

AULA 09

Regressão

Ernesto F. L. Amaral

17 de setembro de 2012

Faculdade de Filosofia e Ciências Humanas (FAFICH)
Universidade Federal de Minas Gerais (UFMG)

Fonte:

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

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:

$$\hat{y} = b_0 + b_1x$$

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

$$y = \beta_0 + \beta_1x$$

- 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 previsor, 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:

$$\hat{y} = b_0 + b_1x$$

- 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	$y = \beta_0 + \beta_1x$	$\hat{y} = b_0 + b_1x$

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

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

$$b_0 = \bar{y} - b_1\bar{x}$$

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.

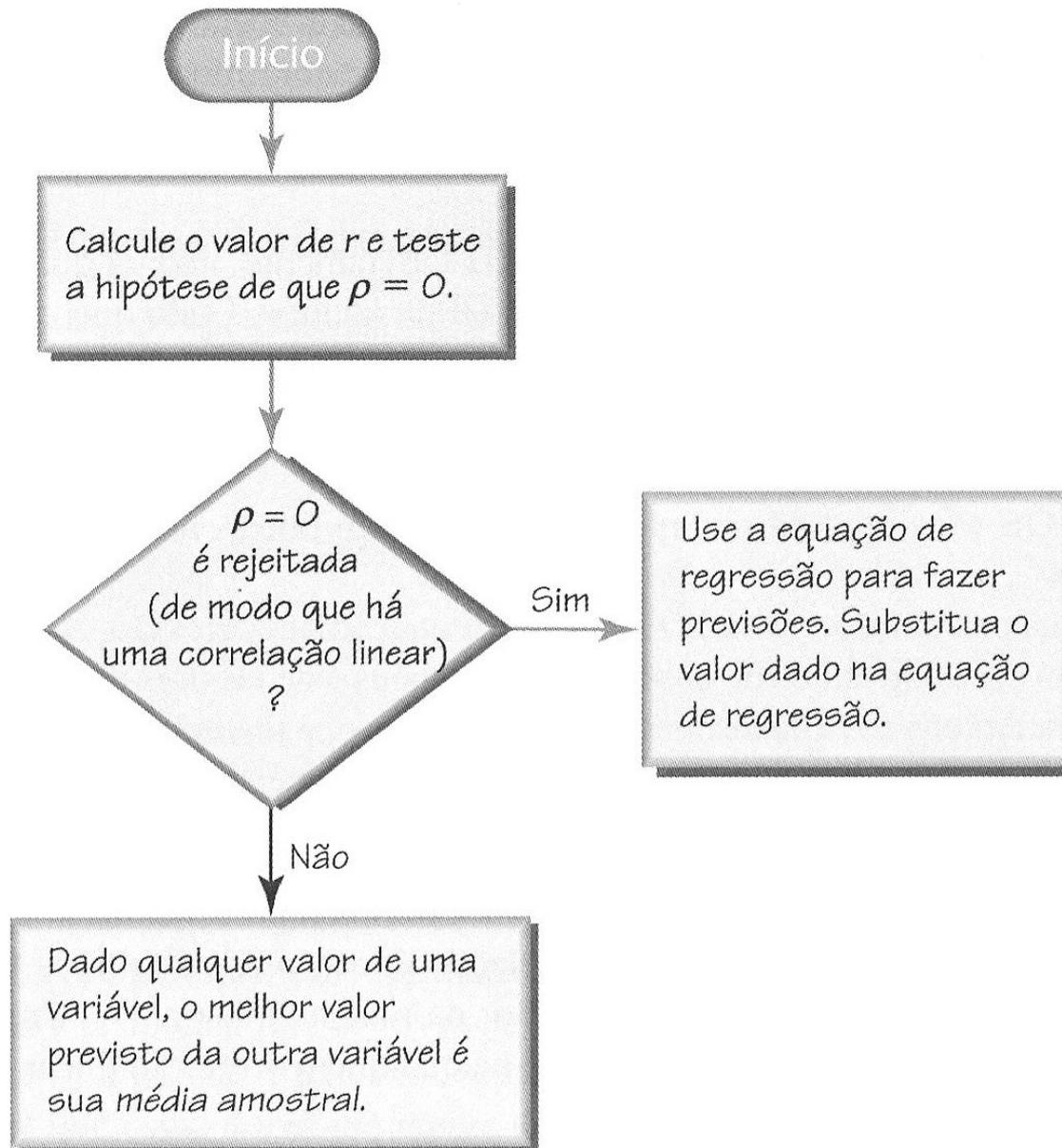
OBSERVANDO A CORRELAÇÃO LINEAR

- 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

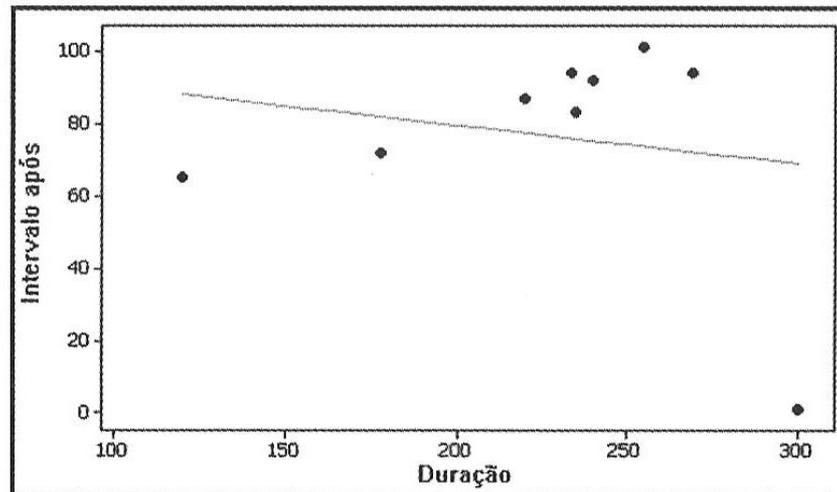
- Se não há qualquer correlação linear, não use a equação de regressão para fazer previsões.
- Quando usar equação de regressão para previsões, permaneça dentro do alcance dos dados amostrais disponíveis.
- Uma equação de regressão com base em dados antigos, não é necessariamente válida no momento atual.
- 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

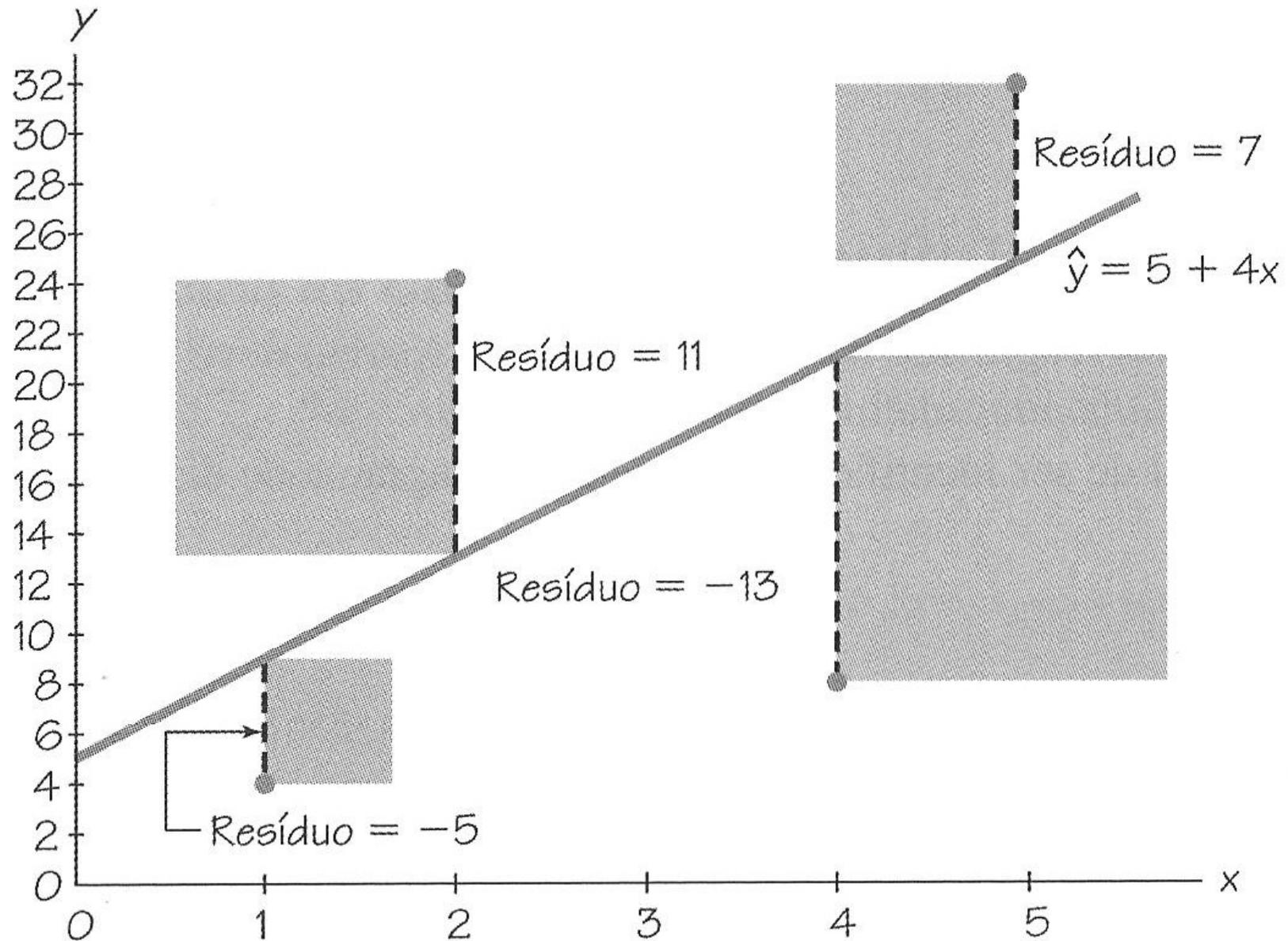
- Há critérios para dizer que a equação de regressão representa a reta que melhor se ajusta aos dados.
- Esse critério se baseia nas distâncias verticais entre os pontos de dados originais e a reta de regressão (resíduos).
- Para uma amostra de dados emparelhados (x, y) , um resíduo é a diferença $(y - \hat{y})$ entre um valor amostral y observado e o valor de \hat{y} , que é o valor de y previsto pelo uso da equação de regressão.

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

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

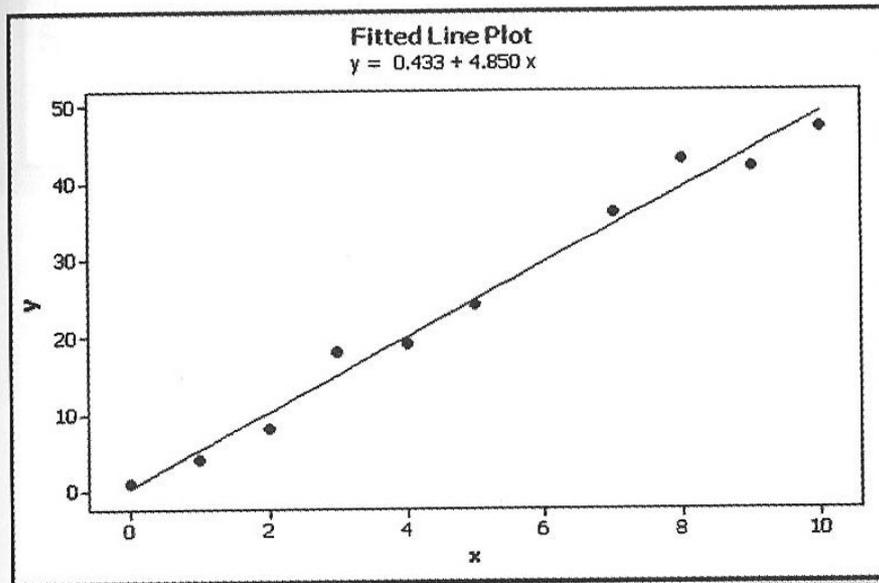
- Gráficos de resíduos podem ser instrumento útil para:
 - Análise dos resultados da correlação e regressão.
 - Verificação dos requisitos necessários para fazer inferências sobre correlação e regressão.
- Para construir gráfico de resíduos, use o mesmo eixo x do diagrama de dispersão, mas use um eixo 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

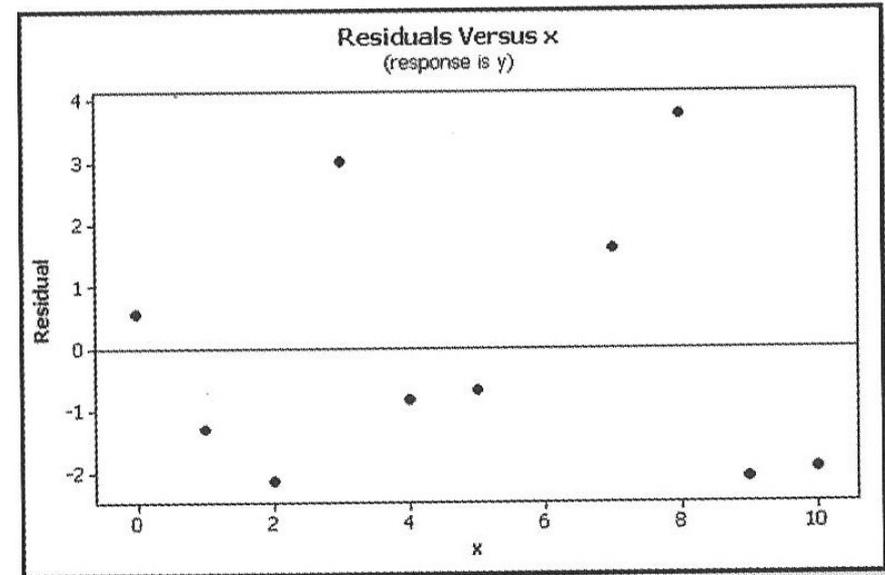
- Se o gráfico de resíduos não revela qualquer padrão, a equação de regressão é uma boa representação da associação entre as duas variáveis.
- Se o gráfico de resíduos revela algum padrão sistemático, a equação de regressão não é uma boa representação da associação entre as duas variáveis.

EXEMPLOS

– Reta de regressão se ajusta bem aos dados.

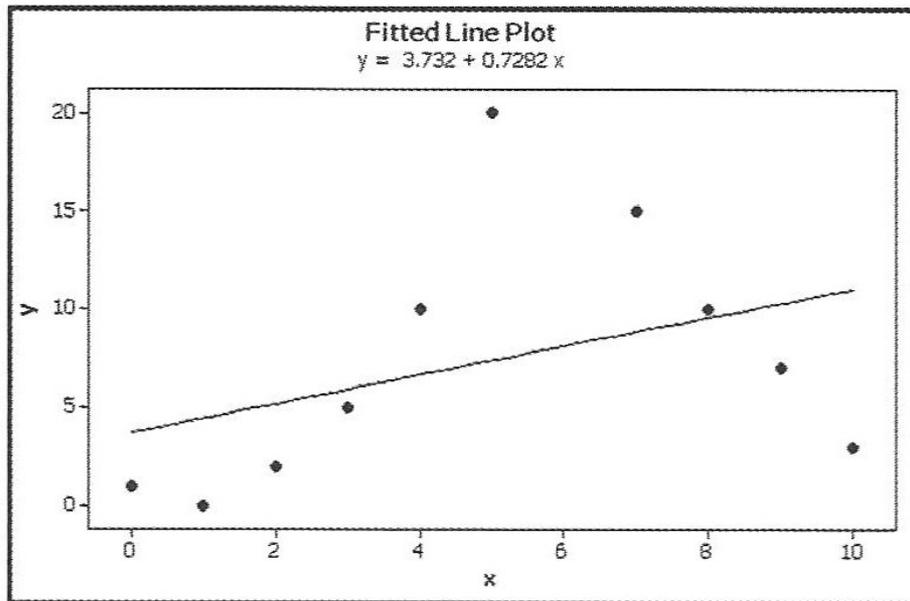


– Gráfico dos resíduos não revela qualquer padrão.

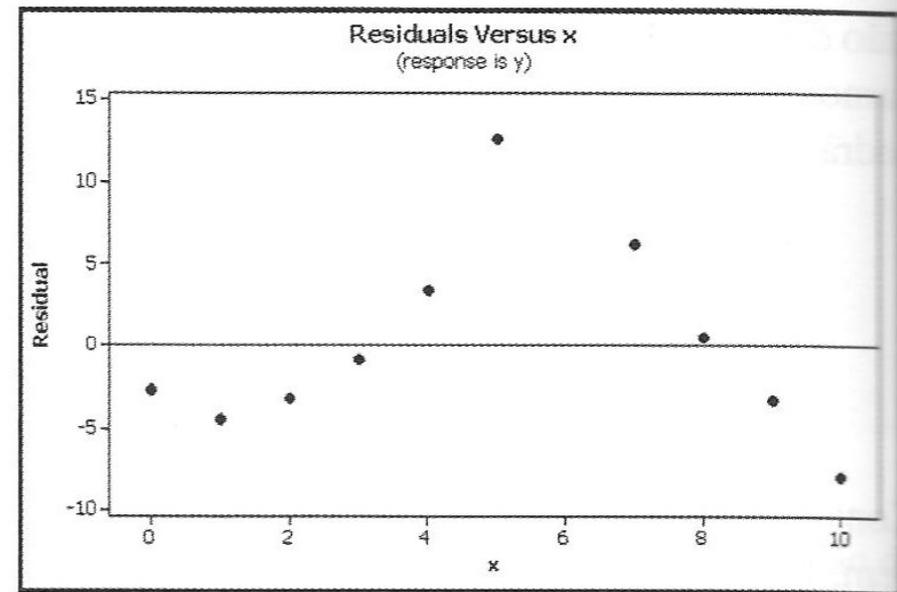


EXEMPLOS

- Diagrama de dispersão mostra que associação não é linear.

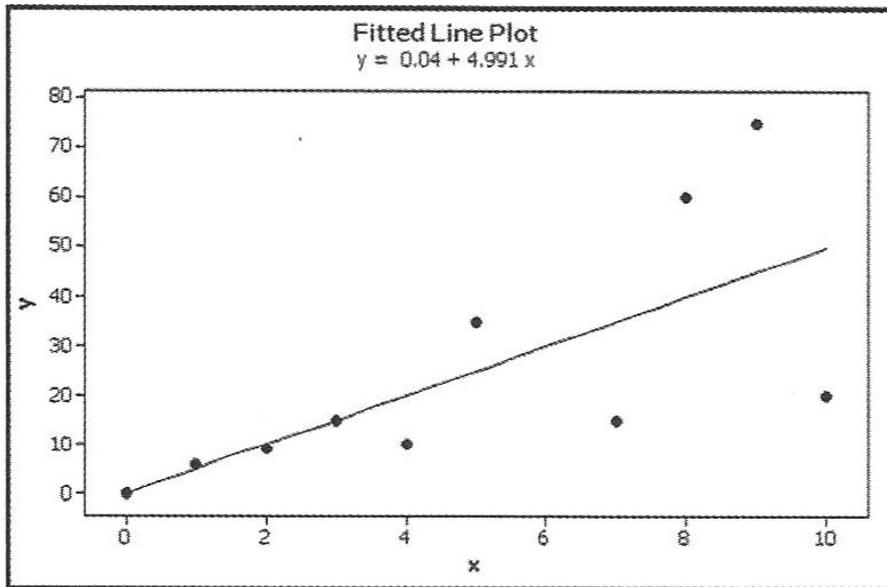


- Gráfico dos resíduos exibe um padrão distinto (não linear).

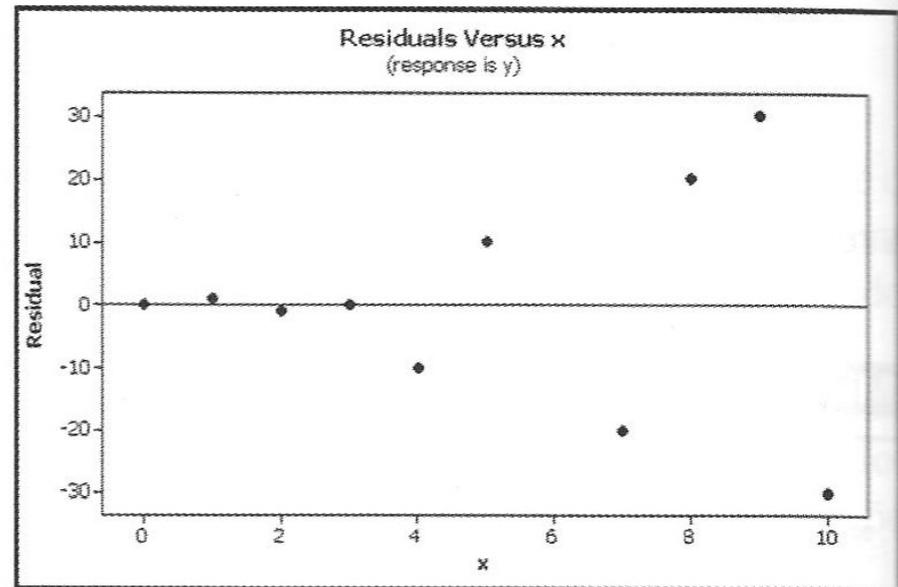


EXEMPLOS

- 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

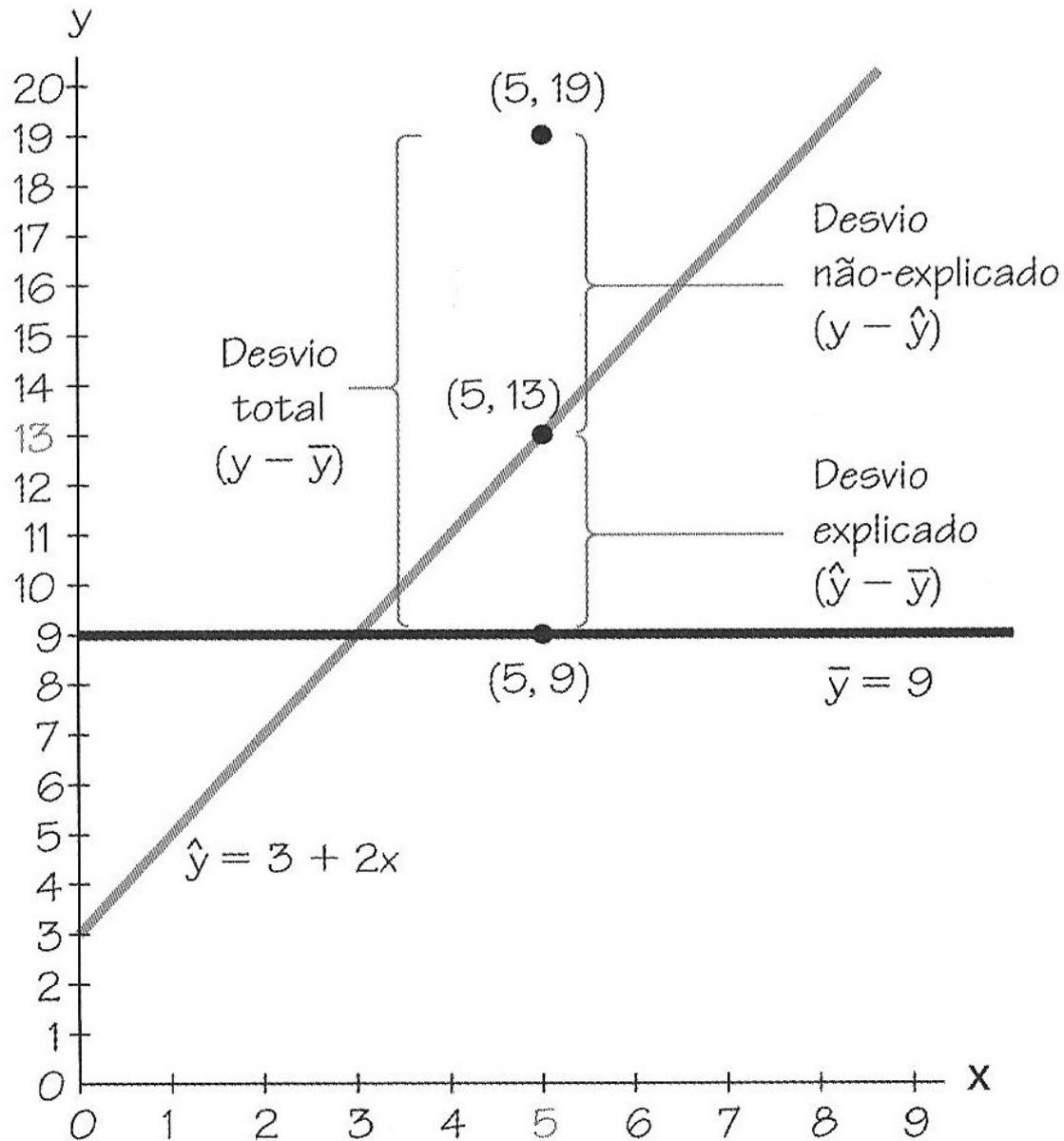
VARIAÇÃO E INTERVALOS DE PREVISÃO

- 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.

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

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.

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

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

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, \dots, x_k).
- Forma geral da equação de regressão múltipla estimada:

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

NOTAÇÃO

$$\hat{y} = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

- n = tamanho amostral
- k = número de variáveis independentes
- \hat{y} = valor previsto de y , calculado com equação de regressão
- x_1, x_2, \dots, 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
- $\beta_1, \beta_2, \dots, \beta_k$ = são coeficientes das variáveis x_1, x_2, \dots, x_k
- b_1, b_2, \dots, b_k = são estimativas amostrais de $\beta_1, \beta_2, \dots, \beta_k$

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, \dots, 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 ($H_0: \beta_1 = \beta_2 = \dots = \beta_k = 0$).
- 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:

$$\ln\left(\frac{p}{1-p}\right) = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$$

- 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.

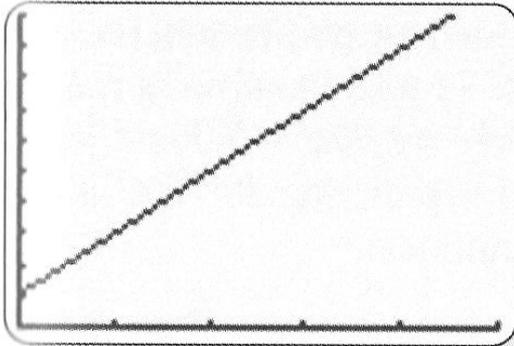
MODELAGEM

MODELAGEM

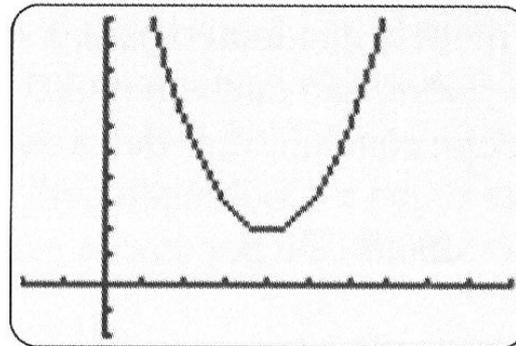
- É importante realizar ajustes no modelo de regressão para que ele se ajuste aos dados do mundo real.
- Não devemos ficar restritos a modelos lineares:
 - Linear: $y = a + bx$
 - Quadrática: $y = ax^2 + bx + c$
 - Logarítmica: $y = a + b \ln(x)$
 - Exponencial: $y = ab^x$
 - Potência: $y = ax^b$
- Em vez de amostras aleatórias, podemos considerar dados coletados ao longo do tempo (séries temporais).

GRÁFICOS DE MODELOS MATEMÁTICOS

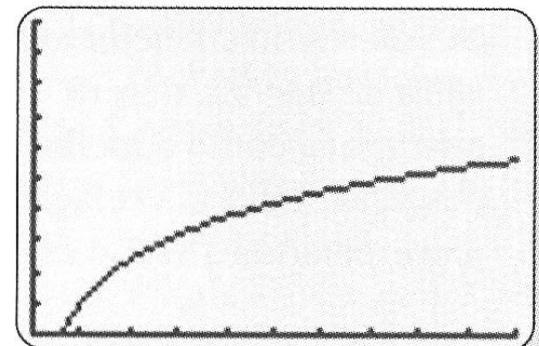
Linear: $y = 1 + 2x$



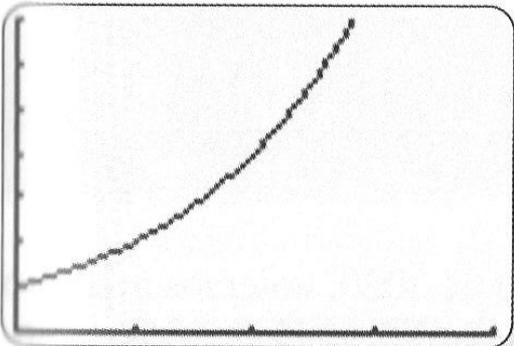
Quadrática: $y = x^2 - 8x + 18$



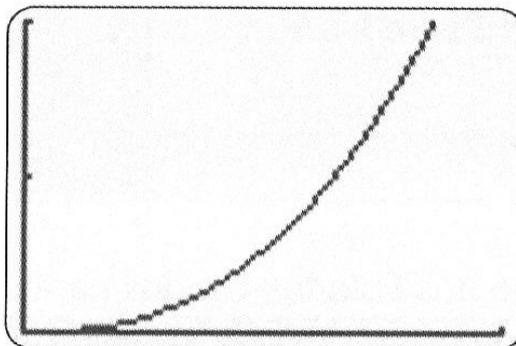
Logarítmica: $y = 1 + 2 \ln x$



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.”