AULAS 16 E 17
Correlação, análise fatorial e regressão

Ernesto F. L. Amaral

08 e 10 de outubro de 2013
Metodologia de Pesquisa (DCP 854B)

Fonte:
Triola, Mario F. 2008. "Introdução à estatística". 10 ª ed. Rio de Janeiro: LTC. Capítulo 10 (pp.408-467).

ESTRUTURA DA AULA

– Correlação.

– Análise fatorial.

– Regressão.

– Variação e intervalos de previsão.

– Regressão múltipla.

– Modelagem.

VISÃO GERAL

- Vamos falar de métodos para:
  - Fazer inferências sobre a relação (**correlação**) entre duas variáveis.
  - Agrupar variáveis que medem dimensões conceituais similares e criar índices (**análise fatorial**).
  - Elaborar uma equação que possa ser usada para prever o valor de uma variável dado o valor de outra (**regressão**).

- Serão considerados dados amostrais que vêm em pares.
  - No capítulo anterior, as inferências se referiam à **média das diferenças** entre pares de valores.
  - Neste capítulo, as inferências têm objetivo de verificar **relação** entre duas variáveis.

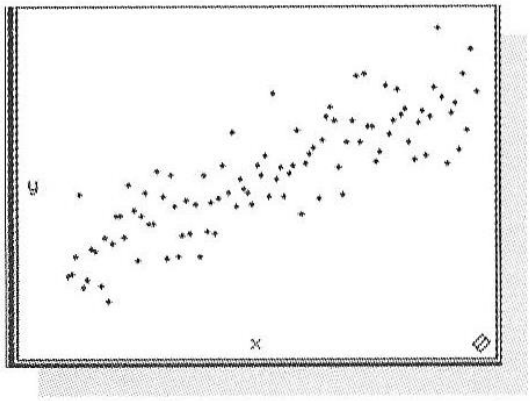
CORRELAÇÃO

CONCEITOS BÁSICOS

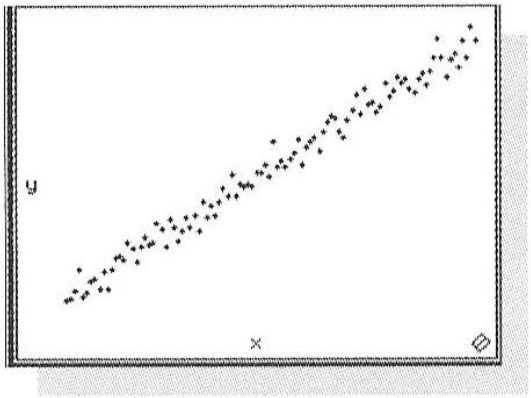
– Existe uma correlação entre duas variáveis quando uma delas está relacionada com a outra de alguma maneira.

– Antes de tudo é importante explorar os dados:
  – Diagrama de dispersão entre duas variáveis.
  – Há tendência?
  – Crescente ou decrescente?
  – *Outliers*?

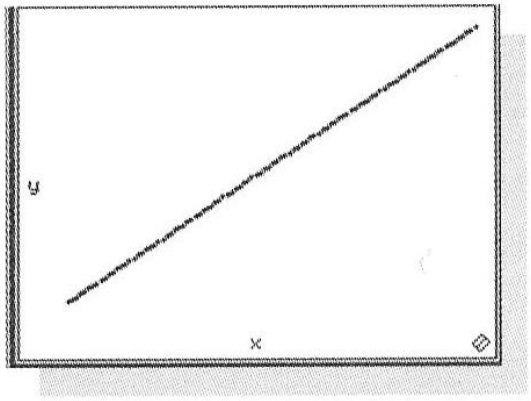
DIAGRAMAS DE DISPERSÃO
(correlação linear)



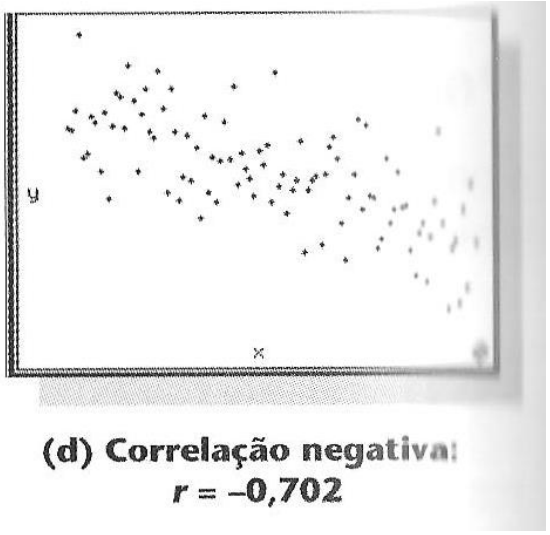
(a) Correlação positiva:
$r = 0{,}851$



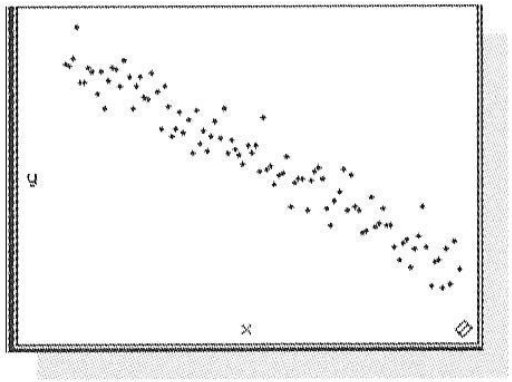
(b) Correlação positiva:
$r = 0{,}991$



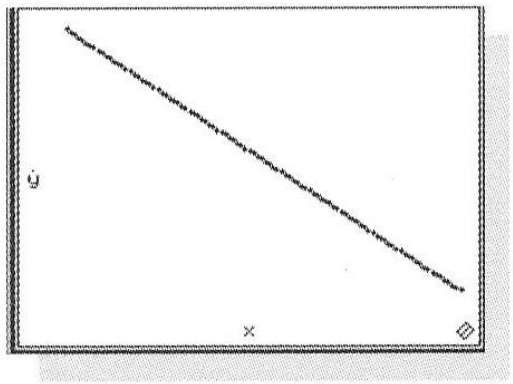
(c) Correlação positiva perfeita:
$r = 1$



(d) Correlação negativa:
$r = -0{,}702$

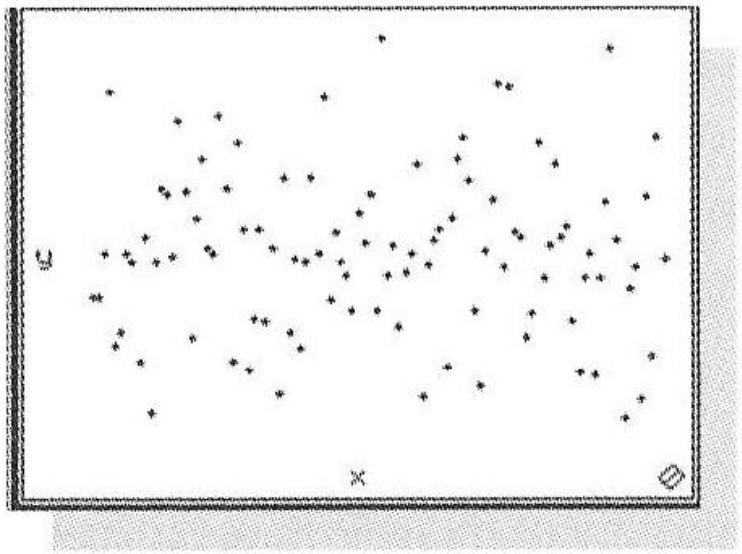


(e) Correlação negativa:
$r = -0{,}965$

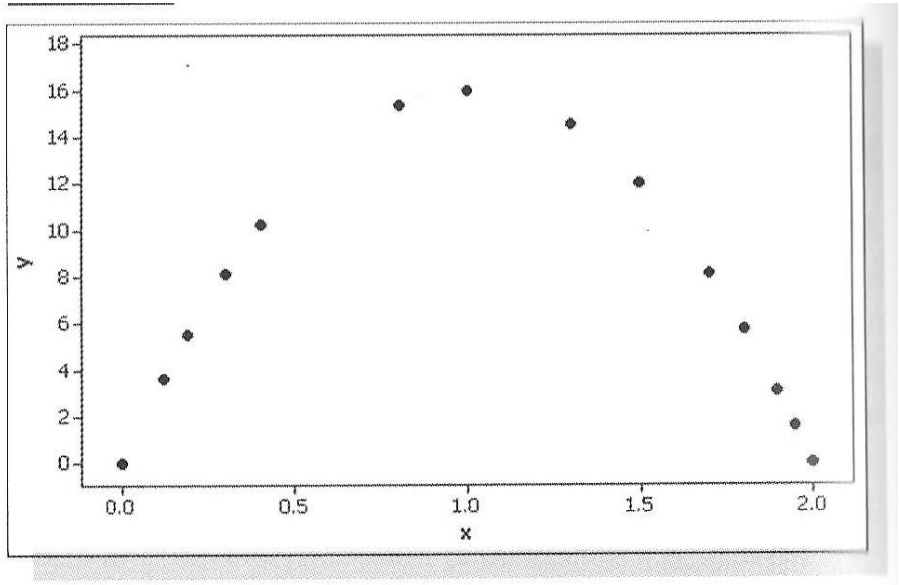


(f) Correlação negativa perfeita:
$r = -1$

DIAGRAMAS DE DISPERSÃO
(não há correlação linear)



(g) Nenhuma correlação: $r = 0$



(h) Relação não-linear: $r = -0,087$

CORRELAÇÃO

– O coeficiente de correlação linear (*r*):

– Medida numérica da força da relação entre duas variáveis que representam dados quantitativos.

– Mede intensidade da relação linear entre os valores quantitativos emparelhados *x* e *y* em uma amostra.

– É chamado de coeficiente de correlação do produto de momentos de Pearson.

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

OBSERVAÇÕES IMPORTANTES

– Usando dados amostrais emparelhados (dados bivariados), estimamos valor de *r* para concluir se há ou não relação entre duas variáveis.

– Serão tratadas relações lineares, em que pontos no gráfico (x, y) se aproximam do padrão de uma reta.

– É importante entender os conceitos e não os cálculos aritméticos.

– *r* é calculado com dados amostrais. Se tivéssemos todos pares de valores populacionais *x* e *y*, teríamos um parâmetro populacional (ρ).

REQUISITOS

– Os seguintes requisitos devem ser satisfeitos ao se testarem hipóteses ou ao se fazerem outras inferências sobre *r* :

- – Amostra de dados emparelhados (*x*, *y*) é uma **amostra aleatória** de dados quantitativos independentes.
  - – Não pode ter sido utilizado, por exemplo, amostra de resposta voluntária.
- – Exame visual do diagrama de dispersão deve confirmar que pontos se aproximam do **padrão de uma reta**.
- – **Valores extremos** (*outliers*) devem ser removidos se forem erros.
  - – Efeitos de outros *outliers* devem ser considerados com estimação de *r* com e sem estes *outliers*.

VALORES CRÍTICOS DO COEFICIENTE DE CORRELAÇÃO DE PEARSON (*r*)

n	$\alpha = 0,05$	$\alpha = 0,01$
4	0,950	0,999
5	0,878	0,959
6	0,811	0,917
7	0,754	0,875
8	0,707	0,834
9	0,666	0,798
10	0,632	0,765
11	0,602	0,735
12	0,576	0,708
13	0,553	0,684
14	0,532	0,661
15	0,514	0,641
16	0,497	0,623
17	0,482	0,606
18	0,468	0,590
19	0,456	0,575
20	0,444	0,561
25	0,396	0,505
30	0,361	0,463
35	0,335	0,430
40	0,312	0,402
45	0,294	0,378
50	0,279	0,361
60	0,254	0,330
70	0,236	0,305
80	0,220	0,286
90	0,207	0,269
100	0,196	0,256

NOTA: Para testar H_0: $\rho = 0$ *versus* H_1: $\rho \neq 0$, rejeite H_0 se o valor absoluto de *r* for maior que o valor crítico na tabela.

– **Arredonde** o coeficiente de correlação linear *r* para três casas decimais, permitindo comparação com esta tabela.

– Interpretação: com 4 pares de dados e **nenhuma correlação** linear entre *x* e *y*, há chance de 5% de que valor absoluto de *r* exceda 0,950.

INTERPRETANDO *r*

– O valor de *r* deve sempre estar entre –1 e +1.

– Se *r* estiver muito próximo de 0, concluímos que não há correlação linear significativa entre *x* e *y*.

– Se *r* estiver próximo de –1 ou +1, concluímos que há uma relação linear significativa entre *x* e *y*.

– Mais objetivamente:

  – Usando a tabela anterior, se valor absoluto de *r* excede o valor da tabela, há correlação linear.

  – Usando programa de computador, se valor *P* é menor do que nível de significância, há correlação linear.

PROPRIEDADES DE *r*

– Valor de *r* está entre: $-1 \leq r \leq +1$

– Valor de *r* não muda se todos valores de qualquer das variáveis forem convertidos para uma escala diferente.

– Valor de *r* não é afetado pela inversão de *x* ou *y*. Ou seja, mudar os valores de *x* pelos valores de *y* e vice-versa não modificará *r*.

– *r* mede intensidade de relação linear, não sendo planejado para medir intensidade de relação que não seja linear.

– O valor de r^2 é a proporção da variação em *y* que é explicada pela relação linear entre *x* e *y*.

ERROS DE INTERPRETAÇÃO

– Erro comum é concluir que correlação implica **causalidade**:
  – A causa pode ser uma variável oculta.
  – Uma variável oculta é uma variável que afeta as variáveis em estudo, mas que não está incluída no banco.

– Erro surge de dados que se baseiam em **médias**:
  – Médias suprimem variação individual e podem aumentar coeficiente de correlação.

– Erro decorrente da propriedade de **linearidade**:
  – Pode existir relação entre *x* e *y* mesmo quando não haja correlação linear (relação quadrática, por exemplo).

TESTE DE HIPÓTESE FORMAL PARA CORRELAÇÃO

– É possível realizar um teste de hipótese formal para determinar se há ou não relação linear significativa entre duas variáveis.

– Critério de decisão é rejeitar a hipótese nula (ρ=0) se o valor absoluto da estatística de teste exceder os valores críticos.

– A rejeição de (ρ=0) significa que há evidência suficiente para apoiar a afirmativa de uma correlação linear entre as duas variáveis.

– Se o valor absoluto da estatística de teste não exceder os valores críticos (ou seja, o valor *P* for grande), deixamos de rejeitar ρ=0.

H_0: ρ=0 (não há correlação linear)

H_1: $\rho \neq 0$ (há correlação linear)

MÉTODO 1: ESTATÍSTICA DE TESTE É *t*

– Estatística de teste representa o valor do desvio padrão amostral dos valores de *r* :

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

– Valores críticos e valor *P*: use tabela A-3 com *n*–2 graus de liberdade.

– Conclusão:

  – Se |*t*| > valor crítico da Tabela A-3, rejeite H_0 e conclua que há correlação linear.

  – Se |*t*| ≤ valor crítico da Tabela A-3, deixe de rejeitar H_0 e conclua que não há evidência suficiente para concluir que haja correlação linear.

MÉTODO 2: ESTATÍSTICA DE TESTE É *r*

– Estatística de teste: *r*

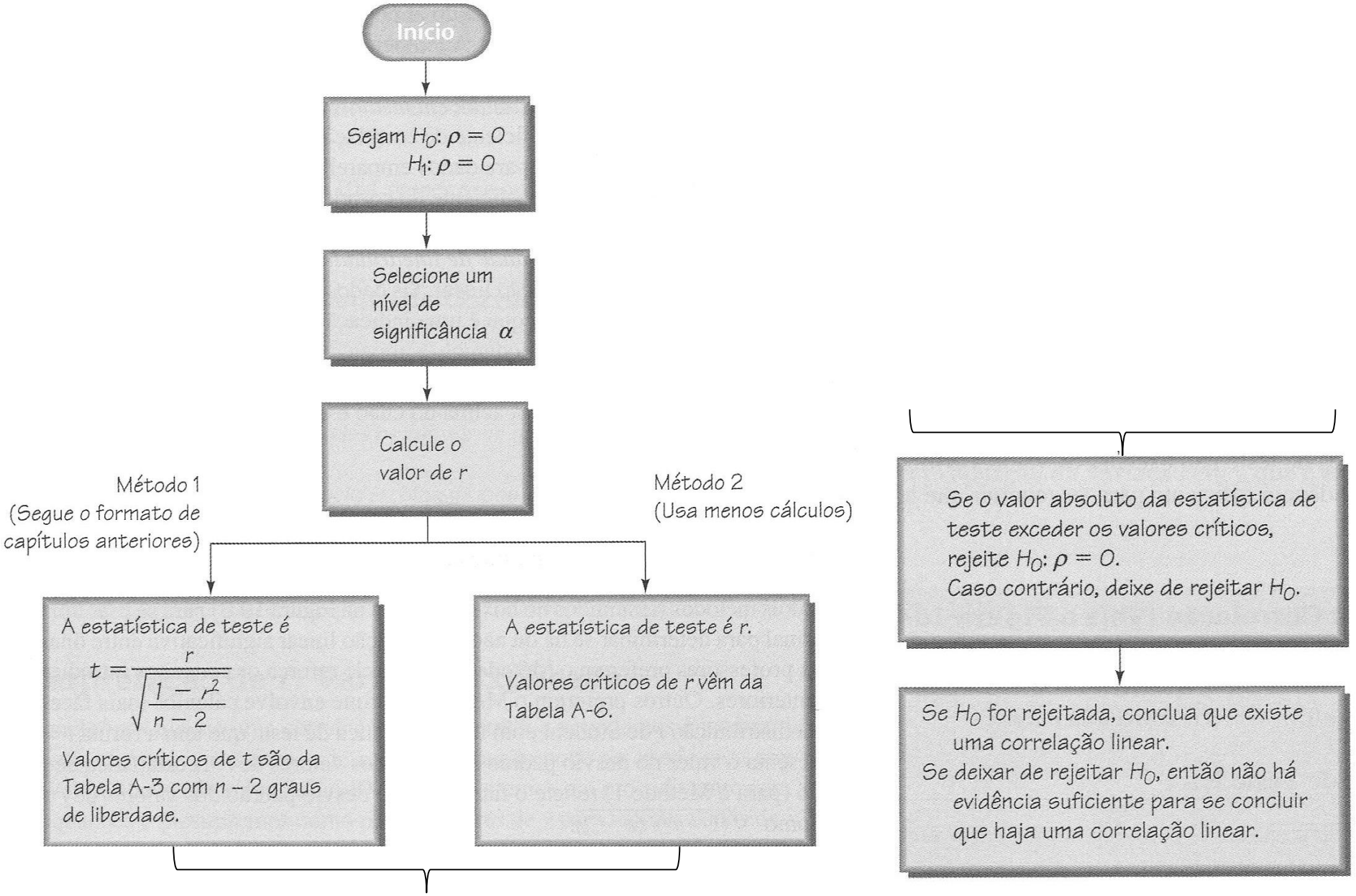
– Valores críticos: consulte Tabela A-6.

– Conclusão:

  – Se $|r|$ > valor crítico da Tabela A-6, rejeite H_0 e conclua que há correlação linear.

  – Se $|r|$ ≤ valor crítico da Tabela A-6, deixe de rejeitar H_0 e conclua que não há evidência suficiente para concluir que haja correlação linear.

TESTE DE HIPÓTESE PARA CORRELAÇÃO LINEAR



TESTES UNILATERAIS

– Os testes unilaterais podem ocorrer com uma afirmativa de uma correlação linear positiva ou uma afirmativa de uma correlação linear negativa.

– Afirmativa de correlação negativa (teste unilateral esquerdo):

$$H_0: \rho = 0$$

$$H_1: \rho < 0$$

– Afirmativa de correlação positiva (teste unilateral direito):

$$H_0: \rho = 0$$

$$H_1: \rho > 0$$

– Para isto, simplesmente utilize α=0,025 (ao invés de α=0,05) e α=0,005 (ao invés de α=0,01).

FUNDAMENTOS

– Essas fórmulas são diferentes versões da mesma expressão:

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2}\sqrt{n(\sum y^2) - (\sum y)^2}}$$

$$r = \frac{\sum\left[\frac{(x - \bar{x})}{s_x}\frac{(y - \bar{y})}{s_y}\right]}{n - 1}$$

$$r = \frac{\sum(x - \bar{x})(y - \bar{y})}{(n - 1)s_x s_y}$$

$$r = \frac{s_{xy}}{\sqrt{s_{xx}}\sqrt{s_{yy}}}$$

FUNDAMENTOS

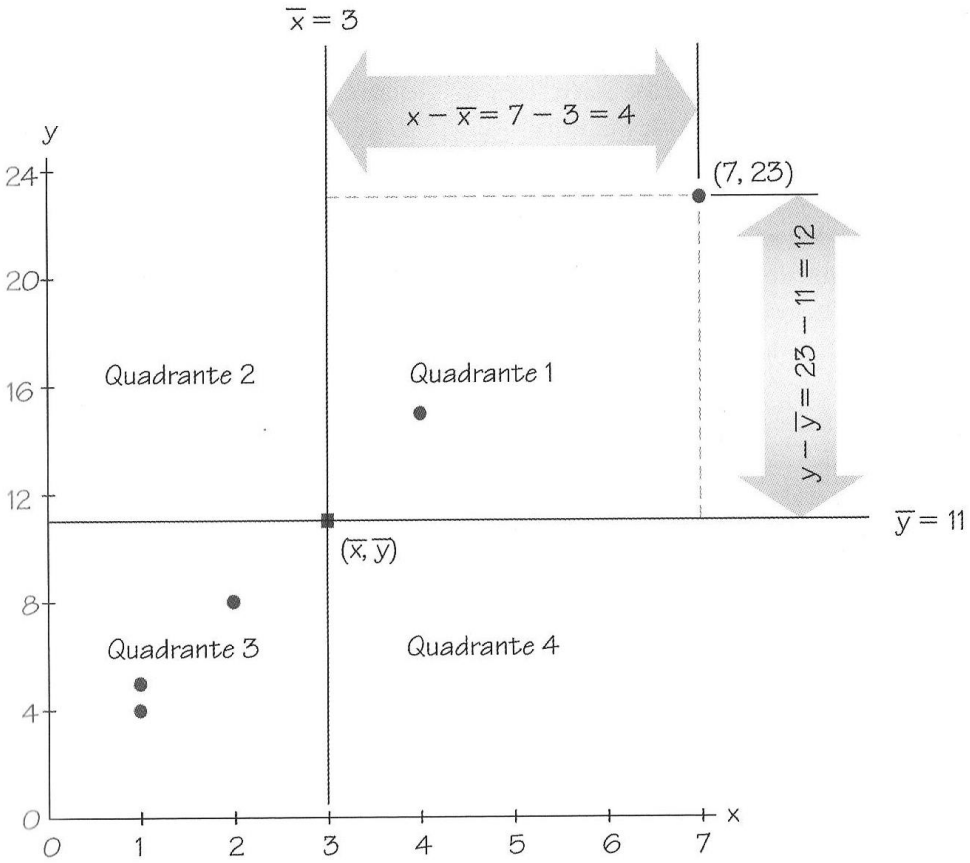
– Dada uma coleção de dados em pares (*x*,*y*), o ponto (*x*-barra, *y*-barra) é chamado de **centróide**.

– A estatística do produto dos momentos de Pearson (*r*) se baseia na soma dos produtos dos momentos:

$$\sum (x - \bar{x})(y - \bar{y})$$

– Se pontos são reta ascendente, valores do produto estarão nos 1º e 3º quadrantes (soma positiva).

– Se é descendente, os pontos estarão nos 2º e 4º quadrantes (soma negativa).



OU SEJA...

– Podemos usar esta expressão para medir como pontos estão organizados:

$$\sum (x - \bar{x})(y - \bar{y})$$

– Grande soma positiva sugere pontos predominantemente no primeiro e terceiro quadrantes (correlação linear positiva).

– Grande soma negativa sugere pontos predominantemente no segundo e quarto quadrantes (correlação linear negativa).

– Soma próxima de zero sugere pontos espalhados entre os quatro quadrantes (não há correlação linear).

PORÉM...

– Esta soma depende da magnitude dos números usados:

$$\sum (x - \bar{x})(y - \bar{y})$$

– Para tornar *r* independente da escala utilizada, usamos a seguinte padronização:

  – Sendo s_x o desvio padrão dos valores amostrais *x*...

  – Sendo s_y o desvio padrão dos valores amostrais *y*...

  – Padronizamos cada desvio pela sua divisão por s_x...

$$\sum \left[\frac{(x - \bar{x})}{s_x} \frac{(y - \bar{y})}{s_y} \right]$$

– Usamos o divisor $n - 1$ para obter uma espécie de média:

$$r = \frac{\sum \left[\frac{(x - \bar{x})}{s_x} \frac{(y - \bar{y})}{s_y} \right]}{n - 1}$$

COMANDOS NO STATA

– Podemos usar os comandos “correlate” ou “pwcorr”, em que ambos mostram a matriz de correlações entre as variáveis.

– O comando “corr” usa “listwise deletion”, em que toda matriz é calculada somente para casos que não possuem nenhum valor em branco (*missing*) em nenhuma variável na lista:

corr x y z

– O comando “pwcorr” usa “pairwise deletion”, em que cada correlação é computada para casos que não possuem nenhum valor em branco para cada par de variáveis:

pwcorr x y z, sig

– Uso do “pwcorr” para obter o mesmo que “corr”:

pwcorr x y z if !missing(x, y, z), sig

ANÁLISE FATORIAL

ANÁLISE FATORIAL

– Os **fatores** ou **construtos** são variáveis hipotéticas, combinações lineares das variáveis observadas, que explicam partes da variabilidade dos dados.

– **Análise fatorial** é usada principalmente com o objetivo de simplificar os dados:

1) Pegar um pequeno número de variáveis (preferencialmente não-correlacionadas) de um grande número de variáveis (em que maioria é correlacionada com a outra).

2) Criar índices com variáveis que medem dimensões conceituais similares.

TIPOS DE ANÁLISE FATORIAL

– **Exploratória:** quando não temos uma idéia pré-definida da estrutura e de quantas dimensões estão presentes em um conjunto de variáveis.

```
factor var1 var2 var3 ... varn
```

– **Confirmatória:** quando queremos testar hipóteses específicas sobre a estrutura de um número de dimensões subjacentes a um conjunto de variáveis. Por exemplo, pensamos que nossos dados possuem duas dimensões e queremos verificar isto.

```
factor var1 var2 var3 ... varn, factors(#)
```

– **Ajuda no Stata:**

```
help factor
```

MATRIZES

– A **matriz de correlação** é uma matriz quadrada cujos elementos são as correlações entre as variáveis analisadas.

– Na diagonal principal todos os elementos são iguais a 1 (um), visto que cada variável é totalmente correlacionada com ela mesma.

– A **matriz de covariância** é uma matriz quadrada cujos elementos fora da diagonal principal são as covariâncias entre as variáveis e na diagonal principal são as variâncias de cada variável.

MATRIZ DE CORRELAÇÃO (“corr” NO STATA)

```
. corr terpro aguarg escoarg lixod eletrica
(obs=134)

             |   terpro   aguarg  escoarg    lixod eletrica
-------------+------------------------------------------------
      terpro |   1.0000
      aguarg |   0.0021   1.0000
     escoarg |  -0.0192   0.2343   1.0000
       lixod |   0.1246   0.4393   0.5296   1.0000
    eletrica |   0.1214   0.2126   0.1253   0.1435   1.0000
```

MATRIZ DE CORRELAÇÃO (“pwcorr” NO STATA)

```
. pwcorr terpro aguarg escoarg lixod eletrica if !missing(terpro,
  aguarg, escoarg, lixod, eletrica), sig
```

	terpro	aguarg	escoarg	lixod	eletrica
terpro	1,0000				
aguarg	0,0021	1,0000			
	0,9810				
escoarg	-0,0192	0,2343	1,0000		
	0,8253	0,0064			
lixod	0,1246	0,4393	0,5296	1,0000	
	0,1514	0,0000	0,0000		
eletrica	0,1214	0,2126	0,1253	0,1435	1,0000
	0,1624	0,0136	0,1493	0,0982	

ANÁLISE DE COMPONENTES PRINCIPAIS

– Na **análise de componentes principais - ACP (*principal component factors - PCF*)**, quase todas variáveis estão altamente correlacionadas ao primeiro fator.

– Ou seja, é definido que a primeira componente (fator) explique a maior parte da variabilidade dos dados e por conseqüência as variáveis estarão mais correlacionadas a ela.

```
factor var1 var2 var3 ... varn, pcf
```

COMPONENTES DA ANÁLISE FATORIAL

```
. factor terpro aguarg escoarg lixod eletrica, pcf
(obs=134)

Factor analysis/correlation                       Number of obs    =      134
    Method: principal-component factors           Retained factors =        2
    Rotation: (unrotated)                         Number of params =        9
    --------------------------------------------------------------------------
        Factor  |   Eigenvalue   Difference        Proportion   Cumulative
    ------------+-------------------------------------------------------------
        Factor1 |      1.90885      0.84141            0.3818       0.3818
        Factor2 |      1.06744      0.17154            0.2135       0.5953
        Factor3 |      0.89590      0.15994            0.1792       0.7744
        Factor4 |      0.73596      0.34410            0.1472       0.9216
        Factor5 |      0.39186            .            0.0784       1.0000
    --------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(10) =   84.37 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances
    ------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
          terpro |   0.1573    0.8300 |      0.2864
          aguarg |   0.6914   -0.0594 |      0.5185
         escoarg |   0.7233   -0.3011 |      0.3862
           lixod |   0.8428   -0.1055 |      0.2786
        eletrica |   0.4155    0.5228 |      0.5541
    ------------------------------------------------
```

factor ideol equality owner respon competition, pcf

Variables

Principal-components factoring

Total variance accounted by each factor. The sum of all eigenvalues = total number of variables.

When negative, the sum of eigenvalues = total number of factors (variables) with positive eigenvalues.

Kaiser criterion suggests to retain those factors with eigenvalues equal or higher than 1.

Since the sum of eigenvalues = total number of variables. Proportion indicate the relative weight of each factor in the total variance. For example, 1.54525/5=0.3090. The first factor explains 30.9% of the total variance

Cumulative shows the amount of variance explained by n+(n-1) factors. For example, factor 1 and factor 2 account for 57.55% of the total variance.

Difference between one eigenvalue and the next.

Uniqueness is the variance that is 'unique' to the variable and not shared with other variables. It is equal to 1 – communality (variance that is shared with other variables). For example, 61.57% of the variance in 'ideol' is not share with other variables in the overall factor model. On the contrary 'owner' has low variance not accounted by other variables (28.61%). Notice that the greater 'uniqueness' the lower the relevance of the variable in the factor model.

```
. factor  ideol equality owner respon competition, pcf
(obs=1125)

Factor analysis/correlation                       Number of obs    =       1125
    Method: principal-component factors           Retained factors =          2
    Rotation: (unrotated)                         Number of params =          9

    Factor  |  Eigenvalue   Difference        Proportion   Cumulative
------------+------------------------------------------------------------
    Factor1 |     1.54524      0.21290           0.3090       0.3090
    Factor2 |     1.33235      0.49085           0.2665       0.5755
    Factor3 |     0.84149      0.12808           0.1683       0.7438
    Factor4 |     0.71341      0.14590           0.1427       0.8865
    Factor5 |     0.56751            .           0.1135       1.0000
------------+------------------------------------------------------------
LR test: independent vs. saturated:  chi2(10) =  398.10 Prob>chi2 = 0.0000

Factor loadings (pattern matrix) and unique variances

    Variable |   Factor1   Factor2 |  Uniqueness
-------------+---------------------+--------------
       ideol |    0.4719    0.4019 |      0.6157
    equality |    0.4066    0.6424 |      0.4220
       owner |    0.6179   -0.5762 |      0.2861
      respon |    0.5807    0.4130 |      0.4922
 competition |    0.6619   -0.5056 |      0.3063
```

Factor loadings are the weights and correlations between each variable and the factor. The higher the load the more relevant in defining the factor's dimensionality. A negative value indicates an inverse impact on the factor. Here, two factors are retained because both have eigenvalues over 1. It seems that 'owner' and 'competition' define factor1, and 'equality', 'respon' and 'ideol' define factor2.

Fonte: Torres-Reyna, Oscar. s.d. *Getting Started in Factor Analysis (using Stata).* (http://dss.princeton.edu/training/)

AUTOVALORES & DIFERENÇA

– Os **autovalores (*eigenvalues*)** são valores obtidos a partir das matrizes de covariância ou de correlação, cujo objetivo é obter um conjunto de vetores independentes, não correlacionados, que expliquem o máximo da variabilidade dos dados.

– Indicam o total da variância causado por cada fator.

– A soma de todos autovalores é igual ao número de variáveis. Quando há valores negativos, a soma dos autovalores é igual ao número total de variáveis com valores positivos.

– O **critério de *Kaiser*** sugere utilizar os fatores com autovalores iguais ou superiores a uma unidade.

– **Diferença (*difference*)** é a subtração entre um autovalor e o próximo autovalor.

PROPORÇÃO & CUMULATIVA

– **Proporção (*proportion*)** indica o peso relativo de cada fator na variância total (variabilidade) dos dados.

– **Proporção cumulativa (*cumulative*)** indica o total da variância explicado por n+(n-1) fatores.

CARGAS FATORIAIS

– **Cargas fatoriais (*factors*)** são as correlações entre as variáveis originais e os fatores.

– Esse é um dos pontos principais da análise fatorial, quanto maior a carga fatorial maior será a correlação com determinado fator.

– Um valor negativo indica um impacto inverso no fator.

– A quantidade de cargas fatoriais é automaticamente calculada pelo Stata com base nos autovalores (iguais ou superiores a uma unidade - critério de *Kaiser*).

– Por sua vez, as **cargas fatoriais relevantes** são aquelas com valores **maiores que 0,5**.

COMUNALIDADE & ESPECIFICIDADE

– As **comunalidades (*communalities*)** são quantidades das variâncias (correlações) de cada variável explicada pelos fatores. Quanto maior a comunalidade, maior será o poder de explicação daquela variável pelo fator.

Desejamos comunalidades superiores a 0,5.

– A **especificidade** (***uniqueness***) ou erro é a parcela da variância (correlação) dos dados que não pode ser explicada pelo fator. É a proporção única da variável não compartilhada com as outras variáveis. É igual a 1 menos a comunalidade. Quanto maior a especificidade, menor é a relevância da variável no modelo fatorial.

Desejamos especificidades inferiores a 0,5.

ROTAÇÃO FATORIAL

– A **ACP** é um método estatístico multivariado que permite transformar um conjunto de variáveis iniciais correlacionadas entre si, num outro conjunto de variáveis não-correlacionadas (ortogonais), as chamadas componentes principais, que resultam de combinações lineares do conjunto inicial.

– Uma **rotação fatorial (*rotation*)** é o processo de manipulação ou de ajuste dos eixos fatoriais para conseguir uma solução fatorial mais simples e pragmaticamente mais significativa, cujos fatores sejam mais facilmente interpretáveis.

– A nova matriz padrão apresenta de forma mais clara a relevância de cada variável em cada fator.

`rotate`

EXEMPLO DE ROTAÇÃO FATORIAL

```
. rotate

Factor analysis/correlation                         Number of obs    =      134
    Method: principal-component factors             Retained factors =        2
    Rotation: orthogonal varimax (Kaiser off)       Number of params =        9

    ------------------------------------------------------------------------------
        Factor  |     Variance   Difference        Proportion   Cumulative
    ------------+-----------------------------------------------------------------
        Factor1 |      1,85704      0,73779            0,3714       0,3714
        Factor2 |      1,11924            .            0,2238       0,5953
    ------------------------------------------------------------------------------
    LR test: independent vs. saturated:  chi2(10) =   84,37 Prob>chi2 = 0,0000

Rotated factor loadings (pattern matrix) and unique variances

    -------------------------------------------------
        Variable |  Factor1   Factor2 |   Uniqueness
    -------------+--------------------+--------------
          terpro |  -0,0536    0,8430 |      0,2864
          aguarg |   0,6845    0,1140 |      0,5185
         escoarg |   0,7754   -0,1122 |      0,3862
           lixod |   0,8426    0,1069 |      0,2786
        eletrica |   0,2728    0,6095 |      0,5541
    -------------------------------------------------

Factor rotation matrix

    --------------------------------
                 | Factor1  Factor2
    -------------+------------------
         Factor1 |  0,9687   0,2481
         Factor2 | -0,2481   0,9687
    --------------------------------
```

rotate

By default the rotation is varimax which produces orthogonal factors. This means that factors are not correlated to each other. This setting is recommended when you want to identify variables to create indexes or new variables without inter-correlated components

Same description as in the previous slide with new composition between the two factors. Still both factors explain 57.55% of the total variance observed.

The pattern matrix here offers a clearer picture of the relevance of each variable in the factor. Factor1 is mostly defined by 'owner' and 'competition' and factor2 by 'equality', 'respon' and 'ideol' .

This is a correlation matrix between factor1 and factor2.

```
. rotate

Factor analysis/correlation                       Number of obs    =      1125
    Method: principal-component factors           Retained factors =         2
    Rotation: orthogonal varimax (Kaiser off)     Number of params =         9

    Factor     Variance   Difference    Proportion   Cumulative
    Factor1     1.45169      0.02579        0.2903       0.2903
    Factor2     1.42590            .        0.2852       0.5755

    LR test: independent vs. saturated:  chi2(10) =  398.10 Prob>chi2 = 0.0000

Rotated factor loadings (pattern matrix) and unique variances

    Variable    Factor1    Factor2   Uniqueness
       ideol     0.0869     0.6138       0.6157
    equality    -0.1214     0.7505       0.4220
       owner     0.8446    -0.0218       0.2861
      respon     0.1610     0.6941       0.4922
  competition     0.8307     0.0603       0.3063

Factor rotation matrix

                Factor1    Factor2
    Factor1      0.7487     0.6629
    Factor2     -0.6629     0.7487
```

NOTE: If you want the factors to be correlated (oblique rotation) you need to use the option `promax` after `rotate`:

```
rotate, promax
```

Type `help rotate` for details.

Fonte: Torres-Reyna, Oscar. s.d. *Getting Started in Factor Analysis (using Stata).* (http://dss.princeton.edu/training/)

TESTE KAISER-MEYER-OLKLIN

– O teste de Kaiser-Meyer-Olklin (KMO) varia entre 0 e 1. Quanto mais perto de 1, melhor.

– Friel (2009) sugere a seguinte escala para interpretar o valor da estatística KMO:

* Entre 0,90 e 1: excelente.
* Entre 0,80 e 0,89: bom.
* Entre 0,70 e 0,79: mediano.
* Entre 0,60 e 0,69: medíocre.
* Entre 0,50 e 0,59: ruim.
* Entre 0 e 0,49: inadequado.

– Pallant (2007) sugere 0,60 como um limite razoável.

– Hair et al. (2006) sugerem 0,50 como patamar aceitável.

`estat kmo`

EXEMPLO DE TESTE KAISER-MEYER-OLKLIN

```
. estat kmo

Kaiser-Meyer-Olkin measure of sampling adequacy

    ----------------------
        Variable |     kmo
    -------------+--------
          terpro |  0,3311
          aguarg |  0,6166
         escoarg |  0,5726
           lixod |  0,5496
        eletrica |  0,6504
    -------------+--------
         Overall |  0,5676
    ----------------------
```

O teste de Kaiser-Meyer-Olklin (0,5676) indica um resultado ruim, mas em patamar razoável (Hair et al. 2006).

CRIANDO NOVAS VARIÁVEIS

– Para criar novas variáveis automaticamente:

```
predict factor1 factor2
```

– Outra opção seria criar manualmente índices para cada conglomerado de variáveis.

– Por exemplo, se temos variáveis binárias:

```
gen factor1 = var1 + var2
```

– Por exemplo, se temos variáveis contínuas e que possuem valores mínimos e máximos compatíveis:

```
gen factor2 = (var3 + var4 + var5) / 3
```

EXEMPLO DE CRIAÇÃO DE ÍNDICES

Problema é que cada fator leva em consideração todas variáveis incluídas no comando, mesmo que com pesos diferenciados para cada fator:

```
factor terpro aguarg escoarg lixod eletrica, pcf
rotate
predict factor1 factor2
```

Baseado nas cargas fatoriais maiores que 0,5 do fator 1:

```
factor aguarg escoarg lixod
rotate
predict factor11
```

Criando índice manualmente:

```
gen factor12 = aguarg + escoarg + lixod
```

predict factor1 factor2

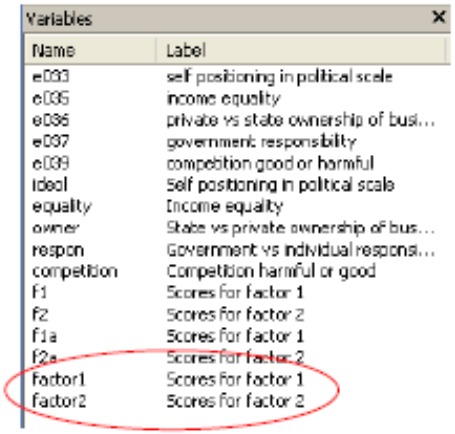
```
predict factor1 factor2 /*or whatever name you prefer to identify the factors*/
```

```
. predict factor1 factor2
(regression scoring assumed)

Scoring coefficients (method = regression; based on varimax rotated factors)
```

Variable	Factor1	Factor2
ideol	0.02868	0.42832
equality	-0.12258	0.53541
owner	0.58610	-0.05873
respon	0.07591	0.48119
competition	0.57225	-0.00014

These are the regression coefficients used to estimate the individual scores (per case/row)



Another option could be to create indexes out of each cluster of variables. For example, 'owner' and 'competition' define one factor. You could aggregate these two to create a new variable to measure 'market oriented attitudes'. On the other hand you could aggregate 'ideol', 'equality' and 'respon' to create an index to measure 'egalitarian attitudes'. Since all variables are in the same valence (liberal for small values, capitalist for larger values), we can create the two new variables as

```
gen market = (owner + competition)/2
gen egalitatiran = (ideol + equality + respon)/3
```

Fonte: Torres-Reyna, Oscar. s.d. *Getting Started in Factor Analysis (using Stata).* (http://dss.princeton.edu/training/)

SUGESTÕES DE LEITURA

Hamilton, Lawrence C. 2006. *Statistics with STATA (updated for version 9)*. Thomson Books/Cole.

Moraes, Odair Barbosa; Alex Kenya Abiko. 2006. *Utilização da Análise Fatorial para a Identificação de Estruturas de Interdependência de Variáveis em Estudos de Avaliação Pós-Ocupação*. XI Encontro Nacional de Tecnologia no Ambiente Construído (ENTAC).

Kim, Jae-on; Charles W. Mueller. 1978. *Factor Analysis. Statistical Methods and Practical Issues*. Sage publications.

Kim, Jae-on; Charles W. Mueller. 1978. *Introduction to Factor Analysis. What it is and How To Do It*. Sage publications.

StatNotes (http://faculty.chass.ncsu.edu/garson/PA765/factor.htm).

StatSoft (http://www.statsoft.com/textbook/stfacan.html).

Torres-Reyna, Oscar. s.d. *Getting Started in Factor Analysis (using Stata).* (http://dss.princeton.edu/training/)

Triola, Mario F. 2008. *Introdução à estatística*. 10ª ed. Rio de Janeiro: LTC. Cap.10.

UCLA (http://www.ats.ucla.edu/stat/stata/output/fa_output.htm).

Vincent, Jack. 1971. *Factor Analysis in International Relations. Interpretation, Problem Areas and Application*. University of Florida Press, Gainsville.

REGRESSÃO

REGRESSÃO

– Após determinar se há ou não correlação linear entre duas variáveis, é preciso descrever a relação entre duas variáveis.

– Podemos usar gráficos e a equação da reta (equação de regressão) que melhor representa a relação.

– Com base em **valores amostrais** emparelhados, estimamos intercepto (b_0) e inclinação (b_1) e identificamos uma reta com a equação:

$$\hat{y} = b_0 + b_1 x$$

– A **verdadeira equação** de regressão é:

$$y = \beta_0 + \beta_1 x$$

– Essa é a mesma equação típica de uma reta: $y = mx + b$.

CONCEITOS BÁSICOS DE REGRESSÃO

– Há variáveis que se relacionam de maneira **determinística**, em que valor de uma variável é automaticamente dado por valor de outra variável, sem erro (ex.: custo é dado pelo preço).

– Porém, estamos interessados em modelos **probabilísticos**, em que uma variável não é completamente determinada por outra variável.

– Equação de regressão expressa relação entre x (variável explanatória, variável previsora, variável independente) e $\hat{y}$ (variável resposta, variável dependente).

– Usamos estatísticas amostrais (b_0 e b_1) para estimar os parâmetros populacionais (β_0 e β_1).

REQUISITOS SIMPLIFICADOS

– Amostra de dados emparelhados (*x*, *y*) é uma amostra aleatória de dados quantitativos.

– Exame do diagrama de dispersão mostra que pontos se aproximam do padrão de uma reta.

– Valores extremos (*outliers*) devem ser removidos se forem erros.

REQUISITOS FORMAIS

– Para cada valor fixo de *x*, os valores correspondentes de *y* têm uma distribuição que tem **forma de sino**.

– Para os diferentes valores fixados de *x*, as distribuições dos valores correspondentes de *y* têm todas a **mesma variância**.

  – Isso é violado se parte do diagrama de dispersão exibir pontos muito próximos da reta de regressão, enquanto outra parte exibir pontos muito afastados da reta.

– Para os diferentes valores fixados de *x*, as distribuições dos valores correspondentes de *y* têm **médias próximas de uma reta**.

– Os valores de *y* são **independentes**.

– Resultados **não são seriamente afetados** se afastamento da normal não for muito extremo.

DEFINIÇÕES

– Utilizando dados amostrais emparelhados, a equação de regressão descreve a relação algébrica entre duas variáveis:

$$\hat{y} = b_0 + b_1 x$$

– O gráfico da equação de regressão é a reta de regressão (reta de melhor ajuste, reta de mínimos quadrados).

Notação	Parâmetro populacional	Estatística amostral
Intercepto	β_0	b_0
Inclinação	β_1	b_1
Equação da reta	$y = \beta_0 + \beta_1 x$	$\hat{y} = b_0 + b_1 x$

– Determinando inclinação (b_1) e intercepto (b_0):

$$b_1 = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2} \qquad b_0 = \bar{y} - b_1 \bar{x}$$

OUTROS PONTOS IMPORTANTES

– A reta de regressão é a que melhor se ajusta aos dados amostrais.

– Arredonde b_1 e b_0 para três dígitos significativos.

EQUAÇÃO DE REGRESSÃO PARA PREVISÕES

– Equações de regressão podem ser úteis para prever valor de uma variável, dado algum valor de outra variável.

– Não baseie previsões em valores muito distantes dos limites dos dados amostrais.

– Se a reta de regressão se ajusta bem aos dados, faz sentido usá-la para previsões.

– Devemos usar equação da reta de regressão apenas se equação de regressão for bom modelo para dados.

OBSERVANDO A CORRELAÇÃO LINEAR

– Devemos usar a equação de regressão para previsões apenas se houver correlação linear.

– Ou seja, a adequação de usar a regressão pode ser avaliada pelo teste da significância do coeficiente de correlação linear (*r*).

– Se não há correlação linear, não usamos a equação de regressão, mas simplesmente a média amostral da variável como seu preditor.

EM SUMA...

– Na previsão de um valor de *y* com base em algum valor dado de *x*:

  – Se não há correlação linear, o melhor valor previsto de *y* é $\bar{y}$.

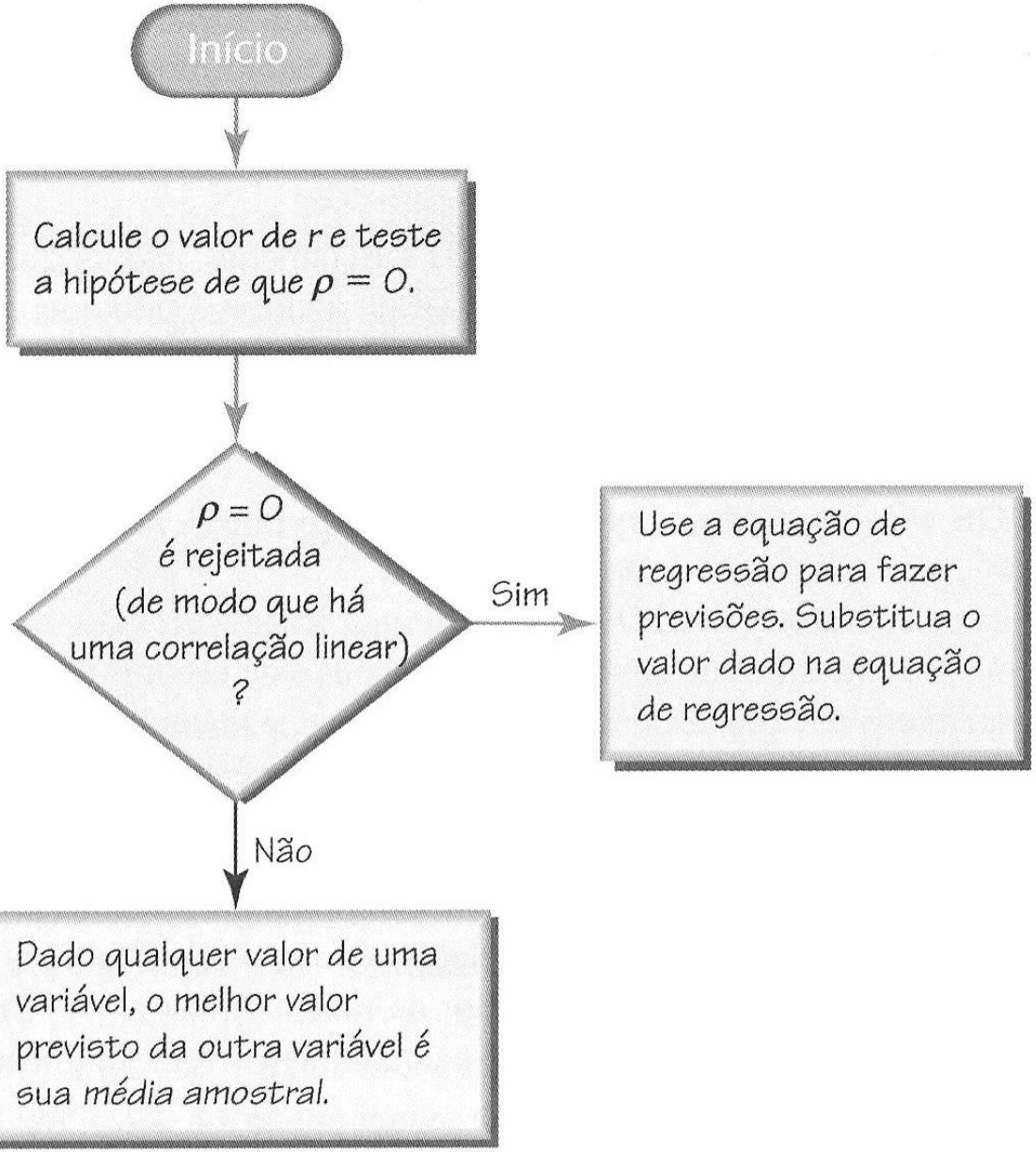
  – Se há correlação linear, melhor valor previsto de *y* é encontrado pela substituição do valor de *x* na equação de regressão.

– O coeficiente de correlação linear (*r*) é a medida de quão bem a reta de regressão se ajusta aos dados amostrais.

– Mesmo que *r* tenha um valor pequeno (0,2), a equação de regressão pode ser modelo aceitável se *r* for significativo.

– Se *r* não for significativo, equação de regressão não deve ser usada para previsões.

PROCEDIMENTO PARA PREVISÃO



DIRETRIZES PARA USO DA EQUAÇÃO DE REGRESSÃO

– Se não há qualquer correlação linear, não use a equação de regressão para fazer previsões.

– Quando usar equação de regressão para previsões, permaneça dentro do alcance dos dados amostrais disponíveis.

– Uma equação de regressão com base em dados antigos, não é necessariamente válida no momento atual.

– Não faça previsões sobre uma população que é diferente da população da qual se extraíram os dados amostrais.

MUDANÇA MARGINAL

– Ao trabalhar com duas variáveis relacionadas por uma equação de regressão, a **mudança marginal** em uma variável (y) é a quantidade que ela varia (b_1) quando outra variável (x) varia em exatamente uma unidade.

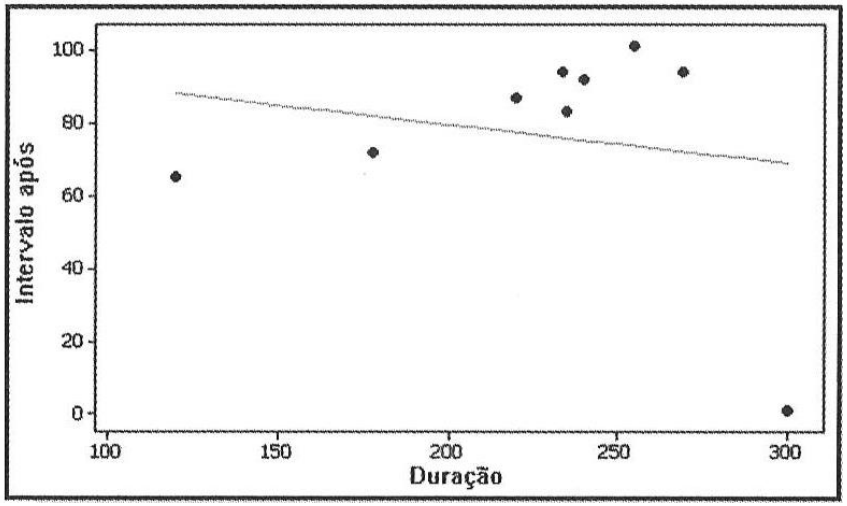
– A inclinação b_1 representa a mudança marginal em y quando x varia em uma unidade.

OUTLIERS E PONTOS INFLUENTES

– Uma análise de correlação e regressão de dados bivariados (pares) deve incluir pesquisa de valores extremos (*outliers*) e pontos influentes.

– Em um diagrama de dispersão, um ***outlier*** é um ponto que se situa muito afastado dos demais pontos amostrais.

– Dados amostrais emparelhados podem incluir um ou mais **pontos influentes**, que são pontos que afetam fortemente o gráfico da reta de regressão.



RESÍDUOS

– Há critérios para dizer que a equação de regressão representa a reta que melhor se ajusta aos dados.

– Esse critério se baseia nas distâncias verticais entre os pontos de dados originais e a reta de regressão (resíduos).

– Para uma amostra de dados emparelhados (x, y), um resíduo é a diferença ($y - \hat{y}$) entre um valor amostral y observado e o valor de $\hat{y}$, que é o valor de y predito pelo uso da equação de regressão.

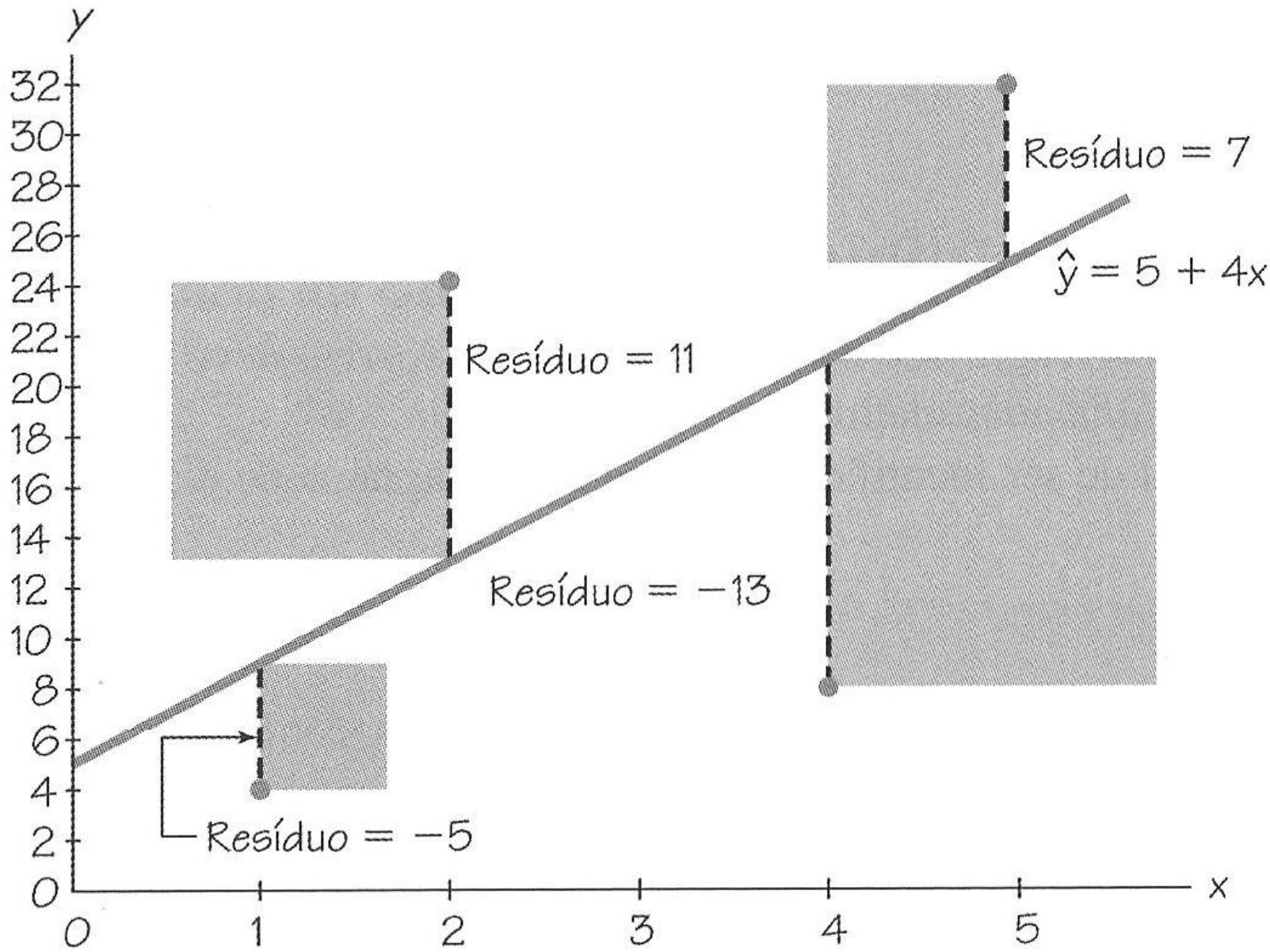
resíduo = y observado – y previsto = $y - \hat{y}$

PROPRIEDADE DOS MÍNIMOS QUADRADOS

– Uma reta satisfaz a propriedade dos mínimos quadrados se a soma dos quadrados dos resíduos é a menor possível.

– A soma das áreas dos quadrados na próxima figura é a menor soma possível.

RESÍDUOS E QUADRADOS DOS RESÍDUOS



GRÁFICOS DOS RESÍDUOS

– Gráficos de resíduos podem ser instrumento útil para:
  – Análise dos resultados da correlação e regressão.
  – Verificação dos requisitos necessários para fazer inferências sobre correlação e regressão.
– Para construir gráfico de resíduos, use o mesmo eixo *x* do diagrama de dispersão, mas use um eixo vertical para os valores dos resíduos.
– Trace uma reta horizontal passando pelo resíduo de valor 0.
– Um gráfico de resíduos é um diagrama de dispersão dos valores de (*x*, *y*) depois que cada um dos valores da coordenada *y* tiver sido substituído pelo valor do resíduo ($y - \hat{y}$).
– Ou seja, é um gráfico dos pontos (x, $y - \hat{y}$).

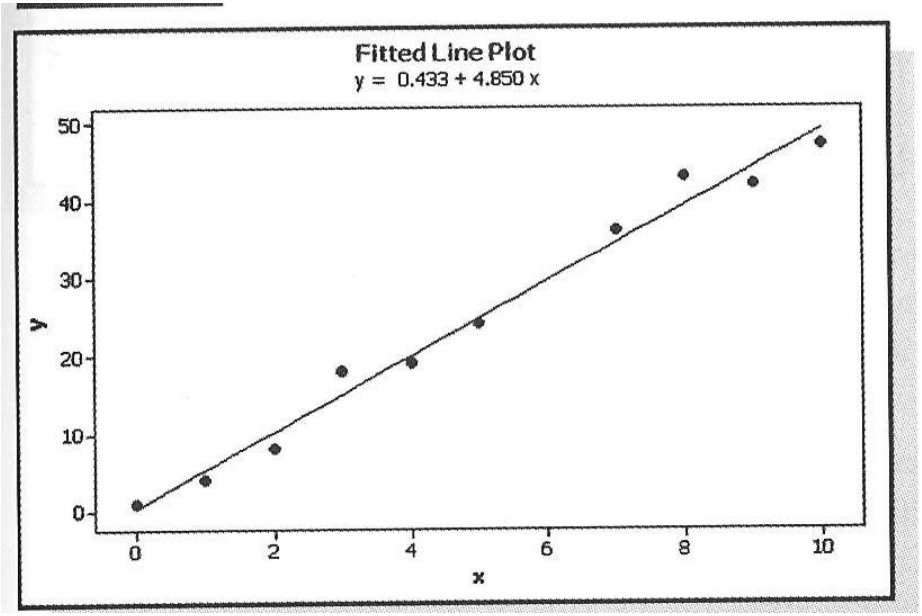
ANÁLISE DOS GRÁFICOS DOS RESÍDUOS

– Se o gráfico de resíduos não revela qualquer padrão, a equação de regressão é uma boa representação da associação entre as duas variáveis.

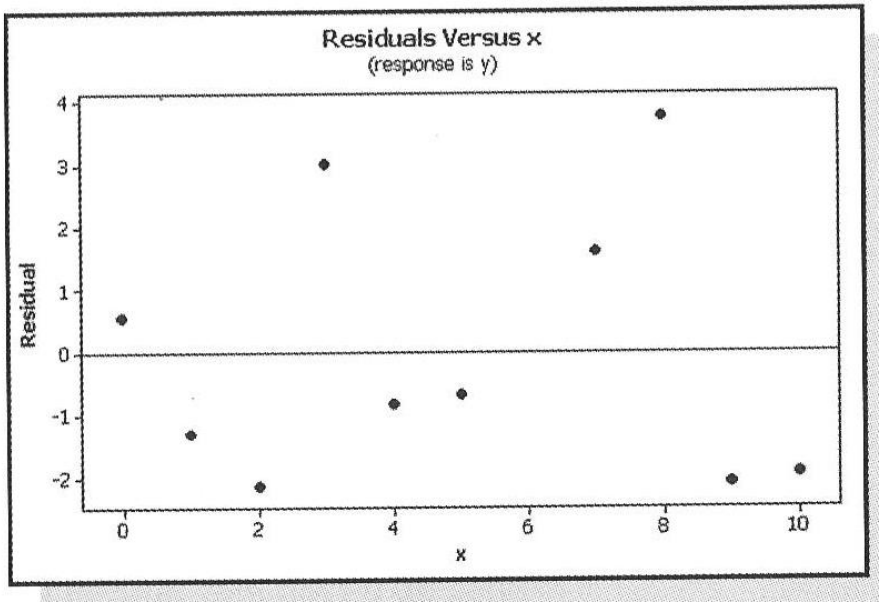
– Se o gráfico de resíduos revela algum padrão sistemático, a equação de regressão não é uma boa representação da associação entre as duas variáveis.

EXEMPLOS

– Reta de regressão se ajusta bem aos dados.

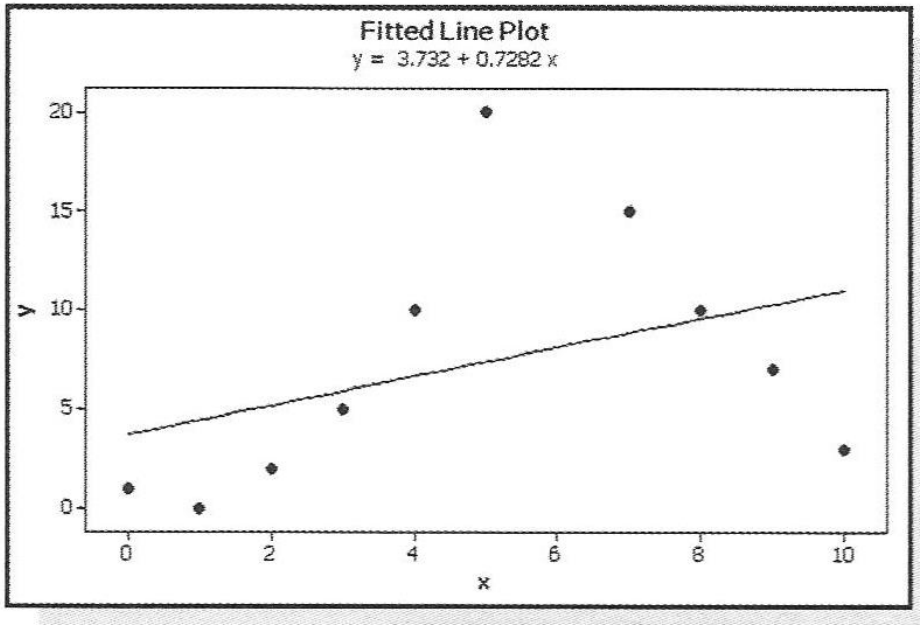


– Gráfico dos resíduos não revela qualquer padrão.

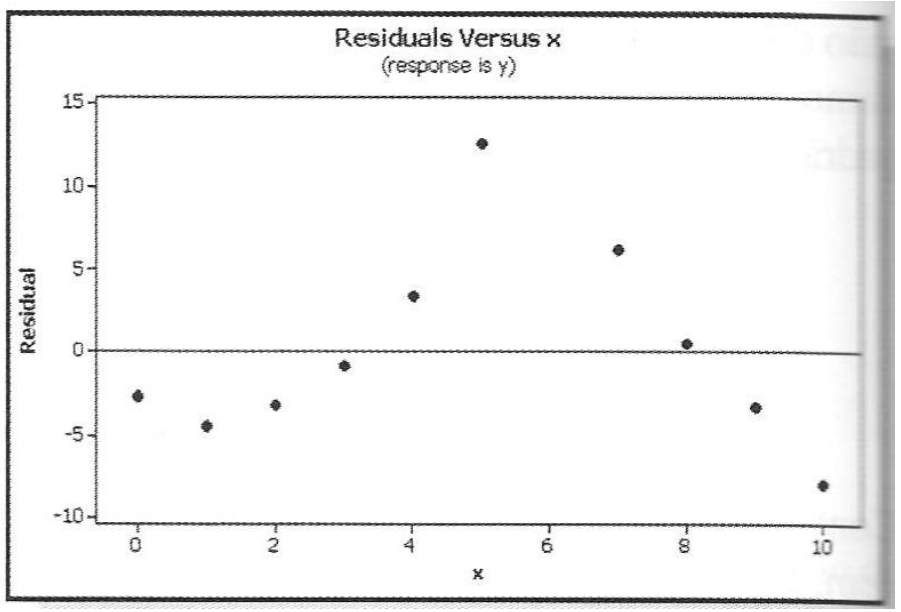


EXEMPLOS

– Diagrama de dispersão mostra que associação não é linear.



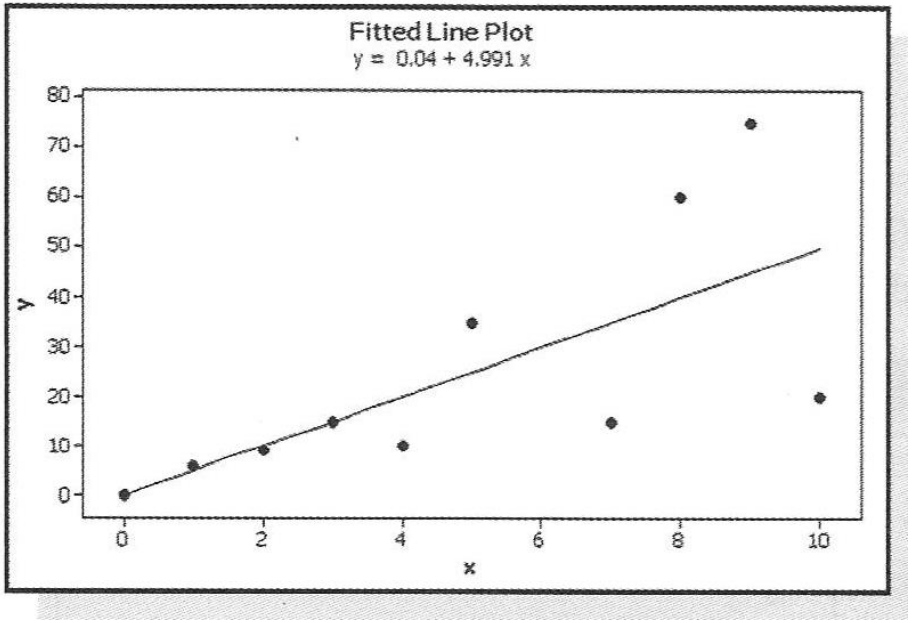
– Gráfico dos resíduos exibe um padrão distinto (não linear).

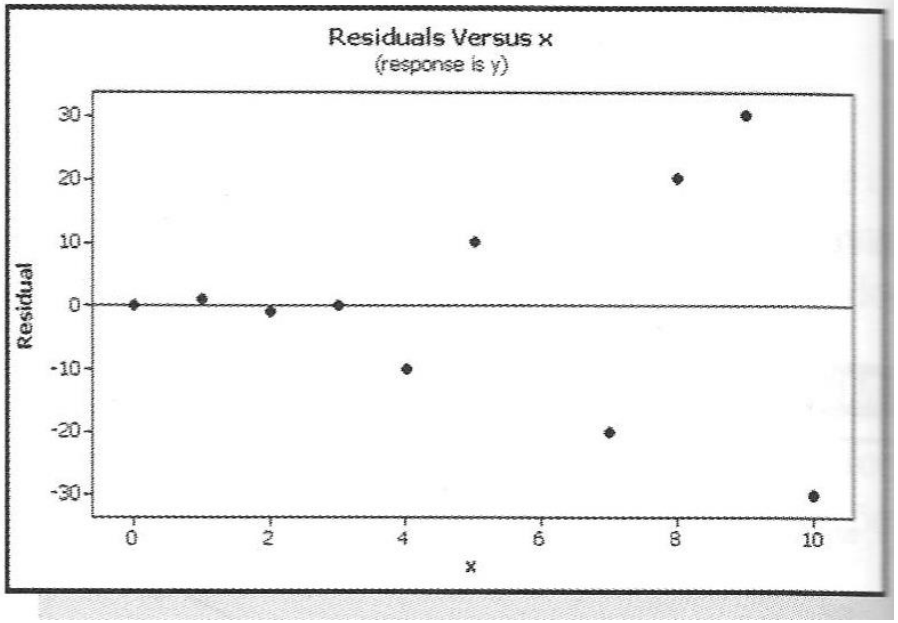


EXEMPLOS

– Diagrama de dispersão exibe variação crescente dos pontos em relação à reta de regressão.

– No gráfico dos resíduos, pontos exibem maior dispersão indo da esquerda para a direita.





– Isso viola requisito de que, para diferentes valores de *x*, distribuição dos valores de *y* tem mesma variância.

VARIAÇÃO E INTERVALOS DE PREVISÃO

VARIAÇÃO E INTERVALOS DE PREVISÃO

– Veremos a variação que pode ser explicada e que não pode ser explicada pela correlação linear entre *x* e *y*.

– Em seguida, construiremos um intervalo de previsão, que é uma estimativa intervalar para o valor previsto de *y*:

  – Estimativas de intervalos de parâmetros são chamados de **intervalos de confiança**.

  – Estimativas de intervalos de variáveis são chamados de **intervalos de previsão**.

DESVIOS TOTAL, EXPLICADO E NÃO-EXPLICADO

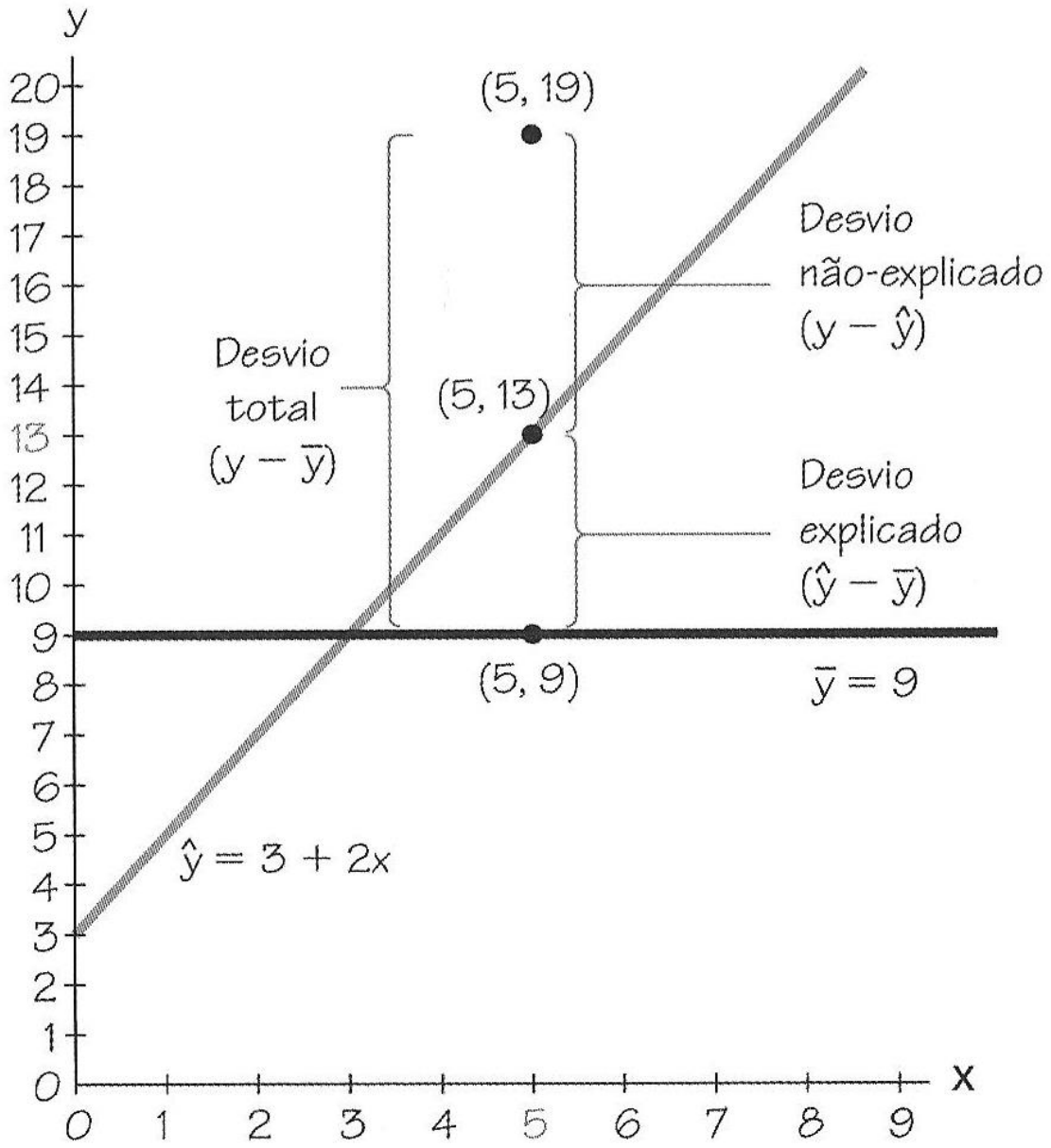
– Suponha que tenhamos um conjunto de pares de dados com o ponto amostral (x, y), que $\hat{y}$ seja o valor previsto de y (obtido pelo uso da equação de regressão) e que a média dos valores amostrais de y seja $\bar{y}$.

– **Desvio total** de (x, y) é a distância vertical $y - \bar{y}$, que é a distância entre o ponto (x, y) e a reta horizontal que passa pela média amostral.

– **Desvio explicado** de (x, y) é a distância vertical $\hat{y} - \bar{y}$, que é a distância entre o valor previsto de y e a reta horizontal que passa pela média amostral.

– **Desvio não-explicado (resíduo)** é a distância vertical $y - \hat{y}$, que é a distância vertical entre o ponto (x, y) e a reta de regressão.

DESVIOS TOTAL, EXPLICADO E NÃO-EXPLICADO



VARIÂNCIAS TOTAL, EXPLICADA E NÃO-EXPLICADA

(desvio total) = (desvio explicado) + (desvio não-explicado)

$$(y - \bar{y}) = (\hat{y} - \bar{y}) + (y - \hat{y})$$

– Se somarmos os quadrados dos desvios usando todos os pontos (x, y), obteremos quantidades de variação.

– A **variância total** se expressa como a soma dos quadrados dos valores do desvio total.

– A **variância explicada** é a soma dos quadrados dos valores do desvio explicado.

– A **variância não-explicada** é a soma dos quadrados dos valores do desvio não explicado.

COEFICIENTE DE DETERMINAÇÃO

– Lembremos que o valor de r^2 é a proporção em *y* que pode ser explicada pela relação linear entre *x* e *y*.

– Este coeficiente de determinação é então a quantidade de variação em *y* que é explicada pela reta de regressão.

$$r^2 = \frac{variação\ explicada}{variação\ total} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

INTERVALOS DE PREVISÃO

– Sabemos que estimativas pontuais têm a séria desvantagem de não fornecerem qualquer informação sobre o nível de precisão.

– Usamos os **intervalos de confiança** para estimar intervalos de parâmetros.

– Agora usaremos **intervalos de previsão** para estimar intervalos de uma variável (valor previsto de *y*).

– O desenvolvimento de um intervalo de previsão requer uma medida da dispersão dos pontos amostrais em torno da reta de regressão.

ERRO PADRÃO DA ESTIMATIVA

– Erro padrão da estimativa é uma medida da dispersão dos pontos amostrais em torno da reta de regressão.

– É utilizado o desvio não-explicado (resíduo).

– O erro padrão da estimativa (s_e) é uma medida das diferenças (distâncias) entre os valores amostrais de *y* observados e os valores previstos $\hat{y}$ que são obtidos com o uso da reta de regressão.

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}} \qquad s_e = \sqrt{\frac{\sum y^2 - b_0 \sum y - b_1 \sum xy}{n - 2}}$$

DESVIO PADRÃO E ERRO PADRÃO DA ESTIMATIVA

– O **desvio padrão** é uma medida de como os valores se afastam de sua média.

– O **erro padrão da estimativa** (s_e) é uma medida de como os pontos amostrais se afastam de sua reta de regressão.

– Valores de s_e relativamente menores refletem pontos que permanecem mais próximos da reta de regressão.

– Valores relativamente maiores ocorrem com pontos mais afastados da reta de regressão.

INTERVALO DE PREVISÃO PARA *y* INDIVIDUAL

– Dado o valor fixo x_0, o intervalo de previsão para um *y* individual é:

$$\hat{y} - E < y < \hat{y} + E$$

– A margem de erro (*E*) é:

$$E = t_{\alpha/2} s_e \sqrt{1 + \frac{1}{n} + \frac{n(x_0 - \bar{x})^2}{n(\sum x^2) - (\sum x)^2}}$$

– Em que:

  – x_0 representa o valor dado de *x*.

  – $t_{\alpha/2}$ tem $n - 2$ graus de liberdade.

  – s_e é encontrado pela fórmula apresentada anteriormente.

REGRESSÃO MÚLTIPLA

REGRESSÃO MÚLTIPLA

– Trataremos de um método para análise de uma relação linear que envolve mais de duas variáveis.

– Mais especificamente, serão abordados:

  – Equação de regressão múltipla.

  – Valor do R^2 ajustado.

  – Valor *P*.

EQUAÇÃO DE REGRESSÃO MÚLTIPLA

– Uma equação de regressão múltipla expressa uma relação linear entre uma variável dependente (y) e duas ou mais variáveis previsoras (x_1, x_2, ..., x_k).

– Forma geral da equação de regressão múltipla estimada:

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

NOTAÇÃO

$$\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

– n = tamanho amostral

– k = número de variáveis independentes

– $\hat{y}$ = valor previsto de y, calculado com equação de regressão

– x_1, x_2, ..., x_k = variáveis independentes

– β_0 = parâmetro populacional que indica intercepto y (valor de y quando todos x_k são zero)

– b_0 = estimativa amostral de β_0

– β_1, β_2,..., β_k = são coeficientes das variáveis x_1, x_2,..., x_k

– b_1, b_2,..., b_k = são estimativas amostrais de β_1, β_2,..., β_k

ERRO ALEATÓRIO

– Para qualquer conjunto específico de valores de x, a equação de regressão está associada a um erro aleatório (ε).

– Admitimos que estes erros:

  – São distribuídos normalmente.

  – Possuem média zero.

  – Possuem desvio padrão de σ.

  – São independentes das variáveis do modelo.

COEFICIENTE DE DETERMINAÇÃO MÚLTIPLA (R^2)

– R^2 é o **coeficiente de determinação múltipla**:

  – Mede o quão bem a equação de regressão múltipla se ajusta aos dados amostrais.

  – Indica a proporção de variação em *y* que pode ser explicada pela variação em x_1, x_2, ..., x_k.

  – R^2 = 1: significa ajuste perfeito.

  – R^2 próximo de 1: ajuste muito bom.

  – R^2 próximo de 0: ajuste muito ruim.

– Na medida em que mais variáveis são incluídas, R^2 cresce.

– O maior R^2 é obtido pela inclusão de todas variáveis disponíveis, mas esta não é a melhor equação de regressão.

COEFICIENTE DE DETERMINAÇÃO AJUSTADO

– Como o R^2 sempre aumenta com a inclusão de variáveis, a comparação de diferentes equações de regressão múltipla é realizada com o ***R*² ajustado** pelo número de variáveis e tamanho amostral:

$$R^2 ajustado = 1 - \frac{(n-1)}{[n-(k+1)]}(1-R^2)$$

– Em que:

  – n = tamanho amostral.

  – k = número de variáveis independentes (x).

OBSERVAÇÕES IMPORTANTES

– O R^2 ajustado auxilia na escolha de modelo sem variáveis independentes redundantes (entre modelos não-aninhados).

– Comparação dos R^2 ajustados pode ser feita para optar entre modelos com formas funcionais diferentes das variáveis independentes:

$$y = \beta_0 + \beta_1 \log(x) + u$$

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + u$$

– Não podemos usar nem o R^2 nem o R^2 ajustado para escolher entre modelos não-aninhados com diferentes formas funcionais da variável dependente.

– Os R^2 medem a proporção explicada do total da variação de qualquer variável dependente.

  – Portanto, diferentes funções da variável dependente terão diferentes montantes de variação a serem explicados.

VALOR *P*

– O valor *P* é uma medida da significância global da equação de regressão múltipla.

– A hipótese nula testada é (H_0: $\beta_1 = \beta_2 = ... = \beta_k = 0$).

– O valor *P* indica a probabilidade de H_0 não ser rejeitada:

  – Se valor *P* for pequeno (<0,05), rejeitamos H_0, o que implica: (1) pelo menos um dos betas não é zero; e (2) a equação de regressão é eficaz na determinação de *y*.

  – Se valor *P* for pequeno, dizemos que a equação de regressão múltipla tem boa significância geral e é adequada para previsões.

– Assim como o R^2 ajustado, o valor *P* é uma boa medida de quão bem a equação se ajusta aos dados amostrais.

DIRETRIZES PARA DETERMINAR MELHOR EQUAÇÃO

- Utilize teoria, hipóteses e estudos anteriores para incluir ou excluir variáveis.
- Considere o valor *P*.
- Considere equações com altos valores de R^2 ajustado e tente incluir poucas variáveis:
  - Não inclua variáveis que não aumentam R^2 ajustado substancialmente.
  - Para um dado número de variáveis independentes, escolha o modelo com maior R^2 ajustado.
  - Se duas variáveis independentes possuem alta correlação linear entre si, não há necessidade de incluir ambas na regressão.

REGRESSÃO PASSO A PASSO (*STEPWISE*)

– Há alguns problemas com a regressão passo a passo:

  – Não resultará necessariamente no melhor modelo, se algumas variáveis independentes forem altamente correlacionadas.

  – Pode resultar em valores inflacionados de R^2.

  – **Não pensamos sobre o problema.**

VARIÁVEIS *DUMMY* E REGRESSÃO LOGÍSTICA

– Muitas aplicações usam variável dicotômica (*dummy*), que assume apenas dois possíveis valores discretos.

– Geralmente representamos estes valores por 0 (fracasso) e 1 (sucesso).

– Se incluirmos uma variável *dummy* como variável independente, podemos usar os métodos anteriores:

  – O coeficiente desta variável indicará a diferença no valor de *y*, quando obtemos sucesso, em relação ao fracasso.

– Se a variável *dummy* for a variável resposta (*y*), devemos usar regressão logística.

REGRESSÃO LOGÍSTICA

– Se a variável dependente é binária, temos esta expressão na regressão logística:

$$ln\left(\frac{p}{1-p}\right) = b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_k x_k$$

– Nesta expressão, *p* representa uma probabilidade.

– Um valor de *p*=0 indica que obtivemos fracasso.

– Um valor de *p*=1 indica que obtivemos sucesso.

– Um valor de *p*=0,2 indica que há chance de 0,2 de obter sucesso e chance de 0,8 de obter fracasso.

MODELAGEM

MODELAGEM

– É importante realizar ajustes no modelo de regressão para que ele se ajuste aos dados do mundo real.

– Não devemos ficar restritos a modelos lineares:

  – Linear: $y = a + bx$

  – Quadrática: $y = ax^2 + bx + c$

  – Logarítmica: $y = a + b\ ln(x)$

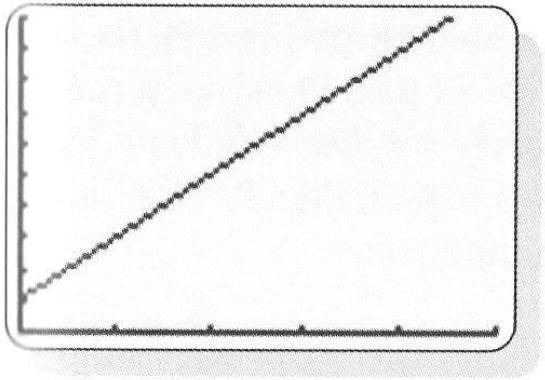
  – Exponencial: $y = ab^x$

  – Potência: $y = ax^b$

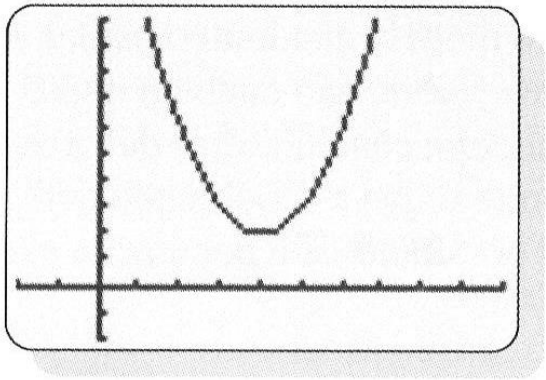
– Em vez de amostras aleatórias, podemos considerar dados coletados ao longo do tempo (séries temporais).

GRÁFICOS DE MODELOS MATEMÁTICOS

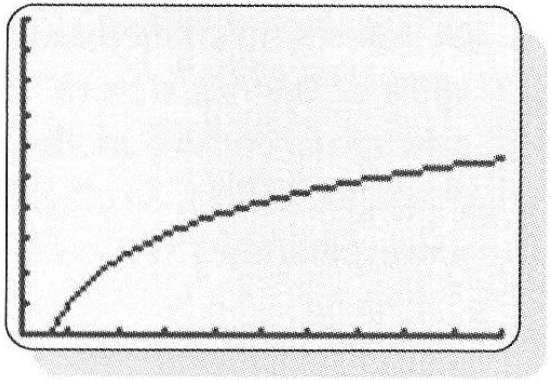
Linear: $y = 1 + 2x$



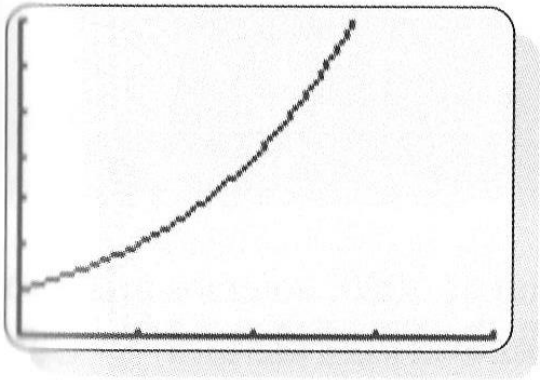
Quadrática: $y = x^2 - 8x + 18$



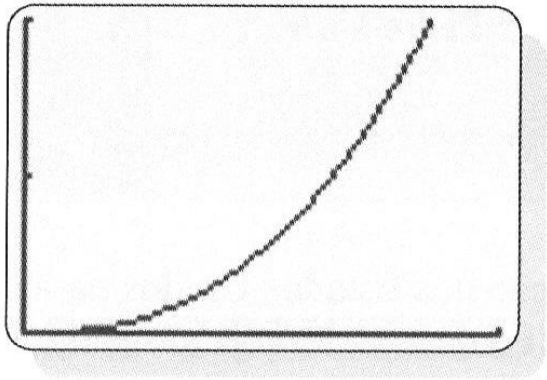
Logarítmica: $y = 1 + 2 \ln x$



Exponencial: $y = 2^x$



Potência: $y = 3x^{2,5}$



ESCOLHA DO MODELO

– O modelo selecionado depende da natureza dos dados:

  – Procure um **padrão no gráfico**: com um diagrama de dispersão entre *x* e *y*, selecione um modelo que se ajuste razoavelmente aos pontos observados.

  – Ache e compare **valores de R^2**: diminua número de modelos possíveis e selecione funções com maiores R^2 (já que indicam melhor ajuste aos pontos observados).

  – **Pense**: use o modelo para calcular valores futuros, passados e para datas omitidas, observando se resultados são realistas.

  – “A melhor escolha de um modelo depende do conjunto de dados que está sendo analisado e requer um **exercício de julgamento**, não apenas computacional.”