasticidade. Usando a ferramenta de regressão do Excel, vamos calcular a regressão dos dados acima

1	RESUMO DOS RESULTADOS								
2									
3	*Estatística de regressão*								
4	R múltiplo	0,30118202							
5	R-Quadrado	0,09071061							
6	R-quadrado ajustado	-0,13661174							
7	Erro padrão	0,06487541							
8	Observações	6							
9									
10	ANOVA								
11		*gl*	*SQ*	*MQ*	*F*	*de significação*			
12	Regressão	1	0,00167948	0,00168	0,39904	0,56188718			
13	Resíduo	4	0,01683527	0,00421					
14	Total	5	0,01851476						
15									
16		*Coeficientes*	*Erro padrão*	*Stat t*	*valor-P*	*95% inferiores*	*% superior*	*ferior 95,0*	*uperior 95,0*
17	Interseção	0,1062077	0,06336857	1,67603	0,16904	-0,0697317	0,28215	-0,06973	0,2821471
18	y	-0,01558527	0,02467212	-0,6317	0,56189	-0,0840861	0,05292	-0,08409	0,0529155

O Coeficiente de determinação ou R-Quadrado, nos diz que há pouca correlação entre as variáveis e o valor de P = 0,56189 ou 59,189% (e o F significação) > 5%. Portando, podemos aceitar a hipótese nula e ausência de hetorocedasticidade

.

Avaliando a normalidade

Os testes de significância e os intervalos de confiança das estimativas do modelo de regressão são baseados no pressuposto da normalidade, isto é , que os resíduos apresentam distribuição normal. A violação da normalidade pode estar ligada a alguns aspectos relacionados ao modelo, tais como: omissão de variáveis explicativas importantes, inclusão de variável explicativa irrelevante para o modelo, utilização de relação matemática incorreta para análise entre as variáveis do modelo.

Utilizaremos o teste não paramétrico do Kolmogorov – Smirnov, para avaliação da normalidade, cuja estatística de teste é a seguinte:

D = máx (i/n- Zi)

Onde: n = tamanho da amostra, sendo i = 1,2,3..............n;

Zi= probabilidade acumulada da distribuição normal padronizada, considerando os valores hi = (ui /s), onde ui são os resíduos ordenados de forma crescente e S é o desvio padrão dos ui.

Para realização do teste, procedemos da seguinte maneira:

- Se $D < D_{Crítico}$, aceitamos a hipótese de que os resíduos se distribuem normalmente;
- Se $D > D_{Crítico}$, rejeitamos a hipótese de que os resíduos se distribuem normalmente.

O valor do $D_{Crítico}$, é obtido a partir da tabela 3 anexo A. Os parâmetros são:

1. O valor do nível significância α, que é de 5%;
2. O tamanho da amostra (n). Em nosso caso, n= 6

Agora vamos construir a planilha de cálculo

I (1)	U1(2)	hi = ui /s	zi(4)	I i/ n	D= 5-4
1	-0,3905	-1,349	0,089	0,1667	0,078
2	-0,2762	-0,954	0,170	0,3333	**0,163**
3	-0,0190	-0,066	0,474	0,5	0,026
4	0,0952	0,329	0,629	0,6667	0,038
5	0,2381	0,822	0,795	0,8333	0,039
6	0,3524	1,217	0,888	1	0,112
Desvio padrão S	0,289514999				

O valor de Zi, foi obtido usando a função estatística do excel Dist.normP(hi).

Verificamos que o maior valor de D= 0,163 (D_{Teste}), portanto como (D_{Teste}), é menor que o $D_{Crítico}$ (0,56) para um nível de confiança de 5%, concluímos que os resíduos se comportam seguindo uma distribuição normal de probabilidades.

Intervalo de Confiança para a Regressão

Quando calculamos o valor de y^ considerando um valor para a variável X, há duas alternativas que levam ao mesmo resultado para y^,mas que produzem resultados diferentes com respeito ao intervalo de confiança. Essas duas alternativas e as fórmulas de cálculo de seus desvios padrões são as seguintes:

a) considerando y^ como valor médio para dado X:

S média de y = Se . $\sqrt{1/n+(x-médiax)^\wedge 2/sxx}$

b) considerando y^ como valor individual para dado X:

Sy = Se . $\sqrt{1+1/n+(x-méidaX)^\wedge 2/Sxx}$

Assim, os intervalos de confiança resultantes para essas duas situações esta descritos a seguir:

Y^médio = y^ ± t . S médiay

Y^individual: y^ ± t . Sy

Para demonstrar o cálculo dos dois intervalos de confiança, retomemos o caso inicial. Vamos estimar X= 4, sabemos que a equação da reta de regressão é:

Y = 4,535 – 0,6286X

Y= 4,535 – 0,6286 x (4)= 2,021

Obtido y^, e considerando t = 2,7764 (α= 5% e Grau de liberdade GL= 6-2=4); Sxx= 17,5 e Se = 0,3236 temos:

S-$_y$ = 0,3236 . $\sqrt{1/6+(4-3{,}5)^\wedge 2/17{,}55}$

S-$_y$ =0,1377

Sy= 0,3236 . $\sqrt{1+1/6+(3-3{,}5)^\wedge 2/17{,}5}$

Sy= 0,3517

Os intervalos de confiança para y^ considerando um X = 4 são:

Y^médio= 2,021 ± 2,7764 . 0,1377 = 2,021 ± 0,3823

Y^individual = 2,021 ± 2,7764 . 0,3517 = 2,021 ± 0,9765

Solução de Exercícios

1. Suponha que os dados seguintes se referem a uma amostra de observações sobre a despesa mensal em bens e em serviços culturais (Y) e o rendimento mensal *per capita* (X) de 14 famílias.

Y	0	44	15	52	39	1	40	2	34	46	22	26	1	57
x	50	250	110	480	190	80	210	90	150	310	120	150	70	650

a) Escreva a equação da reta de regressão.

b) Determine uma estimativa da despesa mensal em bens e serviços para uma família que apresente um rendimento *per capita* igual a 250?

a) **Vamos rever alguns conceitos**

m^ = (média de xy – média de x vezes a média de y)/média de x^2- A média de x elevada ao quadrado. Este é o coeficiente angular.

b^= média de y- m^ vezes a média de x => este é o intercepto

Total	Y	X	Y^2	xy	x^2
1	0	50	0	0	2500
2	44	250	1936	11000	62500
3	15	110	225	1650	12100
4	52	480	2704	24960	230400
5	39	190	1521	7410	36100
6	1	80	1	80	6400
7	40	210	1600	8400	44100
8	2	90	4	180	8100
9	34	150	1156	5100	22500
10	46	310	2116	14260	96100
11	22	120	484	2640	14400
12	26	150	676	3900	22500
13	1	70	1	70	4900
14	57	650	3249	37050	422500
Total	379	2910	15673	116700	985100
Média	27,07143	207,85714	1119,50000	8335,71429	70.364,28571

Vamos resolver:

M^= 8335,71429 – (27,07143 x 207,85714)/ 70364,28571 – $(207,85714)^2$
M^= 0,0997

b^= 27,07143– (0,0997x 207,85714)
b^ = 6,3412

Assim a equação da reta de regressão é :
y = 0,0997x + 6,3412

b) Determine uma estimativa da despesa mensal em bens e serviços para uma família que apresente um rendimento *per capita* igual a 250?

R: **y = 0,0997x + 6,3412**

Y= 0,0997 x (250) + 6,3412 = R$ 31,27
Agora vamos calcular o $\mathbf{r^2}$

. $\mathbf{r^2}$= 1- QE reta/QE méd

QEmed= ∑ (yi-média de y)^2

QE med= n x var (y)

QE reta= ∑ (yi - ^yxi)^2

y^	Resíduo	resíduo^2
11,326	-11,3262	128,283
31,266	12,7338	162,15
17,308	-2,3082	5,32779
54,197	-2,1972	4,82769
25,284	13,7158	188,123
14,317	-13,3172	177,348
27,278	12,7218	161,844
15,314	-13,3142	177,268
21,296	12,7038	161,387
37,248	8,7518	76,594
18,305	3,6948	13,6515
21,296	4,7038	22,1257
13,32	-12,3202	151,787
71,146	-14,1462	200,115

∑ do resíduo ^2 = 1630,83
Q méd= n x var (y)

Var (y) = média de y^2 –(média de y)^2

Var (y) = 1119,50000- $(27,07143)^2$

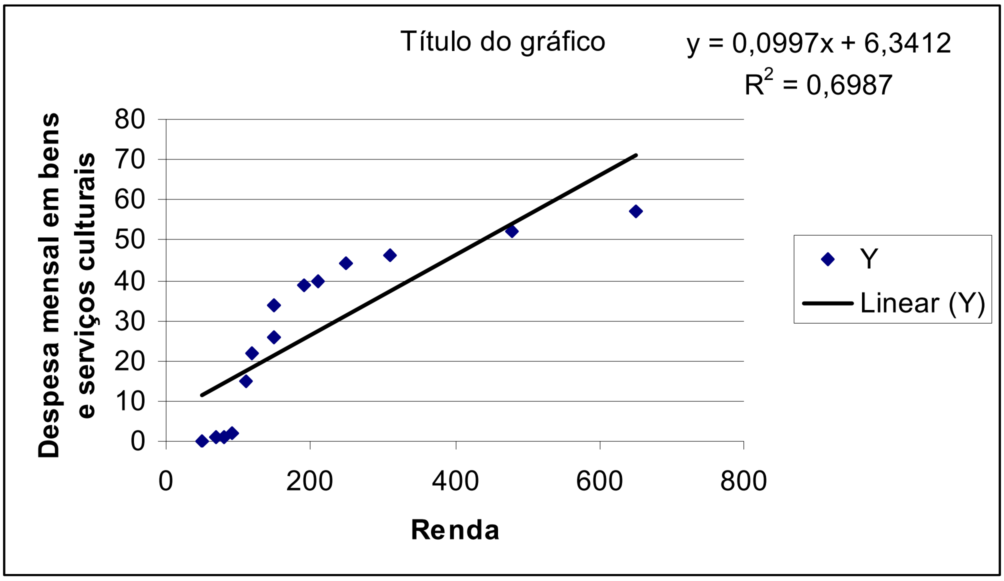
Var (y) = 613,3623

QE Reta= 613,3623 x 14 = 8.587,0725

r^2= 1- 1630,83/8.587,0725= 081 ou 81%, isto nos mostra que 81% das variações dos serviços culturais, podem ser explicados pelo por variações na renda das pessoas.

Cabe mencionar que este valor 0,81 reflete o valor do coeficiente de correlação ao quadrado, este mede o grau de correlação entre duas variáveis e está sempre em -1 e 1, quanto mais positivo forem significa que a correlação e maior, quando mais negativo for significa que a correlação é menor.

r= (0,81)^1/2 = 0,9 ou 90%, demonstrando o alto nível de correlação entre as variáveis



Este gráfico nos mostra que há pouca dispersão nas variáveis, demonstrando que o nível de correlação é alto.

2. Suponha que um automóvel, para analisar o seu consumo de combustível, efetuou 7 viagens, tendo-se registrado a distância percorrida (km) e o consumo (l), obtendo-se, então, os 7 pares de valores seguintes:

Y(distância)	20	40	80	120	160	200	250
X(consumo)	2	3	5	9	12	14	18

a) Escreva a equação da reta de regressão estimada que relaciona distância em relação ao consumo.

b) Com 16 litros de combustível qual das duas distâncias lhe parece mais provável de ser percorrida: 190 km ou 205 km?

c) Sendo o valor do litro de gasolina R$ 2,52, qual o valor gasto (estimado) em um trajeto de 820 km?

R:

Total	x consumo	y Distância	x^2	xy	y^2
1	2	20	4	40	400
2	3	40	9	120	1600
3	5	80	25	400	6400
4	9	120	81	1080	14400
5	12	160	144	1920	25600
6	14	200	196	2800	40000
7	*18*	*250*	324	4500	62500
Média	9,000000	124,285714	111,857143	1551,428571	21557,142857

M^= 1551,428571 – (9 x 124,285714)/(111,857143-9^2)

M^= 14,0277778
b ^= 124,285714 – (14,0277778 x 9)

b^= -1,96428571
A equação da reta de regressão é :
Y= 14,0277778X - 1,96428571

b) Com 16 litros de combustível qual das duas distâncias lhe parece mais provável de ser percorrida: 190 km ou 205 km?

R: y= 14,0277778X - 1,96428571

Y= 14,0277778 x (16) - 1,96428571

Y= 226,4087302
Logo a distância de 205Km é a mais provável a ser percorrida

c) Sendo o valor do litro de gasolina R$ 2,52, qual o valor gasto (estimado) em um trajeto de 820 km?

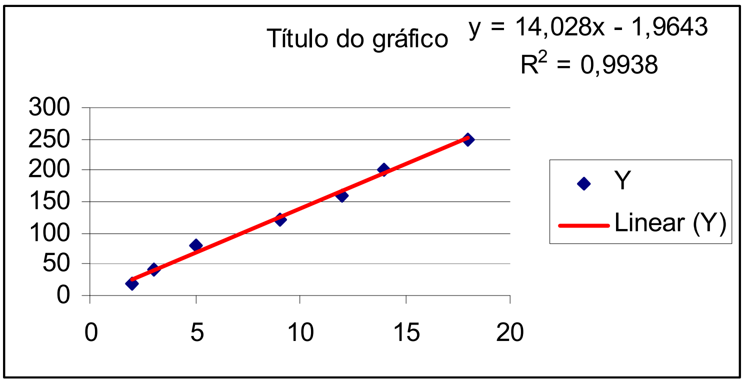
Y= 14,0277778X - 1,96428571

820= 14,0277778X - 1,96428571

820+1,96428571= 14,0277778X

X= 58,59547383 KM
Gasto= 58,59547383 xR$ 2,52= R$ 147,66

A correlação r para este caso é = 0,9968708 ou 99,68707999%, este coeficiente de correlação é praticamente perfeito, pois a cada 1% de variação no consumo ocorre uma variação de 99,68707999% na distância.
Veja o gráfico



Este gráfico nos mostra que há Uma correlação quase perfeita

Bibliografias

London: Chapman-Hall, 1998. 399 p.MOOD, A. M.; GRAYBILL, F. A.; BOES D. C. *Introduction to the theory of statistics***.** 3. ed. New York: McGraw Hill, 1974, 842p.
MORAN, P. A. P. Some theorems on time series II. The significance of the series correlation coefficient. *Biometrika***,** London, v.35, n.3/4, p.255-260, 1948.
NETER, J.; WASSERMAN, W.; KUTNER, H. M. *Applied linear statistical models: regression, analysis of variance, and experimental designs*, 2. ed. Illinois: Richard D. Irwin**,** 1974. 1184p.
YOUNG, L. C**.** On randomnes in ordered sequences. *Ann. Math. Stat.,* Baltimore, v.12, p.296-300, 1941.
STEVESON, William j. Estatística aplicada à administração. Tradução de Alfredo alves de farias. São Paulo: Harper & Row, 1992.
Pesquisa Operacional para decisão em contabilidade e administração: Contabilometria/ Luiz J. Corrar, Carlos Renato Theóphilo, (coordenadores), -- São Paulo: Atlas, 2004.
PINDYDK, Robert S. ; RUBINFELD, Daniel L. Econometric models and Economic Forecasts.2.ed. Cingapura: MCGraw-Hill,1986.

Anexos

TABELA 5 – Distribuição t de Student

Probabilidade: $P(t > t_c) = p/2$ ou $P(t < -t_c) = p/2$

Exemplo: GL=n-1= 15 - $P(t > 2{,}1315) = 0{,}05/2 = 0{,}025$

GL	p=0,90	p=0,80	p=0,70	p=0,60	p=0,50	p=0,40	p=0,30	p=0,20	p=0,10
1	0,1584	0,3249	0,5095	0,7265	1,0000	1,3764	1,9626	3,0777	6,3137
2	0,1421	0,2887	0,4447	0,6172	0,8165	1,0607	1,3862	1,8856	2,9200
3	0,1366	0,2767	0,4242	0,5844	0,7649	0,9785	1,2498	1,6377	2,3534
4	0,1338	0,2707	0,4142	0,5686	0,7407	0,9410	1,1896	1,5332	2,1318
5	0,1322	0,2672	0,4082	0,5594	0,7267	0,9195	1,1558	1,4759	2,0150
6	0,1311	0,2648	0,4043	0,5534	0,7176	0,9057	1,1342	1,4398	1,9432
7	0,1303	0,2632	0,4015	0,5491	0,7111	0,8960	1,1192	1,4149	1,8946
8	0,1297	0,2619	0,3995	0,5459	0,7064	0,8889	1,1081	1,3968	1,8595
9	0,1293	0,2610	0,3979	0,5435	0,7027	0,8834	1,0997	1,3830	1,8331
10	0,1289	0,2602	0,3966	0,5415	0,6998	0,8791	1,0931	1,3722	1,8125
11	0,1286	0,2596	0,3956	0,5399	0,6974	0,8755	1,0877	1,3634	1,7959
12	0,1283	0,2590	0,3947	0,5386	0,6955	0,8726	1,0832	1,3562	1,7823
13	0,1281	0,2586	0,3940	0,5375	0,6938	0,8702	1,0795	1,3502	1,7709
14	0,1280	0,2582	0,3933	0,5366	0,6924	0,8681	1,0763	1,3450	1,7613
15	0,1278	0,2579	0,3928	0,5357	0,6912	0,8662	1,0735	1,3406	1,7531
16	0,1277	0,2576	0,3923	0,5350	0,6901	0,8647	1,0711	1,3368	1,7459
17	0,1276	0,2573	0,3919	0,5344	0,6892	0,8633	1,0690	1,3334	1,7396
18	0,1274	0,2571	0,3915	0,5338	0,6884	0,8620	1,0672	1,3304	1,7341
19	0,1274	0,2569	0,3912	0,5333	0,6876	0,8610	1,0655	1,3277	1,7291
20	0,1273	0,2567	0,3909	0,5329	0,6870	0,8600	1,0640	1,3253	1,7247
21	0,1272	0,2566	0,3906	0,5325	0,6864	0,8591	1,0627	1,3232	1,7207
22	0,1271	0,2564	0,3904	0,5321	0,6858	0,8583	1,0614	1,3212	1,7171
23	0,1271	0,2563	0,3902	0,5317	0,6853	0,8575	1,0603	1,3195	1,7139
24	0,1270	0,2562	0,3900	0,5314	0,6848	0,8569	1,0593	1,3178	1,7109
25	0,1269	0,2561	0,3898	0,5312	0,6844	0,8562	1,0584	1,3163	1,7081
26	0,1269	0,2560	0,3896	0,5309	0,6840	0,8557	1,0575	1,3150	1,7056
27	0,1268	0,2559	0,3894	0,5306	0,6837	0,8551	1,0567	1,3137	1,7033
28	0,1268	0,2558	0,3893	0,5304	0,6834	0,8546	1,0560	1,3125	1,7011
29	0,1268	0,2557	0,3892	0,5302	0,6830	0,8542	1,0553	1,3114	1,6991
30	0,1267	0,2556	0,3890	0,5300	0,6828	0,8538	1,0547	1,3104	1,6973
35	0,1266	0,2553	0,3885	0,5292	0,6816	0,8520	1,0520	1,3062	1,6896
40	0,1265	0,2550	0,3881	0,5286	0,6807	0,8507	1,0500	1,3031	1,6839
45	0,1264	0,2549	0,3878	0,5281	0,6800	0,8497	1,0485	1,3007	1,6794
50	0,1263	0,2547	0,3875	0,5278	0,6794	0,8489	1,0473	1,2987	1,6759
60	0,1262	0,2545	0,3872	0,5272	0,6786	0,8477	1,0455	1,2958	1,6706
70	0,1261	0,2543	0,3869	0,5268	0,6780	0,8468	1,0442	1,2938	1,6669
80	0,1261	0,2542	0,3867	0,5265	0,6776	0,8461	1,0432	1,2922	1,6641
90	0,1260	0,2541	0,3866	0,5263	0,6772	0,8456	1,0424	1,2910	1,6620
100	0,1260	0,2540	0,3864	0,5261	0,6770	0,8452	1,0418	1,2901	1,6602
120	0,1259	0,2539	0,3862	0,5258	0,6765	0,8446	1,0409	1,2886	1,6576
infinito	0,1257	0,2533	0,3853	0,5244	0,6745	0,8416	1,0364	1,2816	1,6449

p=0,10	p=0,05	p=0,04	p=0,03	p=0,02	p=0,01	p=0,002	p=0,001
6,3137	12,706	15,894	21,205	31,821	63,656	318,29	636,58
2,9200	4,3027	4,8487	5,6428	6,9645	9,9250	22,328	31,599
2,3534	3,1824	3,4819	3,8961	4,5407	5,8408	10,214	12,924
2,1318	2,7765	2,9985	3,2976	3,7469	4,6041	7,1729	8,6101
2,0150	2,5706	2,7565	3,0029	3,3649	4,0321	5,8935	6,8685
1,9432	2,4469	2,6122	2,8289	3,1427	3,7074	5,2075	5,9587
1,8946	2,3646	2,5168	2,7146	2,9979	3,4995	4,7853	5,4081
1,8595	2,3060	2,4490	2,6338	2,8965	3,3554	4,5008	5,0414
1,8331	2,2622	2,3984	2,5738	2,8214	3,2498	4,2969	4,7809
1,8125	2,2281	2,3593	2,5275	2,7638	3,1693	4,1437	4,5868
1,7959	2,2010	2,3281	2,4907	2,7181	3,1058	4,0248	4,4369
1,7823	2,1788	2,3027	2,4607	2,6810	3,0545	3,9296	4,3178
1,7709	2,1604	2,2816	2,4358	2,6503	3,0123	3,8520	4,2209
1,7613	2,1448	2,2638	2,4149	2,6245	2,9768	3,7874	4,1403
1,7531	2,1315	2,2485	2,3970	2,6025	2,9467	3,7329	4,0728
1,7459	2,1199	2,2354	2,3815	2,5835	2,9208	3,6861	4,0149
1,7396	2,1098	2,2238	2,3681	2,5669	2,8982	3,6458	3,9651
1,7341	2,1009	2,2137	2,3562	2,5524	2,8784	3,6105	3,9217
1,7291	2,0930	2,2047	2,3457	2,5395	2,8609	3,5793	3,8833
1,7247	2,0860	2,1967	2,3362	2,5280	2,8453	3,5518	3,8496
1,7207	2,0796	2,1894	2,3278	2,5176	2,8314	3,5271	3,8193
1,7171	2,0739	2,1829	2,3202	2,5083	2,8188	3,5050	3,7922
1,7139	2,0687	2,1770	2,3132	2,4999	2,8073	3,4850	3,7676
1,7109	2,0639	2,1715	2,3069	2,4922	2,7970	3,4668	3,7454
1,7081	2,0595	2,1666	2,3011	2,4851	2,7874	3,4502	3,7251
1,7056	2,0555	2,1620	2,2958	2,4786	2,7787	3,4350	3,7067
1,7033	2,0518	2,1578	2,2909	2,4727	2,7707	3,4210	3,6895
1,7011	2,0484	2,1539	2,2864	2,4671	2,7633	3,4082	3,6739
1,6991	2,0452	2,1503	2,2822	2,4620	2,7564	3,3963	3,6595
1,6973	2,0423	2,1470	2,2783	2,4573	2,7500	3,3852	3,6460
1,6896	2,0301	2,1332	2,2622	2,4377	2,7238	3,3400	3,5911
1,6839	2,0211	2,1229	2,2503	2,4233	2,7045	3,3069	3,5510
1,6794	2,0141	2,1150	2,2411	2,4121	2,6896	3,2815	3,5203
1,6759	2,0086	2,1087	2,2338	2,4033	2,6778	3,2614	3,4960
1,6706	2,0003	2,0994	2,2229	2,3901	2,6603	3,2317	3,4602
1,6669	1,9944	2,0927	2,2152	2,3808	2,6479	3,2108	3,4350
1,6641	1,9901	2,0878	2,2095	2,3739	2,6387	3,1952	3,4164
1,6620	1,9867	2,0839	2,2050	2,3685	2,6316	3,1832	3,4019
1,6602	1,9840	2,0809	2,2015	2,3642	2,6259	3,1738	3,3905
1,6576	1,9799	2,0763	2,1962	2,3578	2,6174	3,1595	3,3734
1,6449	1,9600	2,0537	2,1701	2,3264	2,5758	3,0902	3,2905

Tabela: da distribuição de Student

	K= 1		K=2	
n	di	du	di	du
6	0,61	1,4	-	-
7	0,7	1,356	0,467	1,896
8	0,763	1,332	0,559	1,777
9	0,824	1,32	0,629	1,699
10	0,879	1,32	0,697	1,641
11	0,927	1,324	0,758	1,604
12	0,971	1,331	0,812	1,579
13	1,01	1,34	0,861	1,562
14	1,045	1,35	0,905	1,551
15	1,077	1,361	0,946	1,543
16	1,106	1,371	0,982	1,539
17	1,133	1,381	1,015	1,536
18	1,158	1,391	1,046	1,535
19	1,18	1,401	1,074	1,536
20	1,201	1,411	1,1	1,537

Tabela: Durbin – watson.