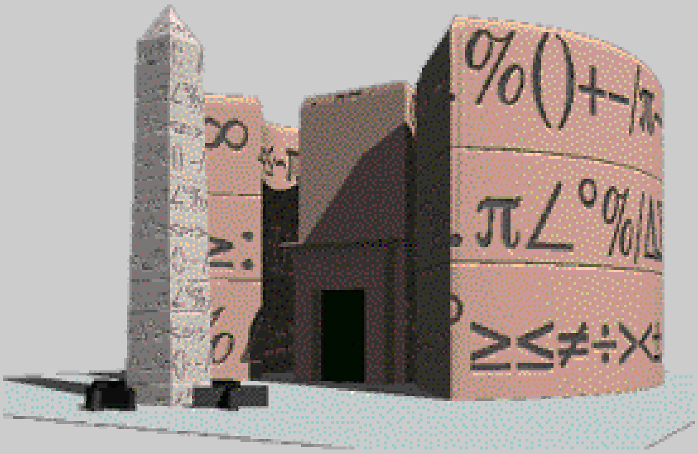
Material Didático

Série

Estatística Básica



Texto V

Correlação e Regressão

Enfoque: Exatas

Prof. Lorí Viali, Dr.

SUMÁRIO

CORRELAÇÃO E REGRESSÃO

1. CORRELAÇÃO

1.1. INTRODUÇÃO

Ao se estudar uma variável o interesse eram as medidas de tendência central, dispersão, assimetria, etc. Com duas ou mais variáveis além destas medidas individuais também é de interesse conhecer se elas tem algum relacionamento entre si, isto é, se valores altos (baixos) de uma das variáveis implicam em valores altos (ou baixos) da outra variável. Por exemplo, pode-se verificar se existe associação entre a taxa de desemprego e a taxa de criminalidade em uma grande cidade, entre verba investida em propaganda e retorno nas vendas, etc.

A associação entre duas variáveis poder ser de dois tipos: **correlacional** e **experimental**. Numa relação experimental os valores de uma das variáveis são controlados pela atribuição ao acaso do objeto sendo estudado e observando o que acontece com os valores da outra variável. Por exemplo, pode-se atribuir dosagens casuais de uma certa droga e observar a resposta do organismo; pode-se atribuir níveis de fertilizante ao acaso e observar as diferenças na produção de uma determinada cultura.

No relacionamento correlacional, por outro lado, não se tem nenhum controle sobre as variáveis sendo estudadas. Elas são observadas como ocorrem no ambiente natural, sem nenhuma interferência, isto é, as duas variáveis são aleatórias. Assim a diferença entre as duas situações é que na experimental nós atribuímos valores ao acaso de uma forma não tendenciosa e na outra a atribuição é feita pela natureza.

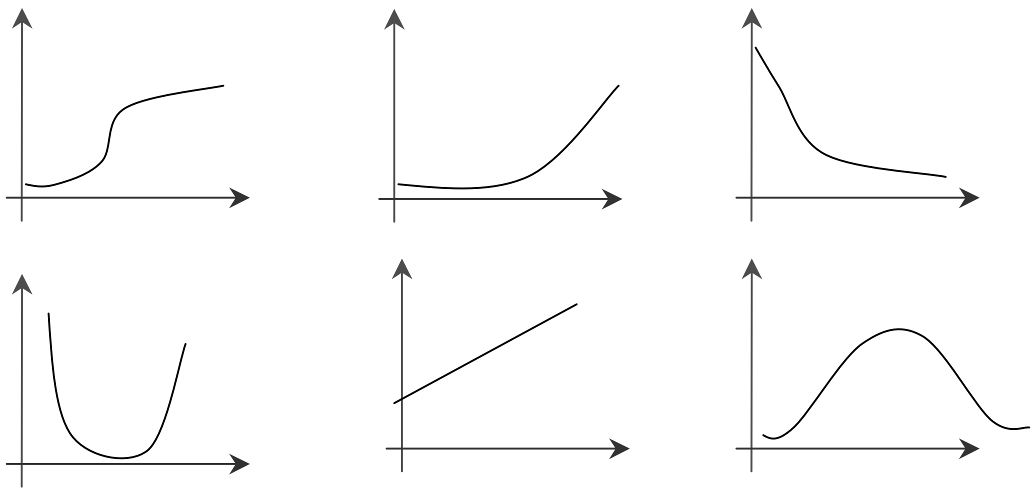


Figura 1.1 - Vários tipos de relacionamento entre as variáveis X e Y

Freqüentemente é necessário estudar o relacionamento entre duas ou mais variáveis. Ao estudo do relacionamento entre duas ou mais variáveis denominamos de **correlação** e **regressão.** Se o estudo tratar apenas de duas variáveis tem-se a correlação e a regressão simples, se envolver mais do que duas variáveis, tem-se a correlação e a regressão múltiplas. A regressão e a correlação tratam apenas do relacionamento do tipo linear entre duas variáveis.

A análise de correlação fornece um número que resume o *grau de relacionamento linear* entre as duas variáveis. Já a análise de regressão fornece uma equação que descreve o comportamento de uma das variáveis em função do comportamento da outra variável.

1.2. PADRÕES DE ASSOCIAÇÃO

Independente do tipo (correlacional ou experimental) a relação entre as variáveis pode ser resumida através de uma equação indicando o padrão de associação entre as duas variáveis. As relações mais comuns encontradas estão ilustradas na figura 1.1.

Quando não é possível perceber uma relação sistemática entre as variáveis é dito que as variáveis são **não correlacionadas**, são **independentes** ou ainda que são **ortogonais**.

1.3. INDICADORES DE ASSOCIAÇÃO

Suponha-se que queiramos determinar se duas variáveis aleatórias estão de alguma forma correlacionadas. Por exemplo, suponha-se que se queira determinar se o desempenho dos empregados no trabalho está de alguma forma associado ao escore obtido num teste vocacional.

Tabela de contingência 2x2. Uma vez que a correlação entre duas variáveis aleatórias reflete o quanto os altos escores de uma delas implicam em altos escores da outra e baixos escores de uma implicam em baixos escores da outra e vice-versa, no caso de uma relação negativa, pode-se começar a análise identificando, justamente quantos elementos de uma das variáveis são altos e quantos são baixos. Para determinar se um escore ou valor é alto ou baixo, pode-se convencionar que qualquer valor acima da mediana é alto e qualquer valor abaixo da mediana é baixo. Classificando desta forma pode-se ter então, para o exemplo, 4 possíveis resultados:

- ✓ Tanto o desempenho no trabalho quanto no teste estão acima da mediana (+ +)
- ✓ O desempenho no trabalho está acima mas o do teste está abaixo da mediana (+ –)
- ✓ Tanto o desempenho no trabalho quanto o do teste estão abaixo da mediana (– –)
- ✓ O desempenho no trabalho está abaixo da mediana mas o teste não (– +)

Estas quatro possibilidades podem ser arranjadas em uma tabela de contingência 2x2, como a mostrada abaixo:

Tabela 1.1 – Desempenho no trabalho e no teste

Desempenho no trabalho	**Escore no teste vocacional**	
	Abaixo da mediana (–)	**Acima da mediana (+)**
Acima da mediana (+)	(–, +) 10 empregados	(+, +) 40 empregados
Abaixo da mediana (–)	(–, –) 40 empregados	(+, –) 10 empregados

Observe–se que se não existir relação entre as duas variáveis deve–se esperar número idêntico de empregados em cada uma das células da tabela, isto é, se a pessoa o escore da pessoa no teste vocacional está acima ou abaixo da mediana não tem nada a ver com o seu escore no desempenho no trabalho estar acima ou abaixo da mediana.

O que pode ser visto na tabela acima é que parece existir uma forte correlação entre as duas variáveis, pois ao invés de igual número em cada célula o que se tem é um número grande de ambas as variáveis acima da mediana e um número grande de escores de ambas as variáveis abaixo da mediana. Das 50 pessoas com escore acima da mediana no teste, 40 deles (80%) apresentaram escore acima da mediana no desempenho do trabalho. Da mesma forma dos 50 que tiverem classificações abaixo da mediana, 40 deles apresentaram escore abaixo da mediana no desempenho do trabalho. Se não houvesse correlação seria de se esperar que dos 50 que tiveram escores acima da mediana no teste 25 tivessem escores acima da mediana no desempenho do trabalho e 25 abaixo.

A tabela 1.2 mostra outras possíveis saídas para este tipo de esquema de classificação cruzada. Novamente 100 elementos são classificados em 4 células de acordo com o critério anterior. A parte (a) da tabela mostra uma associação positiva, a parte (b) uma negativa e a parte (c) que não deve existir associação entre duas variáveis X e Y.

Tabela 1.2 - Indicativos da presença de associação entre duas variáveis X e Y.

(a) Relação positiva			(b) Relação negativa			(c) Sem relação		
	Valor de Y			Valor de Y			Valor de Y	
Valor de X	Abaixo da mediana	Acima da mediana	Valor de X	Abaixo da mediana	Acima da mediana	Valor de X	Abaixo da mediana	Acima da mediana
Acima da mediana	**15**	**35**	Acima da mediana	**35**	**15**	Acima da mediana	**25**	**25**
Abaixo da mediana	**35**	**15**	Abaixo da mediana	**15**	**35**	Abaixo da mediana	**25**	**25**

Diagramas de dispersão. As tabelas de contingência 2x2 fornecem somente a indicação grosseira da relação entre duas variáveis, a não ser o fato de que os valores estão situados acima e abaixo da mediana, qualquer outra informação é desperdiçada. Vamos considerar um exemplo, envolvendo duas variáveis contínuas.

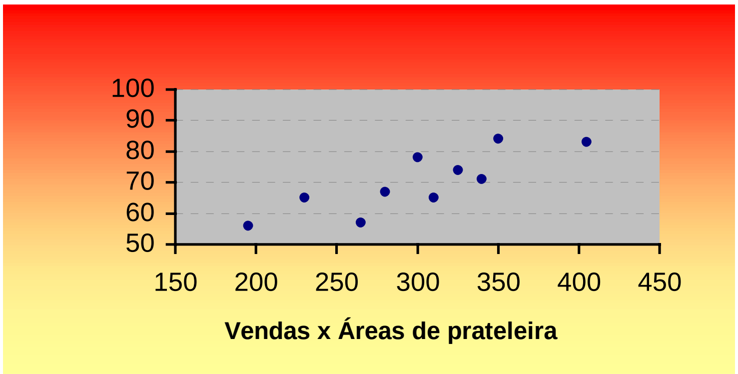
Um comerciante de temperos está curioso sobre a grande variação nas vendas de loja para loja e acha que as vendas estão associadas com o espaço nas prateleiras dedicados a sua linha de produto em cada ponto de venda. Dez lojas foram selecionadas ao acaso através do país e as duas seguintes variáveis foram mensuradas: (1) total de espaço de frente (comprimento x altura em cm^2) dedicados a sua linha de produtos e (2) total das vendas dos produtos, em reais, no último mês. Os dados são apresentados na tabela 1.3.

Tabela 1.3 – Vendas x espaço dedicado aos produtos (em cm^2).

Local	Espaço	Vendas
1	340	71
2	230	65
3	405	83
4	325	74
5	280	67
6	195	56
7	265	57
8	300	78
9	350	84
10	310	65

Pela observação da tabela não é fácil perceber o tipo de relacionamento que possa existir entre as duas variáveis. Para ter uma idéia melhor, as variáveis são colocadas no que é denominado de **diagrama de dispersão.** Uma das variáveis (X) é representada no eixo horizontal e a outra variável (Y) no eixo vertical, conforme figura 1.2.

Figura 1.2 – Diagrama de dispersão das variáveis apresentadas na tabela 1.3.



Uma olhada rápida no diagrama de dispersão mostra a existência de um relacionamento entre as variáveis, com altos valores de uma das variáveis associados a altos valores da outra variável. Se não houvesse relacionamento entre elas, os pontos estariam distribuídos ao acaso no gráfico sem mostrarem alguma tendência.

1.4. O COEFICIENTE DE CORRELAÇÃO

Apesar do diagrama de dispersão nos fornecer uma idéia do tipo e extensão do relacionamento entre duas variáveis X e Y, seria altamente desejável ter um número que medisse esta relação. Esta medida existe e é denominada de **coeficiente de correlação**. Quando se está trabalhando com amostras o coeficiente de correlação é indicado pela letra **r** que é, por sua vez, uma estimativa do coeficiente de correlação populacional: **ρ** (rho).

O coeficiente de correlação pode variar de –1,00 a + 1,00, com um coeficiente de +1, indicando uma correlação **linear** positiva perfeita. Neste caso, as duas variáveis serão exatamente iguais em termos de escores padronizados z, isto é, um elemento apresentando um escore padronizado de 1,5 em uma das variáveis vai apresentar o mesmo escore padronizado na outra variável. Um coeficiente de correlação de –1, indica correlação linear perfeita negativa, com os escores padronizados exatamente iguais em valores absolutos, diferindo apenas no sinal.

Uma correlação de +1 ou –1 é raramente observado. O mais comum é que o coeficiente fique situado no intervalo entre estes dois valores. Um coeficiente de correlação “0”, significa que não existe um relacionamento **linear** entre as duas variáveis.

1.5. HIPÓTESES BÁSICAS

A suposição básica sobre o coeficiente de correlação é que o relacionamento entre as duas variáveis seja linear. Isto é, o coeficiente de correlação é adequado para avaliar somente o relacionamento linear. As duas variáveis podem estar perfeitamente relacionadas, mas se não for de forma linear o valor do coeficiente pode ser zero ou próximo de zero.

Uma segunda hipótese é que as variáveis envolvidas sejam aleatórias e que sejam medidas no mínimo em escala de intervalo. Ele não se aplica a variáveis em escala nominal ou ordinal ou quando uma das variáveis é manipulada experimentalmente, pois neste caso, a escolha dos valores experimentais vai influenciar o valor de r obtido.

Uma terceira hipótese é que as duas variáveis tenham uma distribuição conjunta normal bivariada. Isto é equivalente a dizer que para cada x dado a variável y é normalmente distribuída.

Suponha-se que existam apenas duas variáveis X e Y. Uma amostra da variável "X", assumindo os valores particulares $X_1, X_2, ..., X_n$ e uma amostra da variável "Y" assumindo os valores particulares $Y_1, Y_2, ..., Y_n$ são obtidas e suponha-se ainda que o objetivo é saber se existe algum tipo de relacionamento linear entre estas duas variáveis. Isto poderá ser medido pelo **coeficiente de correlação** que fornece o grau de relacionamento linear entre duas variáveis.

1.6. DEFINIÇÃO

Na população o coeficiente de correlação é representado por ρ e na amostra por **r.** Assim dadas duas amostras, uma da variável X e outra da variável Y, o coeficiente de correlação amostral poderá ser calculado através da seguinte expressão:

$$\mathbf{r} = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum\left(X_i - \overline{X}\right)^2 . \sum\left(Y_i - \overline{Y}\right)^2}} = \frac{n\sum X_i \cdot Y_i - \left(\sum X_i\right).\left(\sum Y_i\right)}{\sqrt{\left[n\sum X_i^2 - \left(\sum X_i\right)^2\right].\left[n\sum Y_i^2 - \left(\sum Y_i\right)^2\right]}}$$

Uma população que tenha duas variáveis não correlacionadas linearmente pode produzir uma amostra com coeficiente de correlação diferente de zero. Para testar se a amostra foi ou não retirada de uma população de coeficiente de correlação não nulo entre duas variáveis, precisamos saber qual é a distribuição amostral da estatística **r**.

1.7. DISTRIBUIÇÃO AMOSTRAL DE R (QUANDO ρ = 0)

A distribuição amostral de **r** depende somente do valor de ρ (coeficiente de correlação populacional) e do tamanho da amostra.

Se for admitido que $\rho = 0$, a distribuição amostral de r (coeficiente de correlação na amostra) será simétrica em torno de "0" com variabilidade dada por:

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

Neste caso, pode-se mostrar que o quociente: $r / \sigma_r = r \Big/ \sqrt{\frac{1 - r^2}{n - 2}}$ tem uma distribuição **t** com **n - 2** graus de liberdade. Isto é: $t = r \Big/ \sqrt{\frac{1 - r^2}{n - 2}}$.

<u>Exemplo:</u>

Quer-se testar se existe ou não correlação linear entre X = toneladas de adubo orgânico por ha e Y = produção da cultura A por ha. Para tanto é realizado um experimento com duração de 5 anos que mostrou os resultados da tabela 1.4. Verificar se existe relacionamento linear entre as duas variáveis.

Tabela 1.4 – Valores das variáveis X e Y

Anos	X	Y
1989	2	48
1990	4	56
1991	5	64
1992	6	60
1993	8	72

Para saber se há ou não correlação linear entre estas duas variáveis na população de onde foi retirada esta amostra é necessário realizar um teste de hipóteses, ou seja, é preciso testar:

H_0: $\rho = 0$ (Não existe relacionamento linear na população)

H_1: $\rho \neq 0$ (Existe relacionamento linear na população)

A tabela 1.5 mostra os cálculos necessários para se obter o coeficiente de correlação para esta amostra das variáveis X e Y.

Tabela 1.5 – Valores das variáveis X e Y e cálculos para obter r

Anos	X	Y	XY	X^2	Y^2
1989	2	48	96	4	2304
1990	4	56	224	16	3136
1991	5	64	320	25	4096
1992	6	60	360	36	3600
1993	8	72	576	64	5184
Total	25	300	1576	145	18320

O valor de **r** será dado então por:

$$\mathbf{r} = \frac{\mathbf{n}\sum \mathbf{X_i}.\mathbf{y_i} - (\sum \mathbf{X_i}).(\sum \mathbf{Y_i})}{\sqrt{[\mathbf{n}\sum \mathbf{X_i^2} - (\sum \mathbf{X_i})^2].[\mathbf{n}\sum \mathbf{Y_i^2} - (\sum \mathbf{Y_i})^2]}} = \frac{5.1576 - 25.300}{\sqrt{(5.145 - 25^2).(5.18320 - 300^2)}} = 0,95$$

A estatística teste será:

$$\mathbf{t} = \mathbf{r} \Big/ \sqrt{\frac{1-\mathbf{r}^2}{\mathbf{n}-2}},$$

que neste caso, tem uma distribuição **t** com n - 2 = 3 graus de liberdade. O valor de **t** (calculado) é:

$$\mathbf{t} = \mathbf{r} \Big/ \sqrt{\frac{1-\mathbf{r}^2}{\mathbf{n}-2}} = 0,95 \Big/ \sqrt{\frac{1-0,95^2}{5-3}} = 5,270$$

O valor tabelado de **t** com 3 g.l. e a 5% de significância, considerando um teste bilateral é: 3,182.

Com estes valores rejeita-se H_0 e pode-se afirmar, com 5% de significância, que as duas variáveis possuem um relacionamento linear na população.

Dado que há fortes evidências de que as duas variáveis possuem um relacionamento linear pode-se então ajustar uma linha de regressão entre elas.

1.8. DISTRIBUIÇÃO AMOSTRAL DE R (QUANDO $\rho \neq 0$)

Para testar a existência de um certo grau de correlação entre duas variáveis X e Y, isto é, para testar

H_0: $\rho = \rho_0$ contra

H_1: $\rho \neq \rho_0$

$\rho > \rho_0$

$\rho < \rho_0$

é necessário determinar a distribuição de "r", quando ρ é diferente de zero. A distribuição de "r" só é simétrica quando ρ é zero, se isto não ocorre a distribuição será assimétrica. Esta falta de normalidade impede que se use o teste tradicional, o teste t, neste caso.

Contudo, mediante uma transformação apropriada, "r" pode ser alterado para uma estatística que é aproximadamente normal. Esta transformação é denominada de **transformação Z de Fischer.**

A expressão para realizá-la é: $r' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$

Esta quantidade tem distribuição aproximadamente normal com média

$\mu = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ e variância $\sigma^2 = 1 / (n - 3)$, quando "n" não for muito pequeno, ou seja, $n \geq 20$

Exemplo:

Suponha que de experiências anteriores pode ser suposto que a correlação entre a idade e a pressão sangüínea sistólica é ρ = 0.85. Para testar a hipótese nula, a 5% de significância, de que ρ é este valor contra a alternativa de que ele é diferente deste valor supõem-se que foi extraída uma amostra de tamanho n = 30 e que forneceu um r = 0,66. Então o teste pode ser realizada através dos seguintes cálculos:

Solução:

$$r' = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \frac{1}{2}\ln\left(\frac{1+0,66}{1-0,66}\right) = 0,7928$$

A distribuição de r' é dada por:

$$\mu = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = \frac{1}{2}\ln\left(\frac{1+0,85}{1-0,85}\right) = 1,2561$$

$$z = \frac{0,7928 - 1,2561}{1/\sqrt{30-3}} = -2,41$$

Para um nível de significância de 5% o valor tabelado de z é -1,96. Rejeita-se, então a hipótese nula. Isto é, pode-se afirmar que o valor da correlação populacional é diferente de 0,85.

1.9. PROPRIEDADES DE R

As propriedades mais importantes do coeficiente de correlação são:

1. O intervalo de variação vai de -1 a +1.

2. O coeficiente de correlação é uma medida adimensional, isto é, ele é independente das unidades de medida das variáveis X e Y.

3. Quanto mais próximo de +1 for "r", maior o grau de relacionamento linear positivo entre X e Y, ou seja, se X varia em uma direção Y variará na mesma direção.

4. Quanto mais próximo de -1 for "r", maior o grau de relacionamento linear negativo entre X e Y, isto é, se X varia em um sentido Y variará no sentido inverso.

5. Quanto mais próximo de zero estiver "r" menor será o relacionamento linear entre X e Y. Um valor igual a zero, indicará ausência **apenas** de relacionamento linear.

2. REGRESSÃO

Uma vez constatado que existe correlação linear entre duas variáveis, pode-se tentar prever o comportamento de uma delas em função da variação da outra.

Para tanto será suposto que existem apenas duas variáveis. A variável **X** (denominada variável controlada, explicativa ou independente) com valores observados X_1, X_2, ..., X_n e a variável **Y** (denominada variável dependente ou explicada) com valores Y_1, $Y_{2,}$ ..., Y_n. Os valores de Y são aleatórios, pois eles dependem não apenas de X, mas também de outras variáveis que não estão sendo representadas no modelo. Estas variáveis são consideradas no modelo através de um termo aleatório denominado "erro". A variável X pode ser **aleatória** ou então **controlada**.

Desta forma pode-se considerar que o modelo para o relacionamento linear entre as variáveis X e Y seja representado por uma equação do tipo:

$\mathbf{Y = \alpha + \beta X + U}$,

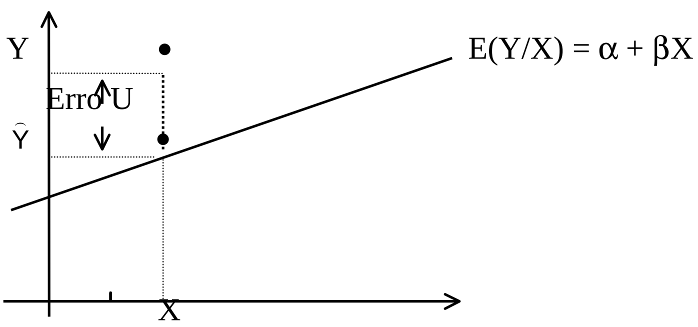
onde "U" é o termo erro, isto é, "U" representa as outras influências na variável Y além da exercida pela variável "X".

Esta equação permite que Y seja maior ou menor do que $\alpha + \beta$**X**, dependendo de "U" ser positivo ou negativo. De forma ideal o termo "U" deve ser pequeno e independente de X, de modo que se possa modificar X, sem modificar "U", e determinar o que ocorrerá, em média, a Y, isto é:

$\mathbf{E(Y/X) = \alpha + \beta X}$

Os dados $\{(X_i, Y_i), i = 1, 2, ..., n\}$ podem ser representados graficamente marcando-se cada par (X_i, Y_i) como um ponto de um plano. Os termos U_i são iguais a distância vertical entre os pontos observados (X_i, Y_i), e os pontos calculados $(X_i, \alpha + \beta X_i)$. Isto está ilustrado na figura 2.1.

Figura 2.1 – O modelo de regressão linear



Um modelo de regressão consiste em um conjunto de hipóteses sobre a distribuição dos termos "erro" e as relações entre as variáveis X e Y.

Algumas destas hipóteses são:

(i) $E(U_i) = 0$;

(ii) $Var(U_i) = \sigma^2$

Na hipótese (i) o que se está supondo é que os U_i são variáveis aleatórias independentes com valor esperado igual a zero e na (ii) que a variância de cada U_i é a mesma e igual a σ^2, para todos os valores de X.

Supõem-se ainda que a variável independente X, *permaneça fixa*, em observações sucessivas e que a variável dependente Y seja função linear de X. Os valores de Y devem ser independentes um do outro. Isto ocorre em geral, mas em alguns casos, como, por exemplo, observações diferentes são feitas no mesmo indivíduo em diferentes pontos no tempo está suposição poderá não ocorrer.

Como o valor esperado de U_i é zero, o valor esperado da variável dependente Y, para um determinado valor de X, é dado pela função de regressão $\alpha + \beta X$ ou seja:

$$E(Y/X) = E(\alpha + \beta X + U) = \alpha + \beta X + E(U) = \alpha + \beta X \qquad \textbf{[1]}$$

já que $\alpha + \beta X$ é constante para cada valor de X dado.

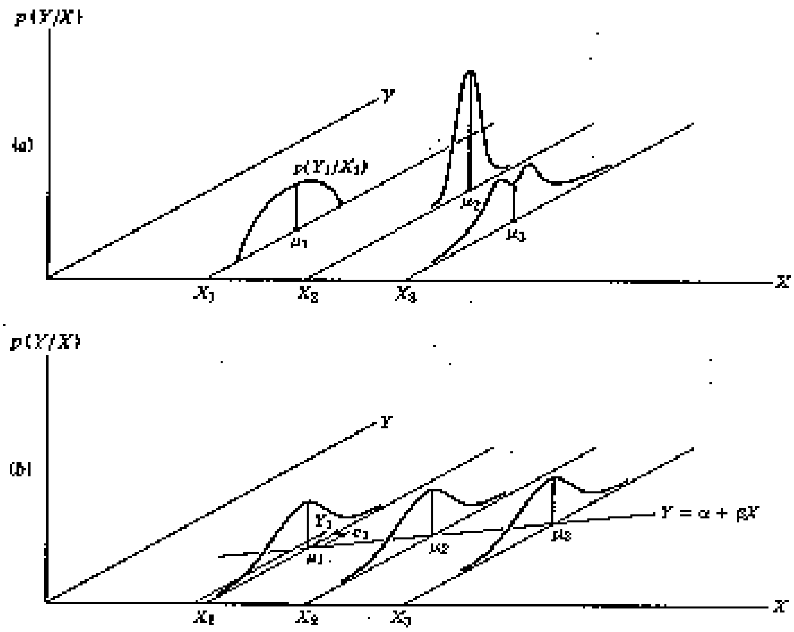
O símbolo E(Y/X) é lido **valor esperado de Y, dado X.** A variância de Y, para determinado valor de X, é igual a:

$$V(Y/X) = V(\alpha + \beta X + U) = V(U) = \sigma^2 \qquad \textbf{[2]}$$

A hipótese de que V(Y/X) é a mesma para todos os valores de X, denominada de <u>homocedasticidade</u>, é útil pois permite que se utilize cada uma das observações sobre X e Y para estimar σ^2. O termo “homo” significa “o mesmo” e “cedasticidade” significa “disperso”.

De [1] e [2] decorre que, para um dado valor de X, a variável dependente Y tem função densidade de probabilidade (condicional) com média $\alpha + \beta X$ e variância σ^2. A figura 2.2, ilustra a função densidade. Na parte superior da figura é ilustrado o caso heterocedástico e na parte inferior o caso homocedástico.

Figura 2.2 – Função densidade de Y dado X



A posição da função densidade f(Y/X) varia em função da variação do valor de X. Note-se que a média da função densidade se desloca ao longo da função de regressão $\alpha + \beta X$.

Em resumo, o modelo de regressão proposto consiste nas seguintes hipóteses:

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

2. $E(Y/X) = \alpha + \beta X$;

3. $V(Y/X) = \sigma^2$;

4. $Cov(U_i, U_j) = 0$, para $i \neq j$;

5. A variável X permanece fixa em observações sucessivas;

6. Os erros U são normalmente distribuídos.

2.1. ESTIMATIVA DOS PARÂMETROS DE REGRESSÃO

Se fosse conhecido toda a população de valores (X_i, Y_i) então seria possível determinar os valores exatos dos parâmetros α, β e σ^2. Como, em geral, se trabalha com amostras se faz necessário, então, estimar estes parâmetros com base nos valores da amostra.

Existem alguns métodos para ajustar uma linha entre as variáveis X e Y o mais utilizado é o denominado **método dos mínimos quadrados (MMQ).** A reta obtida através deste método, não é necessariamente, o "melhor" ajustamento possível, mas possui muitas propriedades estatísticas que são desejáveis.

Sejam **a** e **b** estimadores de α e β e $\mathbf{E_i} = Y_i - a - bX_i$ o desvio observado em relação a reta ajustada, isto é, $\mathbf{E_i}$ é um estimador do termo $\mathbf{U_i}$. O método dos mínimos quadrados exige que os estimadores **a** e **b** sejam escolhidos de tal forma que a soma dos quadrados dos desvios dos mesmos em relação à reta de regressão ajustada seja mínima, isto é:

$$\phi = \sum_{i=1}^{n} E_i^2 = \sum_{i=1}^{n} (Y_i - a - bX_i)^2 = \text{mínimo.}$$

Para tornar mínima esta soma em relação a **a** e **b**, é necessário diferenciar a expressão parcialmente em relação aos valores **a** e **b**. Após algumas simplificações vai-se obter:

$$\sum Y_i = na + b\sum X_i \qquad \textbf{(i)}$$

$$\sum X_i Y_i = a\sum X_i + b\sum (X_i)^2 \qquad \textbf{(ii)}$$

que são denominadas de equações normais da regressão, onde "n" é o número de pares de observações.

Obs.: Para simplificar a notação foram desconsiderados os índices nos somatórios.

Dividindo-se a equação (i) por "n" e isolando o valor de **a** vem:

$$a = \frac{\sum y_i}{n} - b\left(\frac{\sum x_i}{n}\right) = \overline{Y} - b\overline{X}$$

levando-se este resultado na equação (ii) tem-se:

$$b = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{\sum X_i Y_i - \dfrac{\sum X_i \sum Y_i}{n}}{\sum X_i^2 - \dfrac{(\sum X_i)^2}{n}} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

A reta estimada de regressão será então:

$$\boxed{\mathbf{\hat{Y} = a + bX}}$$

com os valores de "a" e "b" obtidos através das seguintes expressões:

$$b = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} \quad \text{e} \quad a = \overline{Y} - b\overline{X}$$

Utiliza-se o valor $\hat{Y}$, porque o valor de Y, obtido a partir da reta estimada de regressão, para um dado valor de X, é uma estimativa do valor **E(Y/X)**, isto é, do valor esperado de Y dado X.

Exemplo:

São fornecidos 5 pares de valores, na tabela abaixo, correspondentes as variáveis **X** e **Y**. A estimativa da reta de regressão entre **X** e **Y**, é obtida utilizando as expressões de **a** e **b** acima e usando os resultados obtidos na tabela 2.1.

Tabela 2.1 - Valores para estimar a linha de regressão

X	Y	X^2	XY
1	3	1	3
2	3	4	6
4	7	16	28
5	6	25	30
8	12	64	96
20	31	110	163

$\overline{X}$ = 20 / 5 = 4;

$\overline{Y}$ = 31/5 = 6,2

b = (5.163 - 20.31) / (5.110 - 400) = 1,30

a = $\overline{Y}$ - b$\overline{X}$ = 6,20 - 1,30.4 = 1

Então a linha estimada será: $\hat{Y}$ = 1.3X + 1

Esta reta é o "melhor" ajustamento para estes dados e seria diferente para cada amostra das variáveis X e Y, retiradas desta mesma população. Esta reta pode ser considerada uma estimativa da verdadeira linha de regressão onde 1,3 seria uma estimativa do valor β (parâmetro angular) e 1 uma estimativa do valor α (parâmetro linear), que são os verdadeiros coeficientes de regressão.

2.2. ESTIMATIVA DA VARIÂNCIA DO TERMO ERRO

O termo erro, U, é uma variável aleatória, supostamente com média zero e variância constante. Então, intuitivamente parece plausível usar os resíduos da reta de regressão pelos método dos mínimos quadrados para se estimar a variância σ^2 dos termos "erro". A variância amostral desses resíduos é igual a:

$\hat{\sigma}^2 = \frac{\sum(E-\overline{E})^2}{n}$, onde $\overline{E} = \sum E / n$. Observe-se entretanto que:

$\sum E = \sum(Y - a - bX) = \sum Y - na - b\sum X$ = 0, pela primeira equação normal (i).

Portanto, $\hat{\sigma}^2$ pode ser escrito como: $\hat{\sigma}^2 = \sum E^2 / n$.

Mas $\hat{\sigma}^2$, neste caso, é um estimador tendencioso. Pode-se obter um estimador não tendencioso, multiplicando $\hat{\sigma}^2$ por n / (n - 2). O novo estimador, não tendencioso, será representado S^2 e sua raiz quadrada:

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

é denominada de "**erro-padrão da estimativa**" ou **"erro-padrão amostral da regressão".**

Obs.: A utilização de "n - 2" é conseqüência do fato de que se deve estimar dois parâmetros, α e β, antes de obter os resíduos E. Como resultado, há somente "n - 2" graus de liberdade associados à quantidade $\sum E^2$.

A expressão acima, para o cálculo do erro amostral da regressão, apresenta o inconveniente de exigir o cálculo de cada valor previsto de Y, através da linha de regressão, tornando sua obtenção muito trabalhosa. Existe, entretanto, uma alternativa para se obter este valor (erro padrão da estimativa) sem a necessidade de calcular todos os valores previstos.

Observe-se que:

$$\sum E^2 = \sum(Y - \hat{Y})^2 \quad \sum(Y - a - bX)^2 = \sum[Y-\overline{Y}+b(\overline{X}-bX)]^2 = \sum(Y-\overline{Y})^2 - 2b\sum(X-\overline{X})(Y-\overline{Y}) + \sum b^2(\overline{X}-X)^2.$$

Fazendo:

$$\boxed{\sum(\mathbf{X}-\overline{\mathbf{X}})^2 = \sum \mathbf{X}^2 - \frac{(\sum \mathbf{X})^2}{\mathbf{n}} = \mathbf{S_{XX}}}$$

$$\boxed{\sum(\mathbf{Y}-\overline{\mathbf{Y}})^2 = \sum \mathbf{Y}^2 - \frac{(\sum \mathbf{Y})^2}{\mathbf{n}} = \mathbf{S_{YY}}}$$

$$\boxed{\sum(\mathbf{X}-\overline{\mathbf{X}})(\mathbf{Y}-\overline{\mathbf{Y}}) = \sum \mathbf{XY} - \frac{\sum \mathbf{X} \sum \mathbf{Y}}{\mathbf{n}} = \mathbf{S_{XY}}}$$

Lembrando que:

$$b = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - (\sum X_i)^2} = \frac{\sum X_i Y_i - \frac{\sum X_i \sum Y_i}{n}}{\sum X_i^2 - \frac{(\sum X_i)^2}{n}}$$, segue que $\mathbf{b = S_{XY}/S_{XX}}$ e que $\mathbf{S_{XY} = bS_{XX}}$

Então vem:

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

Assim:

$$S^2 = \frac{\sum E^2}{n-2} = \frac{\sum(Y-a-bX)^2}{n-2} = \frac{S_{YY} - b^2 S_{XX}}{n-2} = \frac{S_{YY} - b\,S_{XY}}{n-2}$$

Pode-se verificar que S^2 definido desta maneira é um estimador não-tendencioso de σ^2, isto é, $E(S^2) = \sigma^2$.

O erro padrão da regressão será dado, então, por:

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

<u>Exemplo:</u>

Considerando as variáveis X e Y acima e a linha de regressão anterior determinar uma estimativa do erro padrão da regressão.

Os cálculos necessários estão na tabela 2.2.

Tabela 2.2 – Determinação do erro padrão da regressão

X	Y	Y_c	$E = Y - Y_c$	E^2
1	3	2,3	0,7	0,49
2	3	3,6	-0,6	0,36
4	7	6,2	0,8	0,64
5	6	7,5	-1,5	2,25
8	12	11,40	0,6	0,36
20	31	31	0	4,10

O erro padrão da regressão será então:

$$S = \sqrt{\frac{\sum E^2}{n-2}} = \sqrt{\frac{\sum (Y - a - bX)^2}{n-2}} = \sqrt{\frac{4,10}{5-3}} = \sqrt{1,3667} = 1,17$$

Este mesmo cálculo poderá ser efetuado pela expressão definida acima, sem a necessidade de se obter os valores estimados.

Tabela 2.3 – Determinação do erro padrão da regressão

X	Y	X^2	Y^2	XY
1	3	1	9	3
2	3	4	9	6
4	7	16	49	28
5	6	25	36	30
8	12	64	144	96
20	31	110	247	163

Neste caso, tem-se:

$$S_{XX} = \sum X^2 - \frac{(\sum X)^2}{n} = 110 - 20^2/5 = 30$$

$$S_{YY} = \sum Y^2 - \frac{(\sum Y)^2}{n} = 247 - 31^2/5 = 54,80$$

$$S_{XY} = \sum XY - \frac{\sum X \sum Y}{n} = 163 - (20.31)/5 = 39$$

O valor de "b" será:

$b = S_{XY}/S_{XX} = 39/30 = 1,30$

Portanto o erro padrão da regressão será:

$$s = \sqrt{\frac{S_{YY} - b^2 S_{XX}}{n-2}} = \sqrt{\frac{S_{YY} - bS_{XY}}{n-2}} = \sqrt{\frac{54,80 - 1,3.39}{5-2}} = \sqrt{\frac{4,10}{3}} = \sqrt{1,3667} = 1,1690 = 1,17$$

2.3. DISTRIBUIÇÕES DAS ESTIMATIVAS

Observando-se as expressões dos estimadores "a" e "b" da reta estimada, pode-se notar que ambos dependem de Y que é uma variável aleatória com distribuição supostamente normal de média f(X) e desvio padrão σ^2. Como os estimadores "a" e "b" são funções lineares de uma variável aleatória normal, também serão variáveis aleatórias com distribuição normal. O que precisa ser determinado, então, é a média e a variância de cada um deles. Antes disso vai-se determinar uma estimativa de σ^2 a variância da variável Y, que no modelo é suposta a mesma para cada valor de X (homocedasticidade).

2.3.1. DISTRIBUIÇÃO DO ESTIMADOR "B"

Tem-se que:

$b = S_{XY} / S_{XX} = \dfrac{\sum(X-\bar{X})(Y-\bar{Y})}{S_{XX}} = \dfrac{\sum Y(X-\bar{X}) - \sum \bar{Y}(X-\bar{X})}{S_{XX}}$ Mas $\sum(X-\bar{X}) = 0$, logo:

$$b = \frac{\sum Y(X-\bar{X})}{S_{XX}}$$

Mas $Y = \alpha + \beta X + U$, então:

$$b = \frac{\sum Y(X-\bar{X})}{S_{XX}} = \frac{\sum(\alpha + \beta X + U)(X-\bar{X})}{S_{XX}} = \frac{\alpha\sum(X-\bar{X})}{S_{XX}} + \frac{\beta\sum X(X-\bar{X})}{S_{XX}} + \frac{\sum U(X-\bar{X})}{S_{XX}}$$

Como $S_{XX} = \sum(X-\bar{X})^2 = \sum(X-\bar{X})(X-\bar{X}) = \sum X(X-\bar{X}) - \bar{X}\sum(X-\bar{X}) = \sum X(X-\bar{X})$, pois $\sum(X-\bar{X}) = 0$

Vem: $b = \beta + \dfrac{\sum U(X-\bar{X})}{S_{XX}}$

Logo a expectância de "b" será:

$E(b) = E(\beta) + E\left(\dfrac{\sum U(X-\bar{X})}{S_{XX}}\right) = E(\beta) + \dfrac{\sum(X-\bar{X})}{S_{XX}}E(U)$. Mas $E(U) = 0$, por hipótese.

Então:

$E(b) = E(\beta) = \beta$, uma vez que a média de uma constante é a própria constante.

Isto, também, mostra que "b" é um estimador não-tendencioso de β.

Para a variância, tem-se:

$$V(b) = V\left(\beta + \frac{\sum U(X-\bar{X})}{S_{XX}}\right) = V\left(\frac{\sum U(X-\bar{X})}{S_{XX}}\right) = \frac{\sum(X-\bar{X})^2}{(S_{XX})^2}V(U).$$

Tendo em vista que por hipótese do modelo $V(U) = \sigma^2$ e que $\sum(X - \overline{X})^2 = S_{XX}$, segue:

$V(b) = \dfrac{S_{XX} \cdot \sigma^2}{(S_{XX})^2} = \dfrac{\sigma^2}{S_{XX}}$. Portanto, a distribuição da estatística "b" é $N(\beta, \dfrac{\sigma}{\sqrt{S_{XX}}})$.

2.3.2. DISTRIBUIÇÃO DO ESTIMADOR "A"

Quanto à distribuição da variável aleatória "a", tem-se:

$a = \overline{Y} - b\overline{X}$. Mas $\overline{Y} = \sum Y / n$, então:

$$a = \frac{\sum Y}{n} - b\overline{X} = \frac{\sum(\alpha + \beta X + U)}{n} - b\overline{X} = \frac{\sum \alpha}{n} + \beta\frac{\sum X}{n} + \frac{\sum U}{n} - b\overline{X} = \alpha + \beta\overline{X} + \frac{\sum U}{n} - b\overline{X}$$

Assim:

$$E(a) = E(\alpha) + E(\beta\overline{X}) + E(\frac{\sum U}{n}) - E(b\overline{X}) = \alpha + \beta\overline{X} + \frac{\sum E(U)}{n} - \beta\overline{X}, \text{ pois } E(b) = \beta$$

Então $E(a) = \alpha$, pois $E(U) = 0$. Vê-se que "a" é um estimador não-tendencioso de α.

Quanto à variância, tem-se:

$$V(a) = V(\alpha) + V(\beta\overline{X}) + V(\frac{\sum U}{n}) + V(b\overline{X}) = 0 + 0 + \frac{1}{n^2}\sum V(U) + \overline{X}^2\, V(b) = \frac{1}{n^2}\sum \sigma^2 + \overline{X}^2\frac{\sigma^2}{S_{XX}} = =$$

$$\frac{\sigma^2}{n} + \frac{\sigma^2\overline{X}^2}{S_{XX}} = \sigma^2(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}).$$

Portanto a distribuição de "a" é: $N(\alpha,\ \sigma\sqrt{\dfrac{1}{n} + \dfrac{\overline{X}^2}{S_{XX}}})$.

2.4. DECOMPOSIÇÃO DA SOMA DOS QUADRADOS

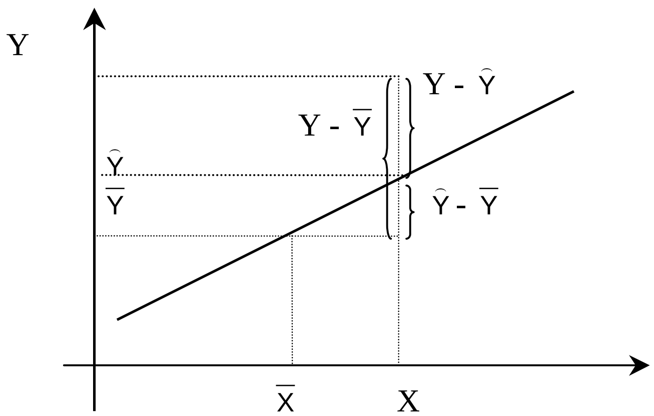


Figura 2.3 – Desvios na regressão

2.4.1. DECOMPOSIÇÃO DOS DESVIOS

Pelo figura 2.3, pode-se perceber que o desvio em relação a Y (desvio total), isto é, $Y - \overline{Y}$ pode ser decomposto em dois outros desvios:

- O desvio explicado pela linha de regressão, isto é, $\hat{Y} - \overline{Y}$ e
- O desvio não-explicado (resíduos) pela linha de regressão, isto é, $Y - \hat{Y}$.

É fácil perceber que a variação total, $\sum(Y - \overline{Y})$, é a soma da variação explicada, $\sum(\hat{Y} - \overline{Y})$, e a não-explicada, $\sum(Y - \hat{Y})$, pois:

$Y - \overline{Y} = Y - \hat{Y} + \hat{Y} - \overline{Y}$, então:

Aplicando somatório a ambos os membros vem:

$$\sum(Y - \overline{Y}) = \sum(Y - \hat{Y}) + \sum(\hat{Y} - \overline{Y})$$

Pode-se verificar também que a propriedade aditiva dos desvios é extensiva à soma dos quadrados desses desvios, ou seja:

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

De fato:

$$\sum(Y - \overline{Y})^2 = \sum(Y - \hat{Y} + \hat{Y} - \overline{Y})^2 = \sum[(Y - \hat{Y}) + (\hat{Y} - \overline{Y})]^2 = \sum(Y - \hat{Y})^2 + \sum(\hat{Y} - \overline{Y})^2 - 2\sum(Y - \hat{Y})(\hat{Y} - \overline{Y})$$

Mas

$$\sum(Y - \hat{Y})(\hat{Y} - \overline{Y}) = \sum(Y - \hat{Y})(a + bX - a - b\overline{X}) = b\sum X(Y - \hat{Y}) - b\overline{X}\sum X(Y - \hat{Y})$$

Pelas condições do método dos mínimos quadrados, tem-se:

$\sum(\hat{Y} - \overline{Y}) = 0$ e $\sum X(Y - \hat{Y}) = 0$, em conseqüência

$\sum(Y - \hat{Y})(\hat{Y} - \overline{Y}) = 0$, logo, segue que:

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

isto é, que a soma dos quadrados dos desvios calculados em torno da média de Y (variação total = VT) é igual à soma dos quadrados dos desvios em torno da linha de regressão (variação residual = VR) mais a soma dos quadrados dos desvios da linha de regressão em torno da média (variação explicada = VE).

2.4.2. CÁLCULO DAS VARIAÇÕES

(a) Variação Total: VT ou s_Y^2

$VT = \sum(Y - \overline{Y})^2 = S_{YY}$, onde $S_{YY} = \sum Y^2 - (\sum Y)^2 / n$

(b) Variação Explicada: VE ou $s_{\hat{Y}}^2$

$$VE = \sum(\hat{Y} - \overline{Y})^2 = \sum(a + bX - \overline{Y})^2 = \sum(\overline{Y} - b\overline{X} + bX - \overline{Y})^2 = \sum[(b(X - \overline{X})]^2 = b^2\sum(X - \overline{X})^2 = b^2S_{XX}$$

Logo:

$$VE = b^2S_{XX} \text{ ou } VE = \left(\frac{S_{XY}}{S_{XX}}\right)^2 S_{XX} = bS_{XY}$$

(c) Variação Residual: VR ou $s^2_{Y/X}$

De acordo com a propriedade aditiva das variações, pode-se calcular VR por diferença. Assim:

$VR = \sum(Y - \hat{Y})^2 = VT - VE$ ou $VR = S_{YY} - bS_{XY}$

2.5. INTERVALOS DE CONFIANÇA

Da mesma forma que foram obtidos intervalos de confiança para a média, variância e proporção de uma população, pode-se determinar os intervalos de confiança para os parâmetros da regressão. Ou seja, pode-se determinar um intervalo de confiança para o coeficiente linear (α), um intervalo de confiança para o parâmetro angular (β) e pode-se ainda determinar um intervalo de confiança para um valor previsto de Y, dado X. Este intervalo pode ser para o valor médio de Y para um dado X, isto é, E(Y/X) ou, então, para um valor individual de Y, isto é, $\hat{Y}$. A estimativa pontual para os dois últimos casos é a mesma. O que vai mudar é o intervalo de confiança correspondente. Isto se deve ao fato de que o modelo desenvolvido é associado principalmente à média do grupo do que a uma informação individual.

2.5.1. INTERVALO PARA O COEFICIENTE LINEAR (α)

Considerando que a distribuição do coeficiente linear é dado por $N(\alpha, \sigma\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}})$. Então, fixada uma confiança de 1 - α, o intervalo será:

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

com t_{n-2} sendo um valor da distribuição “t” com “n - 2” graus de liberdade e S uma estimativa de σ.

2.5.2. INTERVALO PARA O COEFICIENTE ANGULAR (β)

Considerando que a distribuição do coeficiente angular é dado por $N(b, \frac{\sigma}{\sqrt{S_{XX}}})$. Então, fixada uma confiança de 1 - α, o intervalo será:

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

com t_{n-2} sendo um valor da distribuição “t” com “n - 2” graus de liberdade e S uma estimativa de σ.

2.5.3. INTERVALO PARA PREVISÕES

(a) Intervalo para o valor médio de $\hat{Y}$

Tem-se que $\hat{Y} = a + bX$ é um estimador de E(Y/X) ou f(X). Para construir um intervalo de confiança para este valor é necessário conhecer a sua distribuição. Isto é, deve-se conhecer a média e a variância de $\hat{Y}$.

$E(\hat{Y}) = E(a + bX) = E(a) + E(bX) = \alpha + \beta E(X) = \alpha + \beta X = f(X) = E(Y/X)$, pois, neste caso, X é constante para cada valor de Y.

Tem-se: $\hat{Y} = a + bX$, mas $a = \bar{Y} - b\bar{X}$, então:

$\hat{Y} = \bar{Y} - b\bar{X} + bX = \bar{Y} + b(X - \bar{X})$. A variância de $\hat{Y}$, será:

$$V(\hat{Y}) = V[\bar{Y} - b(X - \bar{X})] = V(\bar{Y}) + V[b(X - \bar{X})] = V\left(\frac{\sum Y}{n}\right) + (X - \bar{X})^2\, V(b) = \frac{1}{n^2}\sum V(Y) + (X - \bar{X})^2 \frac{\sigma^2}{S_{XX}} = \frac{\sigma^2}{n} + (X - \bar{X})^2 \frac{\sigma^2}{S_{XX}} = \sigma^2\left[\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}\right].$$

Portanto:

$\hat{Y}$ tem distribuição $N\left(\alpha + \beta X,\ \sigma\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}\right)$

Conhecida a distribuição de $\hat{Y}$, então o intervalo de confiança de "$1 - \alpha$" de probabilidade para f(X) ou E(Y/X) será:

$$P\left(\hat{Y} - t_{n-2}.\ S.\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}} \le E(Y/x) \le \hat{Y} + t_{n-2}.\ S.\sqrt{\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}\right) = 1 - \alpha,$$ onde t_{n-2} é o valor da distribuição t com "n - 2" graus de liberdade.

(b) Intervalo para um valor individual ($\hat{Y}$)

Uma estimativa do valor individual de Y é dado pela reta de regressão $\hat{Y} = a + bX$, para um dado X e o desvio de previsão será dado por $Y - \hat{Y}$, cujas propriedades são:

Para a média:

$E(Y - \hat{Y}) = E(Y) - E(\hat{Y}) = f(X) - f(X) = 0$

Para a variância, tem-se:

$$V(Y - \hat{Y}) = V(Y) + V(\hat{Y}) = \sigma^2 + \sigma^2\left[\frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}\right] = \sigma^2\left[1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}\right].$$

Então:

$Y - \hat{Y}$ tem distribuição $N\left(0,\ \sigma\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}}\right)$

Conhecida a distribuição de $Y_i - \hat{Y}$, então o intervalo de confiança de "$1 - \alpha$" de probabilidade para um valor individual de Y (Y_i) para um dado X, será:

$$\hat{Y} - t_{n-2}.\ S.\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}};\quad \hat{Y} + t_{n-2}.\ S.\sqrt{1 + \frac{1}{n} + \frac{(X - \bar{X})^2}{S_{XX}}},$$ onde t_{n-2} é o valor da distribuição t com "n - 2" graus de liberdade.

2.6. TESTES DE HIPÓTESES

Conhecidas as distribuições dos estimadores dos coeficientes angular e linear, pode-se realizar um teste de hipóteses.

2.6.1. TESTE PARA A EXISTÊNCIA DA REGRESSÃO

Testar a existência da regressão é testar se o parâmetro β é diferente de zero. Desta forma o que se quer testar é:

H_0: $\beta = 0$ contra as alternativas:

H_1: $\beta \neq 0$;

$\beta > 0$ ou

$\beta < 0$

Fixado um nível de significância α a variável teste será a "t" de Student com "n - 2" graus de liberdade, pois sabe-se que:

b tem distribuição Normal com média β e desvio padrão $\frac{\sigma}{\sqrt{S_{XX}}}$, ou seja,

$Z = \dfrac{b-\beta}{\frac{\sigma}{\sqrt{S_{XX}}}}$ tem distribuição normal padrão. Porém como σ não é conhecido é necessário estimá-lo através de S. Então:

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

2.6.2. TESTE PARA O COEFICIENTE LINEAR

Testar o coeficiente linear da regressão "α" é testar o valor inicial da regressão, isto é, é testar o valor de Y quando X = 0. As hipóteses são:

H_0: $\alpha = 0$ contra as alternativas:

H_1: $\alpha \neq 0$;

$\alpha > 0$ ou

$\alpha < 0$

Fixado um nível de significância a variável teste será a "t" de Student com "n - 2" graus de liberdade, pois sabe-se que o estimador "a", tem uma distribuição:

$N(\alpha, \sigma^2(\frac{1}{n} + \frac{\overline{X}^2}{S_{XX}}))$. Então:

$Z = \dfrac{a-\alpha}{\sigma\left(\sqrt{\dfrac{1}{n}+\dfrac{\overline{X}^2}{S_{XX}}}\right)}$ tem distribuição normal padrão. Porém como σ não é conhecido é necessário estimá-lo através de S. Então: $t_{n-2} = \dfrac{a-\alpha}{S\left(\sqrt{\dfrac{1}{n}+\dfrac{\overline{X}^2}{S_{XX}}}\right)}$

2.7. COEFICIENTE DE DETERMINAÇÃO OU DE EXPLICAÇÃO

Além dos testes de hipóteses e dos intervalos de confiança, outro indicador que fornece elementos para a análise do modelo adotado é o coeficiente de determinação ou de explicação, definido por:

$R^2 = VE / VT = \dfrac{b S_{XY}}{S_{YY}}$

O coeficiente de determinação indica quantos por cento a variação explicada pela regressão representa sobre a variação total. Deve-se ter:

$0 \le R^2 \le 1$

Se R^2 for igual a 1, isto significa que todos os pontos observados se situam "exatamente" sobre a reta de regressão. Tendo-se, neste caso, um ajuste perfeito. As variações da variável Y são 100% explicadas pelas variações da variável X, não ocorrendo desvios em torno da função estimada.

Por outro lado, se $R^2 = 0$, isto quer dizer que as variações de Y são exclusivamente aleatórias e explicadas pelas variações de outros fatores que não X.

3. EXERCÍCIOS

(01) Para cada uma das situações abaixo, diga o que é mais adequado: a análise de regressão ou a análise de correlação. Por quê?

(01.1) Uma equipe de pesquisadores deseja determinar se o rendimento na Universidade sugere êxito na profissão escolhida.

(01.2) Deseja-se estimar o número de quilômetros que um pneu radial pode rodar antes de ser substituído.

(01.3) Deseja-se prever quanto tempo será necessário para executar uma determinada tarefa por uma pessoa, com base no tempo de treinamento.

(01.4) Deseja-se verificar se o tempo de treinamento é importante para avaliar o desempenho na execução de uma dada tarefa.

(01.5) Um gerente deseja estimar as vendas semanais com base nas vendas das segundas e terças-feiras.

(02) Suponha que uma cadeia de supermercados tenha financiado um estudos dos gastos com mercadorias para famílias de 4 pessoas. O estudo se limitou a famílias com renda líquida entre 8 e 20 salários mínimos. Obteve-se a seguinte equação:

$\hat{Y}$ = -1,20 + 0,40X, onde $\hat{Y}$ = despesa mensal estimada com mercadorias e X = renda líquida mensal.

(02.1) Estimar a despesa de uma família com renda mensal líquida de 15 s.m.

(02.2) Um dois diretores da empresa ficou intrigado com o fato de que a equação sugerir que uma família com renda de 3 s.m. líquidos mensais não gaste nada em mercadorias. Qual a explicação?

(02.3) Explique por que a equação acima não poderia ser utilizada para estimar

(a) As despesas com mercadorias de famílias de 5 pessoas.

(b) As despesas com mercadorias de famílias com renda de 20 a 40 s.m. líquidos mensais.

(03) Utilize os valores abaixo para estimar as equações de regressão:

(03.1) $\sum X = 200$, $\sum Y = 300$, $\sum XY = 6200$, $\sum X^2 = 3600$ e $n = 20$

(03.2) $\sum X = 7{,}2$, $\sum Y = 37$, $\sum XY = 3100$, $\sum X^2 = 620$ e $n = 36$

(04) Para cada uma das situações abaixo, grafe os valores em um diagrama e se uma equação linear parecer apropriada para explicar os dados, determine os seus parâmetros.

(04.1)

Tamanho do pedido(X)	25	20	40	45	22	63	70	60	55	50	30
Custo Total (Y)	2000	3500	1000	800	3000	1300	1500	1100	950	900	1600

(04.2)

Vendas em mil (X)	201	225	305	380	560	600	685	735	510	725	450	370	150
Lucro em mil (Y)	17	20	21	23	25	24	27	27	22	30	21	19	15

(05) Suponha que uma população se constitua dos seis pontos seguintes:

(1, 2), (4, 6), (2, 4), (2, 3), (3, 5) e (5, 10)

(05.1) Grafe os pontos em um diagrama de dispersão.

(05.2) Determine a equação de regressão: $Y = \alpha + \beta X + u$.

(05.3) Os termos-erro verificam a condição $E(u) = 0$?

(05.4) Selecione uma amostra de tamanho n = 4, da população acima e estime a equação de regressão determinada no item 5.2. Grafe o resultado no mesmo diagrama construído em 5.1.

(06) Verifique que a reta de regressão $\hat{Y} = a + bX$, sempre passa pelo ponto $(\bar{X}, \bar{Y})$.

(07) Os dados abaixo forma colhidos de cinco fábricas diferentes de uma determinada indústria:

Custo total (Y)	80	44	51	70	61
Produção (X)	12	4	6	11	8

(07.1) Estime uma função linear da forma $\hat{Y} = a + bX$ para o custo total dessa indústria.

(07.2) Qual o significado econômico das estimativas "a" e "b"?

(07.3) Teste a hipótese de que o custo fixo da produção do artigo em questão seja igual a 5, contra a alternativa de diferente do que 5, utilizando uma significância de 5%.

(08) Em uma amostra aleatória de 1990, 50 homens americanos entre 35 e 54 anos de idade acusaram a seguinte relação entre renda anual Y (em dólares) e a escolaridade X (em anos). $\hat{Y} = 1200 + 800X$. A renda média foi de 10000 dólares e a escolaridade média foi de 11,0 anos. Sabendo, ainda, que $\sum X^2 = 9000$ e que o desvio padrão residual em relação à reta ajustada foi de 7300 dólares, determine:

(08.1) A renda de uma pessoa que tenha completado 2 anos de educação secundária (x = 10 anos).

(08.2) O intervalo de 95% de confiança para o coeficiente angular populacional..

(08.3) Se a renda para a escolaridade é estatisticamente discernível ao nível de 5%.

(08.4) Se é válida a afirmação que cada ano de escolaridade custa 800 dólares?

(09) Uma pesquisa foi realizada com o objetivo de determinar os efeitos da falta de sono sobre a capacidade de as pessoas resolverem problemas simples. Foram testadas 10 pessoas, mantendo-se cada grupo de 2 pessoas sem dormir por um determinado número de horas. Após cada um destes períodos, cada pessoa teve de resolver um teste com adições simples, anotando-se então os erros cometidos. Os dados resultantes estão na tabela abaixo:

Número de erros (Y)	6, 8	6, 10	8, 14	12, 14	12, 16
Número de horas sem dormir (X)	8	12	16	20	24

(9.1) Determine a estimativa da linha de regressão do número de erros em função do número de horas sem dormir.

(9.2) Determine a dispersão dos termos erro em torno da linha de regressão.

(10) Determine um intervalo de 95% de confiança para o coeficiente angular da reta do exercício acima. Interprete o intervalo obtido.

(11) Realizou-se uma pesquisa de mercado com o objetivo de estudar a relação entre o tempo necessário para um consumidor tomar uma decisão (sobre o que comprar) e o número de embalagens alternativas do mesmo produto apresentadas a esse consumidor. Eliminaram-se as marcas das embalagens, a fim de reduzir o efeito da preferência por uma ou outra marca. Os consumidores fizeram suas escolhas somente com base na descrição do produto, anotada nas embalagens pelos fabricantes. O tempo necessário, Y, para que cada um tomasse sua decisão foi anotado para 15 participantes, resultando nos seguintes dados:

Tempo para decisão, Y (em segundos)	5, 7, 8, 8, 9	7, 8, 9, 9, 10	9, 10, 10, 11, 12
Número de alternativas (X)	2	3	4

(11.1) Determine a reta dos mínimos quadrados de Y em função de X.

(11.2) Determine o erro padrão da estimativa, ou seja, o desvio padrão amostral da regressão.

(11.3) Há evidência suficiente nestes dados de que o tempo de decisão se relaciona linearmente ao número de alternativas oferecidas a esses consumidores?

(12) Na fabricação de um antibiótico, a produção depende do tempo. Os dados indicados na tabela, mostram que um processo resultou na seguinte produção (em quilogramas) de antibióticos por período de tempo (dias) indicados:

Tempo (X = dias)	1	2	3	4	5	6
Produção (Y = em kg.)	23	31	40	46	52	63

(12.1) Por várias razões é conveniente esquematizar a produção em ciclos de 4 dias. Estime o valor médio da produção final de antibiótico produzido em um período de 4 dias. Considere um intervalo de 95% de confiança.

(12.2) Suponha que o processo de produção, no futuro, se desenvolverá em 4 dias. Determine um intervalo de previsão de 95% para a produção. Compare com o intervalo para a produção média de um período de 4 dias que foi obtido em (12.1).

(13) Mediu-se a altura de uma amostra de 5 meninos (em polegadas) na idade de 4 anos e novamente na idade de 18 anos. Os resultados obtidos estão abaixo:

Na idade de 4 anos	40	43	40	40	42
Na idade de 18 anos	68	74	70	68	70

(13.1) Determine o coeficiente de correlação entre as duas categorias de alturas.

(13.2) Teste a hipótese de que existe uma relação linear entre a altura aos 4 anos de idade e a altura aos 18 anos de idade.

(13.3) Se fosse feito o gráfico de toda a população de alturas, calculando-se a correspondente reta dos mínimos quadrados, qual seria o seu coeficiente angular? Responda com um intervalo suficientemente amplo que permita uma aposta de 95%.

(13.4) Repita o item 13.3 só que para o coeficiente linear.

(14) A equação de regressão estimada abaixo resume um estudo da relação entre o uso do fumo e a incidência de câncer pulmonar, relacionando o número X de anos que uma pessoa fumou com a percentagem Y de incidência de câncer pulmonar em cada grupo.

$\hat{Y} = -2 + 1{,}70.X$ e $r = 0{,}60$.

(14.1) Explique o significado das estimativas “-2” e “ 1,70” na equação de regressão.

(14.2) Qual a taxa de incidência de câncer pulmonar para as pessoas que fumam há 20 anos?

(14.3) Se “r” fosse igual a “um” seria possível concluir que o fumo é a única causa de câncer pulmonar?

(14.4) Suponha-se que a equação estimada tenha sido obtida de uma amostra aleatória de 50 fumantes. Teste a hipótese de que o coeficiente de correlação seja igual a zero a uma significância de 1%.

(15) Explique se concorda ou não com as seguintes afirmativas:

(15.1) Um coeficiente de correlação de +1,0 entre duas variáveis X e Y indica que X causa Y, mas um coeficiente de correlação de -1,0 significa que X não causa Y.

(15.2) Se o coeficiente de regressão é zero, o coeficiente de correlação é também zero.

(15.3) Se o coeficiente angular é 1 (um), isto significa que existe perfeita correlação entre X e Y.

(15.4) É possível que o coeficiente de correlação amostral seja positivo, quando não existe, de fato, nenhuma correlação entre as variáveis X e Y.

(15.5) Não se pode utilizar a técnica da regressão pelo método dos mínimos quadrados quando a relação básica entre X e Y não for linear.

(16) Um estudo de duas safras forneceu as seguintes informações:

Safra A: $\hat{Y} = 200 + 0{,}8X$, $r = 0{,}70$ e $S = 30$ Safra B: $\hat{Y} = 50 + 1{,}20X$, $r = 0{,}9$ e $S = 20$, onde Y é a produção por alqueire e X é a quantidade de chuva (em polegadas) no período da safra.

(16.1) Se não houvesse chuva, estas duas equações poderiam ser usadas para predizer a quantidade produzida nas duas safras? Por quê?

(16.2) Qual das duas safras tira mais proveito do aumento das chuvas? Por quê?

(16.3) Para qual das duas safras é possível predizer a produção com melhor aproximação? Por quê?

(17) Os dados abaixo foram obtidos de cinco fábricas diferentes de uma determinada indústria.

Custo total (Y = em milhões)	80	44	51	70	61
Produção (X = toneladas)	12	4	6	11	8

(17.1) Determine um intervalo de confiança de 90% para o custo fixo dessa indústria.

(17.2) Determine um intervalo de confiança de 95% para o custo marginal dessa indústria.

(17.3) Faça uma previsão, através de um intervalo, para o custo total médio dessa indústria, para uma produção de 15t, utilizando uma confiança de 95%.

(17.4) Faça uma previsão, através de um intervalo, para o custo total dessa indústria, para uma produção de 15t, utilizando uma confiança de 95%.

(17.5) é possível afirmar, com uma significância de 1%, que o custo total dessa indústria está linearmente relacionado ao nível de produção?

(17.6) Testar se o custo fixo pode ser considerado menor do que 30.

(17.7) Testar se o custo marginal pode ser considerado menor do que 5.

(18) Qual é o tamanho mínimo da amostra necessária para que se possa concluir que um coeficiente de correlação de 0,32 difere significativamente de zero ao nível de 0,05?

(19) Um coeficiente de correlação, baseado em uma amostra de tamanho n = 18, foi calculado como sendo 0,32. Pode-se concluir aos níveis de significância (19.1) 0,05 e (19.2) 0,01, que o coeficiente de correlação, correspondente na população é diferente de zero?

(20) Se o coeficiente de correlação entre X e Y é 0,80, que percentagem da variação total permanece não-explicada pela equação de regressão?

(21) Examine os cinco pares de pontos dados na tabela

X	-2	-1	0	1	2
Y	4	1	0	1	4

(21.1) Qual é a relação matemática entre X e Y?

(21.2) Determine o valor de r.

(21.3) Mostre que calculando-se a linha de regressão de Y em relação a X tem-se b = 0.

(21.4) Por que, aparentemente, não existe relação entre X e Y como estão indicando b e r?

(22) Os dados abaixo representam o número de rendas pessoais tributáveis e o registro de automóveis de passageiros, em uma determinada região.

X = número de rendas tributáveis (em milhares)	192	80	162	246	310
Y = Número de carros de passageiros (milhares)	23	11	13	31	91

(22.1) Verificar se existe correlação entre as duas variáveis.

(22.2) Determine a equação de regressão de Y em função de X, caso o coeficiente de correlação seja significativamente diferente de zero.

(22.3) Faça uma previsão do número de carros se o número de contribuintes tributáveis for de 500 mil.

(22.4) Determine a equação de regressão de X em função de Y.

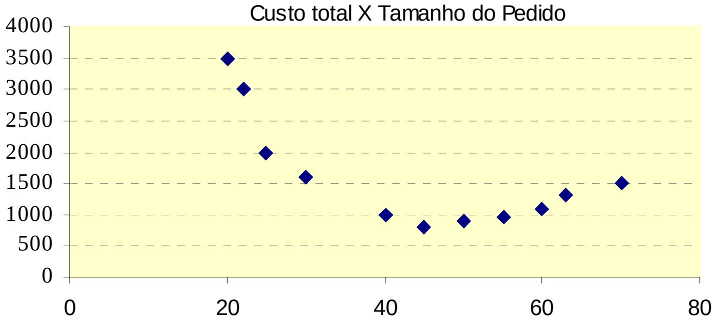
4. RESPOSTAS

(01) **(01.1)** Correlação **(01.2)** Regressão **(01.3)** Regressão
(01.4) Correlação **(01.5)** Regressão

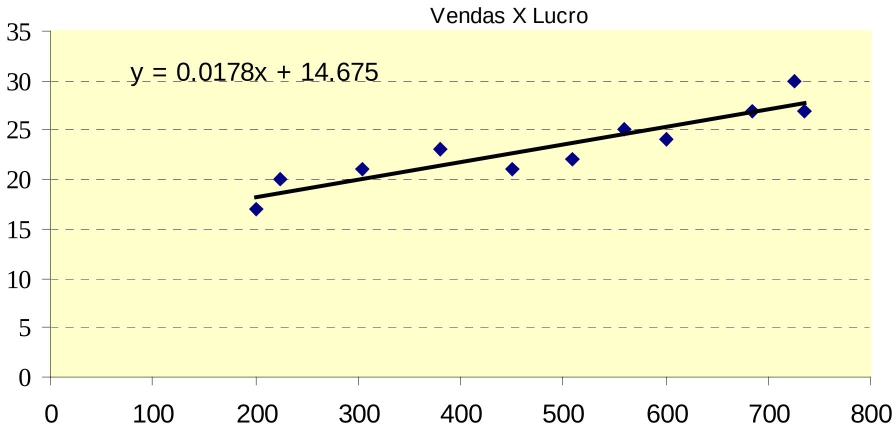
(02) **(02.1)** 4,80 s.m.

(03) **(03.1)** $\hat{Y} = -5 + 2.X$ **(03.2)** $\hat{Y} = -35 + 5.X$

(04) **(04.1)** Neste caso, com base no diagrama, uma linha reta não é adequada.



(04.2) Neste caso, uma linha é adequada e sua equação está sobre o gráfico abaixo.



(05) **(05.3)**

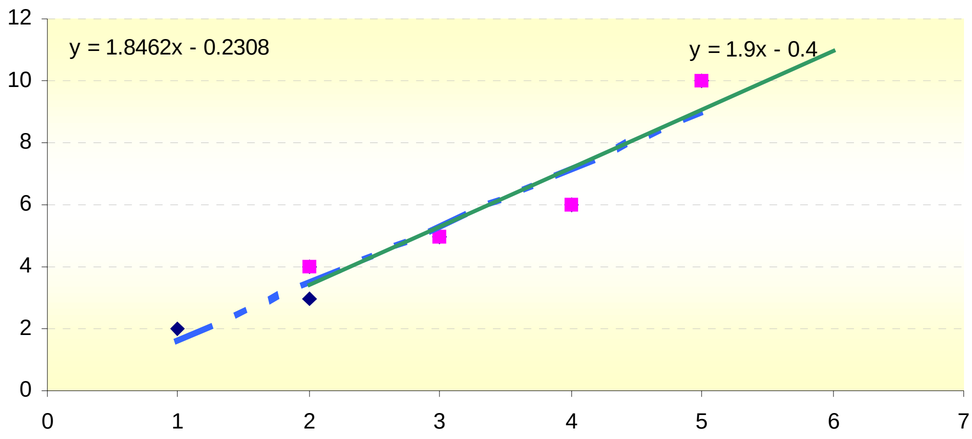
População

X	Y	Yc	Erro
1	2	1.62	0.38
4	6	7.15	-1.15
2	4	3.46	0.54
2	3	3.46	-0.46
3	5	5.31	-0.31
5	10	9.00	1.00
17	**30**	**30.00**	**0.00**

Amostra

X	Y
4	6
2	4
3	5
5	10

(05.1), (05.2) e (05.4)



(06) Basta mostrar que o ponto ($\bar{X}$, $\bar{Y}$) satisfaz a equação de regressão $\hat{Y}$ = a + bX. Se substituirmos X por $\bar{X}$ na equação o resultado deverá ser $\bar{Y}$. Mas a + b.X = a + b.$\bar{X}$ = $\bar{Y}$ - b$\bar{X}$ + b.$\bar{X}$ = $\bar{Y}$. Uma vez que a = $\bar{Y}$ - b$\bar{X}$.

(07) (07.1) $\hat{Y}$ = 4,2589 + 26,2770.X

(07.2) a = Custo fixo **b** = Custo marginal.

(07.3) s = 0,37. O intervalo de confiança de 95% para o "custo fixo" é: [3,09; 5,42] que contém o valor "5". Portanto não se pode afirmar, a 5% de significância que o custo fixo seja diferente do que 5 unidades.

(08) (08.1) $\hat{Y}$ = 9200 **(08.2)** 800 ± 270,02 **(08.3)** t_{48} = 2,009 (t_c = 5,952) **(08.4)** Não

(09) (09.1) $\hat{Y}$ = 3 + 0,48X **(09.2)** 2,24 **(09.3)** 17,25 ± 4,36

(10) [0,19; 0,77]

(11) (11.1) $\hat{Y}$ = 4,30 + 1,50X (r = 0,73) **(11.2)** S = 1,24 **(11.3)** t_{13} = 3,83

(12) (12.1) [44,69; 47,99] **(12.1)** [42,14; 50,54]

(13) (13.1) r = 0,87 **(13.2)** t_3 = 3,00

(13.3) 1,50 ± 1,59 **(13.4)** 8,50 ± 65,26

(14) (14.1) "-2" seria a taxa de incidência de câncer pulmonar que não está relacionada ao hábito de fumar, ou de quem nunca fumou. "1,70" é a variação na taxa de câncer pulmonar para cada ano que a pessoa fumou.

(14.2) $\hat{Y}$ = -2 + 1,70.20 = 32.

(14.3) Não, pois "r" indica associação na amostra e pode ser o mesmo na população.

(14.4) t_{48} = 5,20 que é significativo a 1%.

(15) (15.1) Tanto um coeficiente de "+1" quanto um de "-1" indicam correlação perfeita entre as variáveis.

(15.2) Coeficiente de regressão igual a zero implica em correlação também zero.

(15.3) Não necessariamente, pois neste caso "1" é o valor de inclinação da linha e não grau de associação linear entre as duas variáveis.

(15.4) Sim é possível.

(15.5) A técnica dos mínimos quadrados pode ser utilizado para ajustar vários tipos de equação.

(16) (16.1) Neste caso, a interpretação deve ser mais cuidadosa, pois tanto o excesso de chuvas quanto a falta vão distorcer os dados e estas equações podem não ser mais válidas.

(16.2) A safra B tira mais proveito, provavelmente por ser uma cultura que precisa de mais chuvas.

(16.3) Para a safra B pois existe uma melhor aderência dos dados a equação.

(17) (17.1) 26,28 ± 7,56 **(17.2)** 4,26 ± 1,17 **(17.3)** [81,46; 98,86]

(17.4) [78,45; 101,87] **(17.5)** t_3 = 11,57 **(17.6)** t_c = -1,159 e t_t -2,353, Aceito H_0.

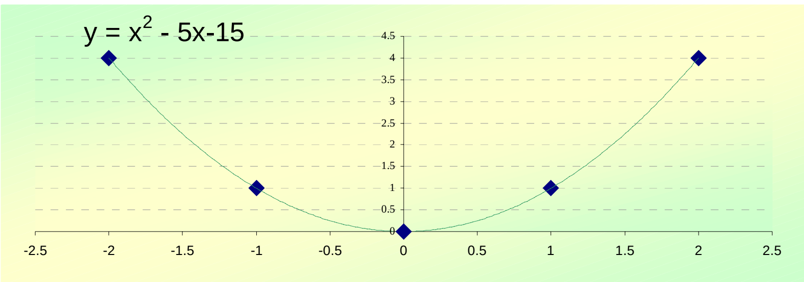
(17.7)) t_c = -2,010 e t_t -2,353, Aceito H_0.

(18) n = 36

(19) t_c = 1,35. Este valor não é significativo nem 5% e nem a 1%.

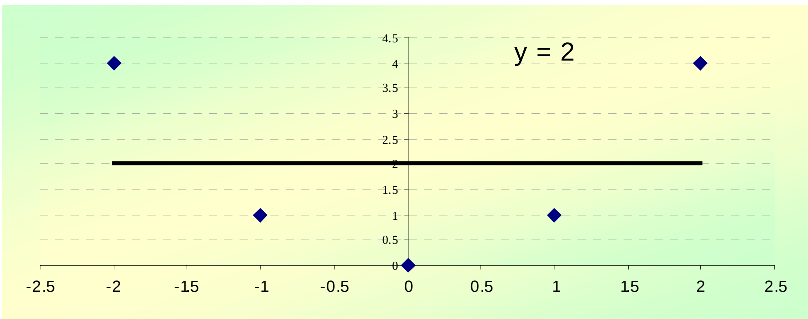
(20) ρ^2 = 64%, portanto não-explicada será: 1 - ρ^2 = 36%

(21) (21.1)



(21.2) r = 0

(21.3)



(21.4) Porque a correlação mostra apenas o relacionamento linear e, neste caso, o relacionamento é do tipo parábola (equação do segundo grau).

(22) (22.1) r = 0,8544

(22.2) $\hat{Y}$ = -30,4980 + 0,3247X

(22.3) $\hat{Y}$ = 132 mil

(22.4) $\hat{X}$ = 122,01 + 2,25.Y

5. REFERÊNCIAS

[BUS86] BUSSAB, Wilton O, MORETTIN, Pedro A. *Estatística Básica*. 3ª ed. São Paulo, Atual, 1986.

[DOW89] DOWNING, Douglas, CLARK, Jeff. *Statistics the Easy Way*. Barron's Educational Series, Inc. New York, 1989.

[FON76] FONSECA, Jairo Simon da, MARTINS, Gilberto de Andrade, TOLEDO, Geraldo Luciano. *Estatística Aplicada*. São Paulo: Editora Atlas, 1976.

[FON80] FONSECA, Jairo Simon da, MARTINS, Gilberto de Andrade. *Curso de Estatística*. São Paulo: Editora Atlas S. A., 1980.

[HOF80] HOFFMAN, Rodolfo. *Estatística para Economistas*. São Paulo. Livraria Pioneira Editora, 1980.

[KLE78] KLEIBAUM, David G., KUPPER, Lawrence L. *Applied Regression Analysis and Other Multivariable Methods.* North Scituate, Massachusetts: Duxbury Press, 1978.

[MAR87] MARKLAND, Robert E., SWEIGART, James R. *Quantitative Methods: Applications to Managerial Decision Making*. New York: John Wiley & Sons, 1987. 827p.

[MAS90] MASON, Robert D., DOUGLAS, Lind A. *Statistical Techniques in Business And Economics*. IRWIN, Boston, 1990.

[MEY78] MEYER, Paul L. *Probabilidade: aplicações à Estatística.* Tradução do Prof. Ruy C. B. Lourenço Filho. Rio de Janeiro, Livros Técnicos e Científicos Editora S.A., 1978.

[MIL90] MILLER, Charles D., HEEREN, Vern E., HORNSBY Jr., E. John. *Mathematical Ideas*. USA: Harper Collins Publishers, 1990.

[REA93] *The Statistics Problem Solver.* Research and Education Association, Piscataway, New Jersey, 1993.

[ROT91] ROTHENBERG, Ronald I. *Probability and Statistics*. Hartcourt Brace Jovanovich, Publishers, Orlando, Florida, 1991.

[SAL82] SALVATORE, Dominick. *Estatística e Econometria*. Tradução Newton Boer, revisão técnica Marco Antônio S. de Vasconcelos. São Paulo: McGraw-Hill do Brasil, 1982.