Pesquisa Quantitativa

Lupércio França Bessegato
Mestrado em Administração/UFJF

Regressão e Correlação

Roteiro Geral

1. Introdução
2. Coleta de dados
3. Modelos probabilísticos
4. Distribuições amostrais e estimação
5. Testes de significância
6. Comparações de médias
7. Tabelas de contagem
8. **Regressão e correlação**
9. Referências

Roteiro do Módulo

8. Regressão e Correlação:
   a) Modelagem de relação
   b) Correlação
   c) Regressão linear simples
   d) Regressão linear múltipla
   e) Verificação do modelo

Modelagem de Relação

Relação entre Variáveis

- Relação entre duas variáveis:
  - √ Quantitativa vs. qualitativa
  - √ Quantitativa vs. quantitativa
  - √ Qualitativa vs. qualitativa
- Uso de figuras para transmitir informação sobre as principais características da relação

Relação entre Variáveis

- Verificação visual – diagrama de dispersão:
  - √ Apresenta forma, direção e força da relação entre duas variáveis quantitativas
- Relações lineares
  - √ São particularmente importantes
    - – Padrão simples e comum
- Relação linear simples
  - √ Pontos se situam próximos de uma linha reta

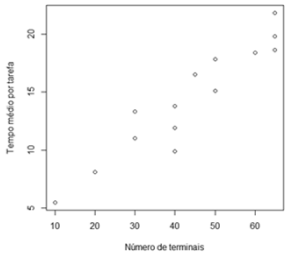
Representação Gráfica dos Dados

- Gráfico de dispersão
  - √ Exploração de relações entre variáveis quantitativas

Exemplo

- Tempos de processamento e cargas de trabalho no computador
  - √ Pesquisa sobre tempos de processamento de usuários de sistema multiusuário
  - √ Quando mais pessoas utilizam o sistema, maior o tempo de execução de uma tarefa particular
  - √ Variáveis:
    - Número de terminais
    - Tempo médio por tarefa

- Diagrama de dispersão:



  - √ Há uma tendência (padrão)
    - Não é uma relação exata
  - √ Há uma dispersão em torno da tendência

Modelagem de Relação

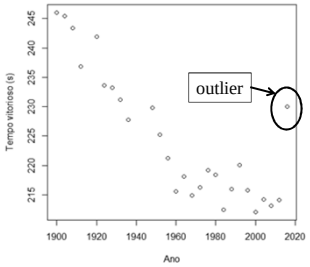
- Regressão:
  - √ Predição ou explicação do comportamento de uma variável em termos do comportamento de outra
- Variável resposta (Y)
  - √ Variável cujo comportamento queremos predizer ou explicar
- Variáveis explicativas (X's)
  - √ Variáveis que ajudam a entender, explicar ou predizer o comportamento de Y
  - √ Em geral, são controladas ou não aleatórias

- Questão para o exemplo anterior:
  - √ Como o tempo por tarefa se altera quando mudamos a quantidade de terminais?
    - Variável resposta (Y): tempo médio por tarefa
    - Variável explicativa (X): número de terminais

Exemplo

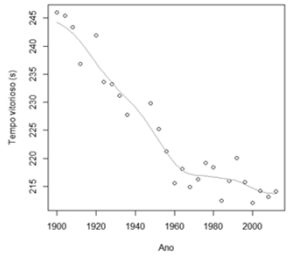
- Corrida de 1.500 m para homens
  - √ Tempos dos vencedores dos Jogos Olímpicos de 1900 a 2016
  - √ Variáveis:
    - ano:
    - tempo: marca vitoriosa (em seg.)

- Diagrama de dispersão



  - √ Marcas vitoriosas diminuem conforme ano aumenta
  - √ Tendência:
    - Linear até a década de 70, depois estabiliza

- Diagrama de dispersão – com suavização
  - √ Sem o *outlier* (2016)



  - √ Muda percepção da tendência decrescente
    - Muda de intensidade na década de 70

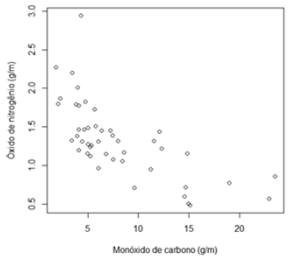
- Questão
  - √ Como a marca vitoriosa muda com o ano?
    - Variável resposta (Y): tempo (s)
    - Variável explicativa (X): ano

Exemplo

- Poluentes em emissões de gases
  - √ Estudo sobre poluentes dos gases de escapamentos emitidos por automóveis
  - √ Amostra com 46 veículos idênticos
  - √ Variáveis:
    - HC: hidrocarboneto
    - CO: monóxido de carbono
    - NOX: óxidos de nitrogênio

17

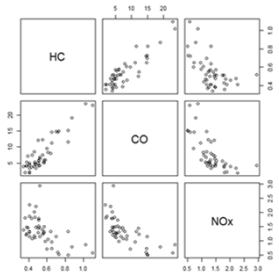
- Diagrama de dispersão



- √ Há um padrão não muito forte
- √ Veículos que emitem um nível baixo de CO tendem a emitir níveis relativamente altos de NOx

18

- Diagramas de dispersão de todas as variáveis



- √ Relação entre CO e HC é forte e positiva
- √ Todas as 3 medidas são aleatórias (não controladas)

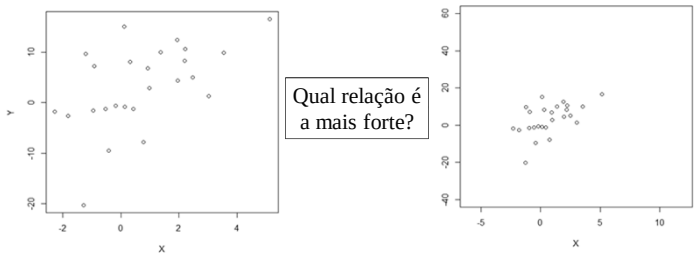
19

- Questão:
  - √ Como predizer os níveis de emissão de óxido de nitrogênio a partir de seus níveis de emissão de monóxido de carbono?
    - Variável resposta (Y) = Nox
    - Variável explicativa (X) = CO

20

Importante

- Verificação visual pode não ser confiável:



√ Gráficos do mesmo conjunto de dados
√ Padrão linear da esquerda aparenta mais força
– Pontos cercados por espaço vazio

Tendência e Dispersão

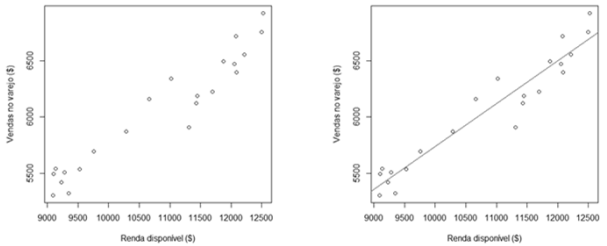
- Tendência:
  √ Padrão observado no diagrama de dispersão
- Dispersão:
  √ Flutuação dos dados em torno da tendência
  √ Às vezes é tão grande que impede a visualização de tendência

Relação de Regressão

Relação de Regressão	=	tendência + dispersão residual

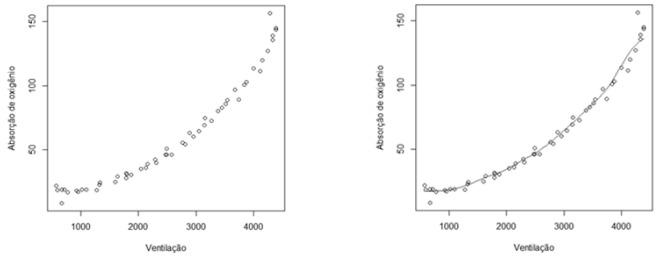
√ Às vezes a variação em torno da tendência é pequena é a relação é boa para predição

- Vendas no varejo vs. Renda disponível:



√ Aparenta ter tendência linear
√ Dispersão homogênea
– Flutuação não depende do valor da Renda

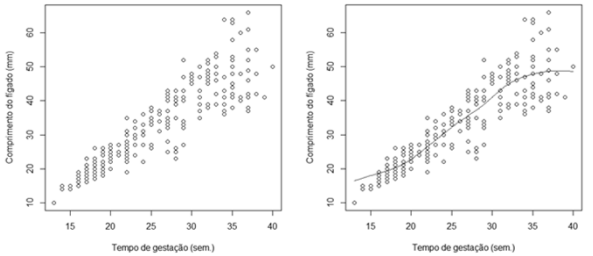
- Absorção de oxigênio na respiração:



√ Tendência claramente não linear
  – Como verificar se é exponencial?
√ Dispersão baixa e homogênea

25

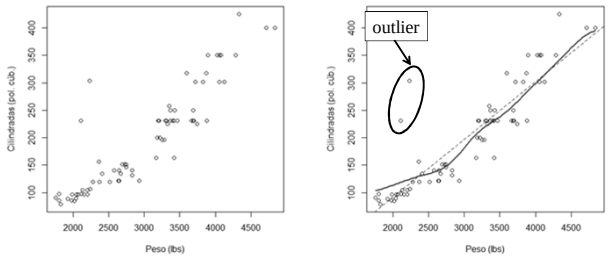
- Comprimentos de fígados de fetos:



√ Tendência não linear
  – Crescente e se estabiliza a partir da semana 32 (≈)
√ Dispersão não é homogênea (funil)

26

- Modelos de carros (USA):



√ Aparentemente tendência é não linear
√ Dispersão não aparenta ser homogênea
√ Prováveis outliers

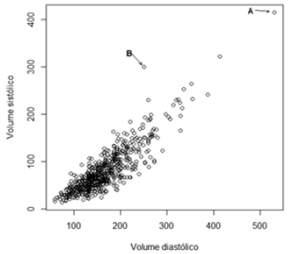
27

Valores Atípicos

- Identificação visual:
  √ Observações muito diferentes do restante dos dados
  √ Valores surpreendentemente distantes da curva de tendência (no contexto de regressão)
- Podem afetar bastante as conclusões
- Suspeita inicial:
  √ Tratar-se de um erro.

28

- Pacientes com doença coronariana:

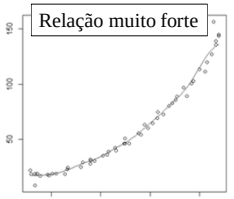


√ B: valor atípico
  – Excêntrico
√ Não tem 'má' aparência (conforme tendência)
  – Mas apresenta valor muito alto

29

Força das Relações

- Relações fortes:
  √ Se dispersão residual for pequena em comparação com a amplitude dos valores assumidos pela curva de tendência



  √ Curva de tendência explica maior parte da variação de Y

30

Associação entre Variáveis

- Variáveis estão associadas:
  √ Se os padrões forem muito fortes para serem explicados somente pelo acaso
- Variáveis positivamente associadas:
  √ Y tende a crescer com X
- Variáveis negativamente associadas:
  √ Y tende a diminuir conforme X aumenta

31

Predição

- Utilizar os valores das variáveis explicativas para predizer a resposta
- Só podemos fazer predição com confiança dentro do domínio de variação dos dados
  √ Quanto mais afastados dos dados, maiores as oportunidades de cometer erros

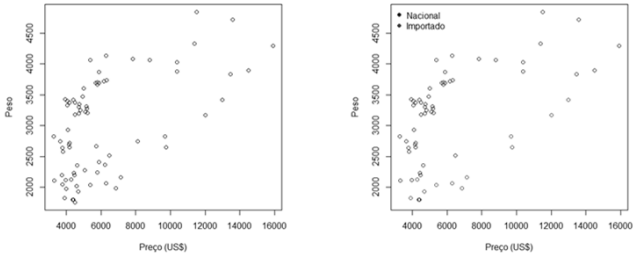
É necessária muita cautela ao predizer fora do domínio de variação dos dados

32

Outros Padrões

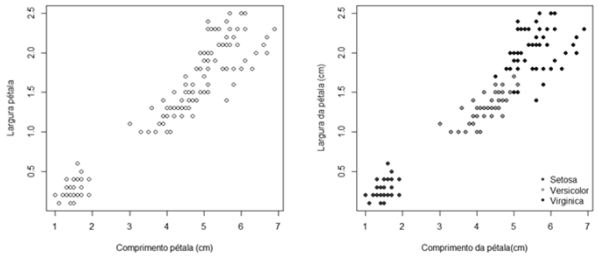
- Visualização gráfica:
  - √ Padrões tendência-mais-dispersão não são os únicos padrões a serem observados
- Outros padrões possíveis
  - √ Subconjuntos dos dados com relações distintas
  - √ Existência de clusters
    - – O que define esses grupos?

- Peso vs. preço de 74 modelos de carros:



√ Relação entre peso e preço é diferente para os dois tipos de carros (nacionais e importados)

- Largura vs. Comprimento de pétalas de iris:



√ Separação em dois grandes conglomerados
- – Inferior: espécie diferente
- – Superior: mistura de duas espécies

Correlação e Regressão

- Regressão:
  - √ Escolhe uma variável como resposta e usa a explanatória para explicar ou predizer seu comportamento
- Correlação:
  - √ Trata simetricamente ambas as variáveis
    - – Mesma importância
  - √ Aplicável somente se ambas as variáveis forem aleatórias

Correlação e Regressão

- Correlação:
  - √ Descreve quão próximo os dados seguem uma tendência linear
- Análise de regressão:
  - √ Fornece uma fórmula direta descrevendo a tendência entre variáveis

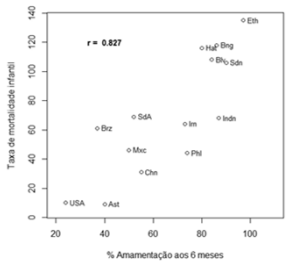
Questão de Causação

- Procurar causas:
  - √ Tentar compreender porque diferentes indivíduos produzem mais valores de Y
- Modelos de regressão:
  - √ Informação de variáveis explicativas para explicar o comportamento de uma resposta
  - √ Contribui para buscar causas
- Possíveis causas:
  - √ Variáveis relacionadas com a resposta

Exemplo

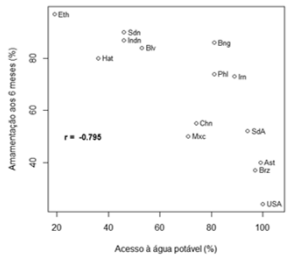
- Taxas de mortalidade infantil e amamentação materna:
  - √ Estudo de 14 países (dados de 1989)
  - √ Variáveis:
    - Taxas de mortalidade, em mortes por 1.000 pessoas
    - % de mães que amamentam aos 6 meses

- Taxa de mortalidade infantil em 14 países



  - √ Forte padrão linear de crescimento
    - Maiores % de amamentação têm maiores mortalidades
  - √ Amamentação é perigosa?

- Amamentação e água potável:

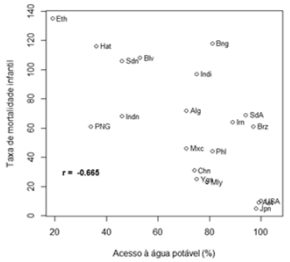


√ % maiores de amamentação no peito correspondem a menores acessos à água potável

√ Taxas mais altas de amamentação relacionadas com pobreza e deficiência de condições sanitárias?

41

- Água potável e mortalidade:



√ Conjunto completo

– Há missing data em amamentação

42

Importante

- Dados observacionais:
  - √ X e Y podem não ter nenhuma relação causal direta
  - √ Pode haver variável oculta que é a causa real de mudanças em Y, mas também está associada a X
    - – Por exemplo: água potável
  - √ Tempo é frequentemente uma variável oculta

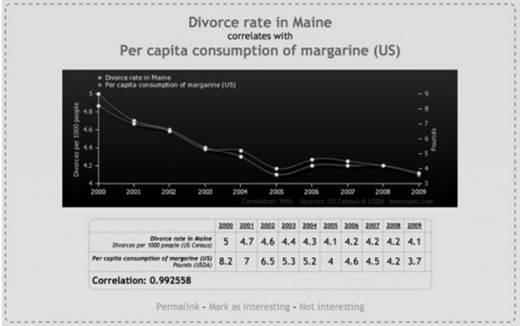
Em dados observacionais, relações fortes não são necessariamente causais

43

- Experimento aleatorizados:
  - √ Maneira de concluirmos confiavelmente sobre relações causais
  - √ Muitas vezes é impossível, fisica ou eticamente realizar experimentos que responderiam conclusivamente se a relação é causal ou não

44

r = 0,9926 indica uma relação alta entre variáveis?



	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
Divorce rate in Maine Divorces per 1000 people (US Census)	5	4.7	4.6	4.4	4.3	4.1	4.2	4.2	4.2	4.1
Per capita consumption of margarine (US) Pounds (USDA)	8.2	7	6.5	5.3	5.2	4	4.6	4.5	4.2	3.7

Correlation: 0.992558

Qual a relação entre taxa de divórcio e consumo de margarina?

Correlação

Correlação

- Descreve quão forte é a associação linear entre duas variáveis
  - √ Quão próximos os dados seguem uma *tendência linear*

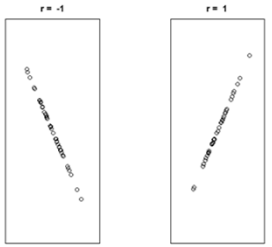
Coeficiente de Correlação Amostral

- Mede quão próximos de uma reta estão os pontos no gráfico X-Y.
- Dados:
  - √ $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{s_X}\right)\left(\frac{y_i - \bar{y}}{s_Y}\right) .$$

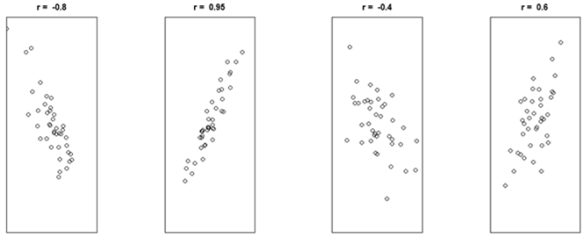
  - √ Valor independe da escolha do X e do Y
  - √ Medida adimensional
    - Cálculo baseia-se nas variáveis padronizadas
  - √ Ambas as variáveis devem ser quantitativas

- Gráficos de dispersão para valores de *r:*



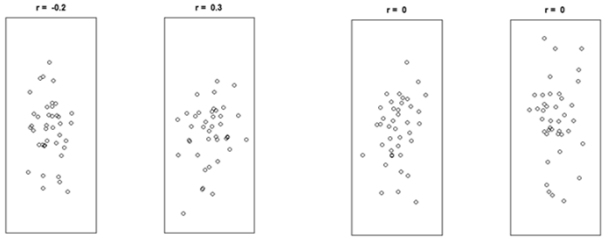
√ Relação linear exata (|r| = 1)

- Gráficos de dispersão para valores de *r:*



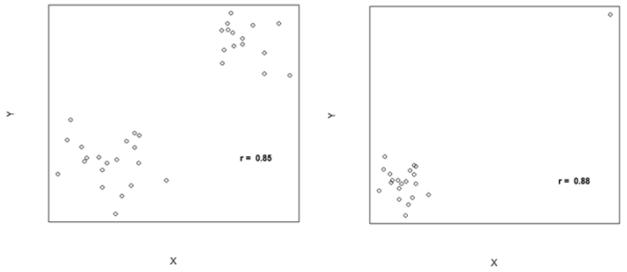
√ À medida que |r| torna-se menor, a relação linear parece cada vez mais fraca

- Gráficos de dispersão para valores de *r:*



√ Em r = 0 não há mais tendência linear

- Gráfico de dispersão de outros padrões:



√ Sem tendência dentro dos 2 clusters
- R é resultado da disposição dos conglomerados

√ Conglomerado sem correlação e um outlier

Propriedades

i. r assume valores entre –1 e 1
ii. Valores positivos de r
   - Tendência ascendente
   - Valores grandes de X associados a valores grandes de Y
iii. Valores negativos de r:
   - Tendência descendente
   - Valores grandes de X estão associado a valores pequenos de Y

Propriedades

iv. r = +1
   - Pontos caem exatamente numa reta ascendente
v. r = –1
   - Pontos caem exatamente numa reta descendente
vi. r = 0
   - Não há tendência linear ascendente ou descendente
vii. |r|: tamanho de r
   - Mede o quão próximo os pontos estão de cair sobre uma reta

Importante

- r pode ser influenciado por pontos que se afastam muito do padrão geral das observações
  - √ Mesmo poucos
- r não é uma descrição completa dos dados de duas variáveis

Teste de Significância

- ρ: coeficiente de correlação populacional
  - √ verdadeira relação entre duas variáveis na população da qual os dados foram amostrados
- Hipóteses:
  - √ H_0: $\rho = 0$
  - √ H_1: $\rho \neq 0$

- Estatística de teste:

$$t_0 = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \sim \mathrm{t}_{n-2} \ .$$

√ Teste relacionado com inclinação da reta de regressão das duas variáveis

√ Suposição do teste:

– Variáveis são conjuntamente normais, com mesma variância

√ Valor de r (e da estatística t_0) pode ser influenciado pelo tamanho da amostra

√ Importante ponderar a significância prática das correlações estatisticamente significantes

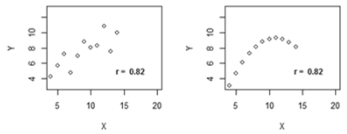
58

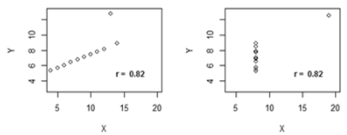
- Para estimar quão grande é a verdadeira correlação é necessário que p-valor ofereça evidência contra H_0
- IC's para ρ são difíceis de serem obtidos

59

Cuidados com o Uso da Correlação

- r não deve ser calculado para estes gráficos





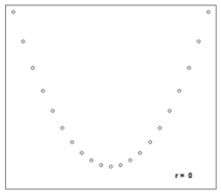
X	Y
$\bar{x} = 9$	66,45
$s_X = 11$	89,65
$r = 0{,}816$	
$Y = 3 + 0{,}5X$	

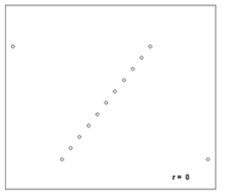
√ Mesmas estatísticas descritivas amostrais

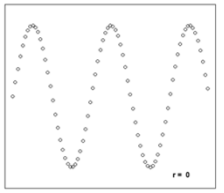
√ Correlação igual com relações muito distintas entre as variáveis X e Y

60

- Padrões com r = 0







√ Em todos os casos as variáveis estão associadas deterministicamente

√ No caso, r = 0 implica em deslocamentos não discerníveis para cima ou para baixo?

61

Importante

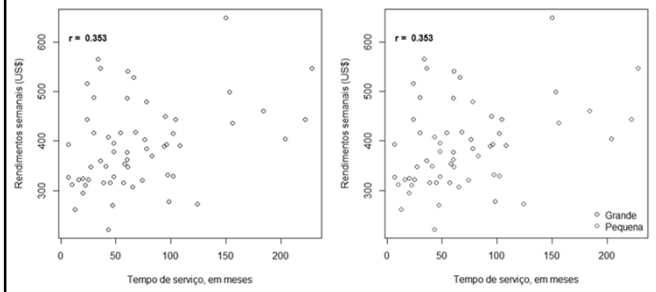
- Sempre lembrar que:
  - √ Correlação não implica necessariamente causação!

Regressão Linear Simples

Exemplo

- Salários aumentam com a experiência?
  - √ Amostra com 59 mulheres casadas, trabalhando no serviço de atendimento a clientes em banco
  - √ Variáveis:
    - Salário semana (US$)
    - Tempo de serviço com atual empregador, em meses
    - Tamanho d banco (Grande = mais de 100 empregados, Pequeno = caso contrário)
  - √ Há relação clara entre salários e tempo de serviço?

- Gráfico de dispersão dos dados:

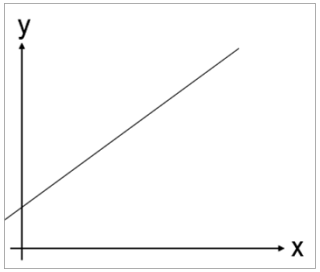


  - √ Relação moderadamente linear sem outliers
  - √ Salários variam de $221 a $649
    - Bancárias diferem entre si quanto ao tempo de serviço

Função Linear

$$f(x) = ax + b\ .$$

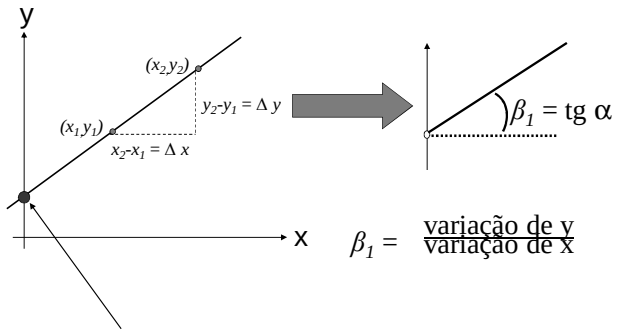
- f(x) se modifica a uma taxa constante em relação à sua variável independente
  - √ *a* e *b* são constantes
  - √ *b* : intercepto
  - √ *a* : coeficiente angular (inclinação)



Tendência Linear do Dados

- Equação da tendência linear: $\hat{y} = \beta_0 + \beta_1 x\ .$
- Parâmetros:
  - √ β_0: intercepto
  - √ β_1: declividade da reta
- Escolha da melhor reta para resumo dos dados:
  - √ Escolher os valores de β_0 e β_1

Intercepto e Coeficiente Angular



$\beta_1 = \dfrac{\text{variação de y}}{\text{variação de x}}$

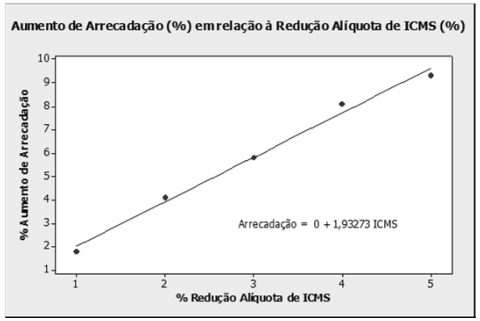
β_0: intersecção da reta com o eixo y

Interpretação dos Parâmetros

- β_1 : declividade da reta
  - √ define o aumento ou diminuição da variável *Y* por uma unidade de variação de *X*
- β_0 = intercepto em y
  - √ define o valor médio de *Y* sem a interferência de *X* (com *X=0*).
  - √ Nem sempre a interpretação de β_0 é aplicável

LFB2

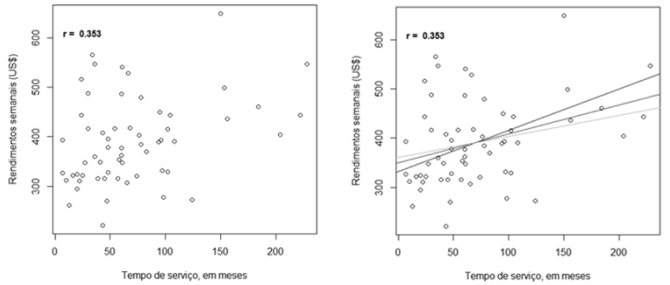
Exemplo



β_1: a cada redução de 1% na alíquota há *1,9 %* de aumento de arrecadação

β_0*=0:* Se não há redução na alíquota não há de aumento de arrecadação

- Escolha da reta da tendência linear:



√ Não há uma única reta possível de representar a tendência linear de crescimento dos rendimentos

√ Como escolher a melhor reta?

71

Predição do Valor de Y

- $\hat{y}_i$: valor predito de y da observação i (x_i)

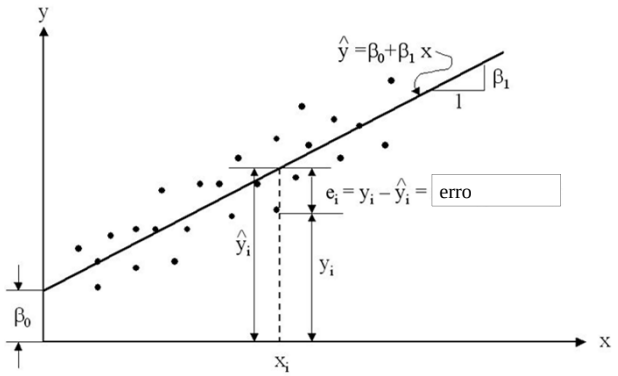
$$\hat{y}_i = \beta_0 + \beta_1 x_i \ .$$

- Erro de predição: $y_i - \hat{y}_i$.

√ Ocorre ao usarmos $\hat{y}_i$ para predizer o resultado real y_i.

72

- Erro de predição ocorre em todos os pontos



73

Slide 70

LFB2 Refazer gráfico

Lupercio Bessegato; 31/10/2007

Critério para Escolha

- Reta que se ajusta bem aos dados
  - √ Aquela para a qual os erros de predição (distâncias verticais) forem tão pequenos quanto possível
    - – Em algum sentido de média global

Critério de Mínimos Quadrados

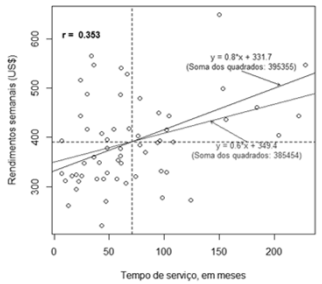
- Critério mais usado
- Reta que melhor se ajusta aos dados:
  - √ minimiza a soma dos quadrados dos erros de predição

$$\sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \ .$$

  - √ Escolher β_0 e β_1 que atendam esse critério
  - √ Ele pode ser usado para outros tipos de curvas

Exemplo

- Retas:
  - √ Reta 1: $\hat{y} = 349,4 + 0,6x$.
  - √ Reta 2: $\hat{y} = 331,7 + 0,8x$.
- Soma dos quadrados:
  - √ Reta 1: 385.454
  - √ Reta 2: 395.355



Reta 1 se ajusta melhor aos dados que a reta 2

Reta de Mínimos Quadrados

- Reta que melhor se ajusta aos dados
  - √ Intercepto: $\beta_0 = \hat{\beta}_0$
  - √ Inclinação: $\beta_1 = \hat{\beta}_1$
- Melhor ajuste no sentido da soma dos quadrados dos erros de medição

- Estimativas de mínimos quadrados do intercepto e da inclinação:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \; . \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \; .$$

Expressão alternativa $\hat{\beta}_1 = r\frac{s_Y}{s_X}$.

- Coeficientes da regressão:
  - √ $\hat{\beta}_0$ e $\hat{\beta}_1$
- Reta de mínimos quadrados

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \; .$$

  - √ Passa pelo ponto $(\bar{x}, \bar{y})$
    - Centróide dos dados

Exemplo

- Salários aumentam com a experiência?
  - √ Ajuste da reta de regressão aos dados:

```
> salario.fit <- lm(salarios ~ tempo, data = bancarias)
> coef(salario.fit)

(Intercept)       tempo
349.3780625   0.5904531
```

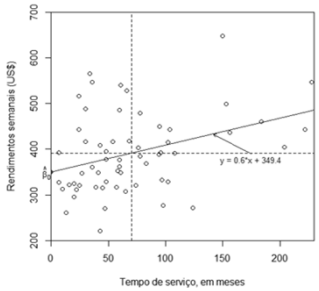
$\hat{\beta}_0 = 349,4$

$\hat{\beta}_1 = 0,59$.

  - √ Reta de mínimos quadrados

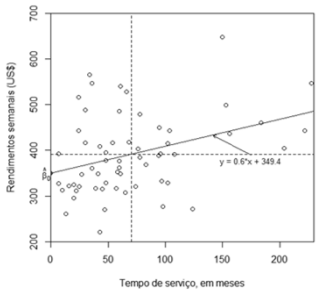
```
Salario = 349,4 + 0,59 Tempo de serviço
```


- Diagrama de dispersão e a reta de regressão:



  - √ Reta passa pelo ponto $(\bar{x}, \bar{y})$: (70,5; 391,0)
  - √ Intercepto: $\hat{\beta}_0 = 349,4$

- Interpretação



  - √ $\hat{\beta}_0 = 349,4$
    - Rendimento médio para bancárias sem experiência
  - √ $\hat{\beta}_1 = 0,59$
    - Aumento do rendimento médio para cada semana de experiência

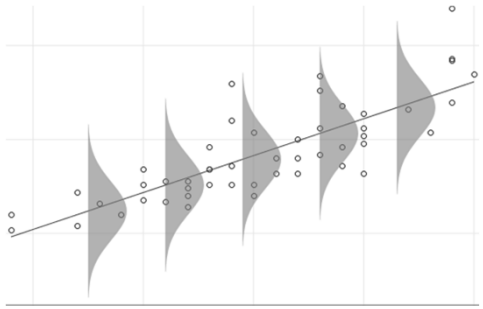
Modelo Linear Simples – Inferência

Relação linear	=	tendência linear + dispersão residual

- Modelo linear simples:
  - √ Médias de Y pertencem a uma reta
    - – Dependem do valor de X
  - √ Dados observados flutuam em torno da reta de acordo com uma normal
  - √ Variação em torno da reta é constante e igual a σ.

83

- Modelo linear simples



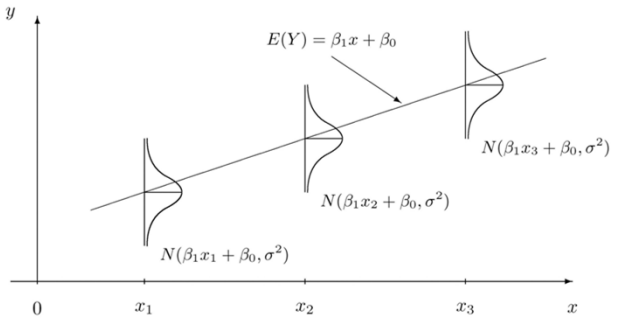
84

Modelo Formal

- Quando $X = x, Y \sim \text{normal}(\mu_Y, \sigma)$
  - √ onde $\mu_Y = \beta_0 + \beta_1 x$
- Seja $\epsilon = Y - \mu_Y$, então $\mathrm{E}(\epsilon) = 0$ e $\mathrm{dp}(\epsilon) = \mathrm{dp}(Y) = \sigma$.
- Modelo estatístico: $Y = \mu_Y + \epsilon$, onde $\epsilon \sim \mathrm{N}(0, \sigma)$.
- Modelo mais comum: $Y = \underbrace{\beta_0 + \beta_1 x}_{\text{tendência linear}} + \underbrace{\epsilon}_{\text{erro aleatório}}$.

86

- Modelo estatístico de regressão linear



87

Hipóteses do Modelo

i. Há uma relação linear entre x e o valor médio de Y, quando X = x

$$\mu_Y = \beta_0 + \beta_1 x \ .$$

ii. Essa é a verdadeira reta e os valores de seu intercepto β_0 e inclinação β_1 são desconhecidos

iii. As estimativas de mínimos quadrados (EMQ) $\hat{\beta}_0$ e $\hat{\beta}_1$ estimam os verdadeiros valores de β_0 e β_1

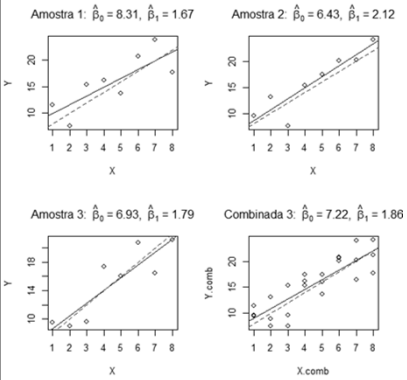
Hipóteses do Modelo

iv. Os erros aleatórios ϵ são normalmente distribuídos
- √ Têm mesmo desvio padrão σ
- √ São estatisticamente independentes

Simulação

- Modelo: Y = 6 + 2x + ϵ, com $\epsilon \sim N(0,3)$
  - √ N = 8 com X={1, 2, ..., 8}
  - √ Cálculo das estimativas de mínimos quadrados $\hat{\beta}_0$ e $\hat{\beta}_1$

- Estimativas da reta de regressão:
  - √ Reta verdadeira: $\beta_0 = 6$ e $\beta_1 = 2$



- Diferentes amostras dão EMQ distintas
  - √ $\hat{\beta}_0$ e $\hat{\beta}_1$ são aleatórios
- Dados combinados:
  - √ Reta estimada mais próxima da verdadeira
  - √ Baseada em mais dados

Estimativas de Mínimos Quadrados - Propriedades

- As estimativas de mínimo quadrado são não viciadas e normalmente distribuídas
  - √ Desvios padrão das EMQ

$$\mathrm{dp}(\hat{\beta}_0) = \frac{\sigma}{S_X}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n(n-1)}}\ . \qquad \mathrm{dp}(\hat{\beta}_1) = \frac{\sigma}{S_X}\sqrt{\frac{1}{n-1}}\ .$$

  - √ S_X: desvio padrão dos valores de X
  - √ σ: desvio padrão populacional dos erros aleatórios

$$\mathrm{dp}(\hat{\beta}_0) = \frac{\sigma}{S_X}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n(n-1)}}\ . \qquad \mathrm{dp}(\hat{\beta}_1) = \frac{\sigma}{S_X}\sqrt{\frac{1}{n-1}}\ .$$

- Comentários:
  - √ À medida que σ cresce, aumenta a flutuação dos dados
    - Dados com mais ruído produzem estimativas de mínimos quadrados mais variáveis
  - √ Muito ruído = muita flutuação em torno da tendência

Estimação de σ

- σ é desconhecida
  - √ $dp(\hat{\beta}_0)$ e $dp(\hat{\beta}_1)$ não podem ser calculados
- Solução:
  - √ Substituir σ por sua estimativa s_U

$$s_U = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}\ .$$

    - s_U é estimativa não viciada de σ

Estimativas de Mínimos Quadrados – Erros Padrão

- Substituindo σ por sua estimativa s_U, têm-se:

$$\mathrm{ep}(\hat{\beta}_0) = \frac{s_U}{s_X}\sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n(n-1)}}\ . \qquad \mathrm{ep}(\hat{\beta}_1) = \frac{s_U}{s_X}\sqrt{\frac{1}{n-1}}\ .$$

Inferência sobre β_0 e β_1

- Intervalos de confiança para o verdadeiro valor da inclinação β_1:

$$\text{estimativa} \pm t_{\text{gl}} \text{erros padrão}$$

$$\hat{\beta}_1 \pm t_{\text{gl}} \times \text{ep}(\hat{\beta}_1) \ .$$

√ Graus de liberdade: $\text{gl} = n - 2$.

- Estatística de teste:
  - √ H_0: $\beta_1 = c$ versus $\beta_1 \neq c$

$$t_0 = \frac{\text{estimativa} - \text{valor admitido por hipótese}}{\text{erro padrão}}$$

$$= \frac{\hat{\beta}_1 - c}{\text{ep}(\hat{\beta}_1)}.$$

  - √ Em geral, há mais interesse na declividade que no intercepto
  - √ Testar a inexistência de relação linear é testar H_0: $\beta_1 = 0$

Exemplo

- Há relação linear entre salário e tempo de serviço?

```
> summary(salario.fit)

Call:
lm(formula = salarios ~ tempo, data = bancarias)

Residuals:
    Min      1Q  Median      3Q     Max
-153.77  -50.54  -15.88   37.61  211.05

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 349.3781    18.0965  19.306  < 2e-16 ***
tempo         0.5905     0.2070   2.853  0.00603 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 82.23 on 57 degrees of freedom
Multiple R-squared:  0.1249,    Adjusted R-squared:  0.1096
F-statistic: 8.138 on 1 and 57 DF,  p-value: 0.006029
```


- Conclusão:
  - √ Existe uma forte evidência de que o salário médio das bancárias cresce com o aumento do tempo de serviço
    (p-valor = 0,006)
  - √ A variação dos salários ao longo da reta de regressão, em função do tempo de serviço explica somente 12% dessa variação
    - 88% restantes devem-se a outras diferenças entre essas funcionárias
      (R^2 = 0,125)

- Intervalo com 95% de confiança para β_1:
  - √ Saída computacional:

```
> confint(salario.fit)     # intervalos de confiança para os parâmetros)

 2.5 %     97.5 %
(Intercept) 313.1404731 385.615652
tempo         0.1759936   1.004913
```

  - √ Cálculo manual
    - Graus de liberdade: $\text{gl} = 59 - 2 = 57$.
    - Estimativas (saída): $\hat{\beta}_1 = 0,5905$

$$\text{ep}(\hat{\beta}_1) = 0,2070 .$$

$$\hat{\beta}_1 \pm t_{57} \times \text{ep}(\hat{\beta}_1)$$
$$0,5905 \pm 2,002 \times 0,2070$$
$$[0,176; 1,005] .$$

- Interpretação para o IC [0,176; 1,005]
  - √ Declividade informa sobre a mudança de Y associada a uma unidade de X
  - √ Há um aumento do salário médio semanal de \$0,2 a \$1 para cada semana de serviço.

ANOVA para Regressão

- Equação da Anova para Regressão

Variação total em y	=	Variação ao longo da reta	+	Variação em torno da reta

SQ Total	=	SQ Reg	+	SQ Res

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y})^2 .$$

Exemplo

- Há relação linear entre salário e tempo de serviço?

```
> anova(salario.fit)

Analysis of Variance Table

Response: salarios
          Df  Sum Sq Mean Sq F value   Pr(>F)
tempo      1   55034   55034  8.1384 0.006029
Residuals 57  385454    6762
```

  - √ Somas de quadrados: $\text{SQReg} = 55.034$
$\text{SQRes} = 385.453$.
  - √ Coeficiente de determinação: $R^2 = \frac{\text{SQReg}}{\text{SQRes}} = \frac{55.034}{385.453} = 0,1240$.
    - Tempo de serviço explica apenas 12% da variação dos salários (ao longo da reta). 88% restantes devem-se a outros fatores

Predição

- Predizer a média de Y (μ_Y) quando X = x_p

$$\mu_Y = \beta_0 + \beta_1 x_P \ .$$

- Exemplo:
  - √ Predizer o rendimento médio da subpopulação de bancárias com tempo de serviço de 125 meses:
    - $x_p = 125$

$$\mu_Y = 349,4 + 0,5905 \times 125 = \$423 \text{ por semana} \ .$$

  - √ Predizer o rendimento de uma única bancária:
    - – Predição é a mesma!: $423

Intervalo de Confiança para a Média (μ_Y)

$\hat{y}_P$: estima o valor médio de Y, quando $X = x_p$

- √ $\hat{y}_P$ é estimativa não viciada de μ_Y
- Intervalo de confiança para a média μ_Y

$$\hat{y}_p \pm t_{\text{gl}} \times \text{ep}(\hat{y}_p) \ .$$

- √ Graus de liberdade: $\text{gl} = n - 2$.
- √ ep($\hat{y}_P$): erro padrão de $\hat{y}_P$.

- Intervalo de predição para Y_P:
  - √ Tenta predizer o próximo valor real de Y para $X = x_P$
  - √ $\hat{y}_P$: valor a ser observado quando $X = x_P$

$$\hat{y}_p \pm t_{\text{gl}} \times \text{epp}(\hat{y}_p) \ .$$

  - √ Graus de liberdade: $\text{gl} = n - 2$.
  - √ epp($\hat{y}_P$): erro padrão de predição de $\hat{y}_P$.

Incerteza na Predição de Y_P

- Fontes de incerteza na predição de Y_P:
  - √ Incerteza sobre os verdadeiros valores de β_0 e β_1
  - √ Incerteza devido à flutuação em torno da reta
- Consequência:
  - √ Intervalo de predição do próximo valor real (Y_P) é sempre mais largo que o intervalo de confiança para a média (μ_P)

Erros Padrão das Predições

- Erro padrão da estimativa da média (μ_P):

$$\text{ep}(\hat{y}_P) = s_U \sqrt{\frac{1}{n} + \frac{(x_P - \bar{x})^2}{(n-1)S_X^2}} \ .$$

- Erro padrão de predição do próximo valor real (Y_P)

$$\text{epp}(\hat{y}_P) = s_U \sqrt{1 + \frac{1}{n} + \frac{(x_P - \bar{x})^2}{(n-1)S_X^2}} \ .$$

√ $\text{ep}(\hat{y}_p)$ é sempre maior que $\text{epp}(\hat{y}_p)$.

√ Estimativas distantes de $\bar{x}$, são mais imprecisas

Exemplo

- Relação entre salários e experiência:

√ x_P = 125 semanas

- Intervalo de confiança para μ_P:

```
> pred.tempo <- data.frame(tempo = 125)
> predito <- predict(salario.fit, int = "c", newdata = pred.tempo, se.fit = T)
> predito$fit#valor predito e IC
       fit      lwr      upr
1 423.1847 392.0403 454.3291
> predito$se.fit# erro padrão da estimativa
[1] 15.55303
```

$$\text{ep}(\hat{y}_P) = 15,55 \ .$$

$$t_{57}(0,975) = 2,002 \ .$$

- Intervalo de predição para Y_P:

```
> pred.tempo <- data.frame(tempo = 125)
> predito <- predict(salario.fit, int = "p", newdata = pred.tempo, se.fit = T)
> predito$fit#valor predito e IP
       fit      lwr      upr
1 423.1847 255.5957 590.7737
```

$$\text{epp}(\hat{y}_P) = \sqrt{\text{ep}(\hat{y}_P)^2 + \hat{\sigma}^2}$$
$$= 15,55^2 + 82.23^2 = 83,69 \ .$$

- Interpretação:

√ Intervalo de 95% de confiança para μ_p:

– [392; 454]

– Com 95% de confiança, a média de salário para bancárias com 125 meses de experiência é não menor que $392 e não maior que $454

√ Intervalo de predição de 95% para Y_P:

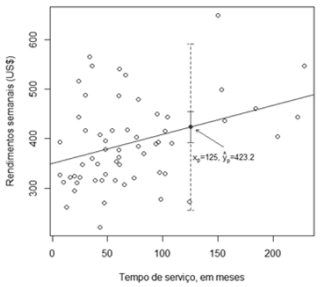
– [256; 591]

– Intervalo de predição de uma nova observação de salário semanal para bancária com 125 meses de tempo de serviço.

- Intervalo de confiança da média e intervalo de predição

√ x_p = 125 semanas

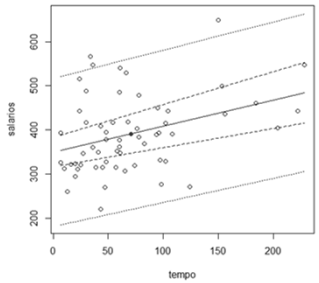
√ Intervalos centrados em $423,20



√ Intervalo de predição é mais largo

√ Abrange dispersão vertical dos rendimentos das bancárias

- Bandas de confiança e de predição de 95%



√ Intervalos são mais estreitos no centro dos dados e se alargam à medida que se afasta do centro, em cada direção

Verificação do Modelo

- Uso dos resíduos para verificação do modelo:
  - √ $\hat{u}_i$: estimativas dos erros (não observados)

$$\hat{u}_i = y_i - \hat{y}_i \ .$$

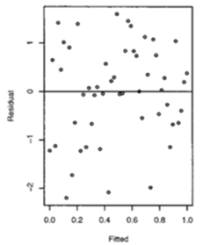
- Análise gráfica dos resíduos para avaliar indícios de violação das hipóteses de:
  - √ Normalidade dos erros
  - √ Homocedasticidade dos erros
  - √ Especificação do modelo
  - √ Independência dos erros
  - √ Outliers

Análise dos Resíduos do Modelo

i. Verificação da normalidade dos erros:
   - √ Gráfico de probabilidade normal dos resíduos e teste formal de normalidade

ii. Gráfico de resíduos ($\hat{u}_i$) versus x_i e resíduos versus valor ajustado ($\hat{y}_i$):
   - √ Indicações sobre variabilidade dos dados
   - √ Problemas com a especificação do modelo

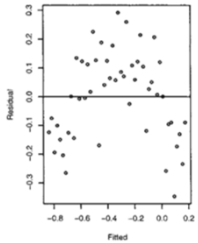
- Padrão ‘normal’ dos resíduos:



√ Faixa horizontal sem problemas

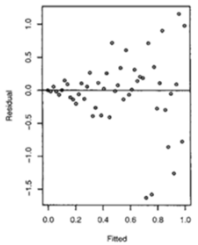
√ Formas ovais não indicam problema

- Qualquer tendência nos gráficos:
  - √ Indicação que modelo ajustado não resumiu adequadamente a tendência ou padrão dos dados



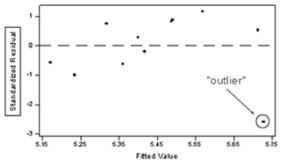
  - √ Tendência não linear nos resíduos

- Leque ou funil em qualquer direção:



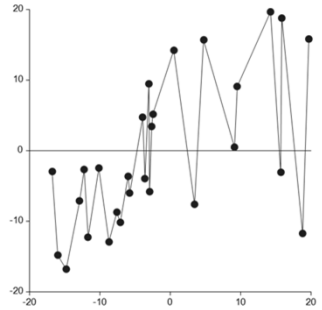
  - √ Indica que a variabilidade dos erros não é constante

- Valores atípicos:
  - √ São apontados claramente nesses gráficos



iii. Gráfico dos resíduos versus tempo (ou ordem) nos quais as observações foram feitas:
  - √ Podem revelar rupturas em experimentos
- Procurando falta de independência:
  - √ Possível se houver informação sobre a sequência de coleta dos dados ao longo do tempo
- Deve-se procurar a existência de correlação serial:
  - √ qualquer relação entre resíduos sucessivos

- Gráfico de $\hat{u}_i$ versus $\hat{u}_{i-1}$:

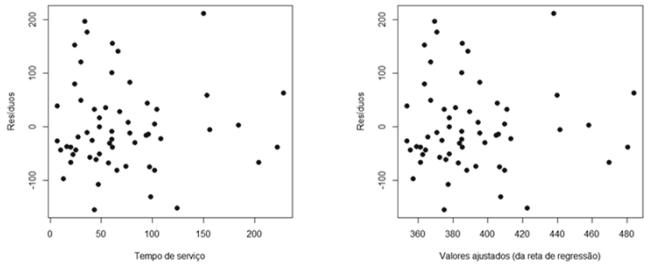


√ Espera-se uma faixa horizontal sem padrão

√ Pode-se usar o teste de Durbin-Watson

122

- Gráficos dos Resíduos versus explicativa valores ajustados:



√ Há poucas observações acima de 120 semanas

√ Pontos abaixo de 125 não apresentam padrão irregular

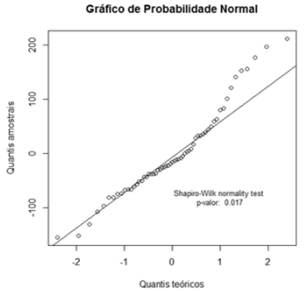
125

Exemplo

- Análise de resíduos do modelo linear da relação entre salários e experiência:
- Gráficos:
  - √ Resíduos vs. explicativa
  - √ Resíduos vs. valores ajustados
  - √ Normalidade

124

- Gráfico de probabilidade normal

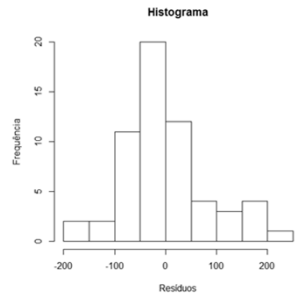


√ Gráfica indica não normalidade

√ Há evidência contra a hipótese de normalidade

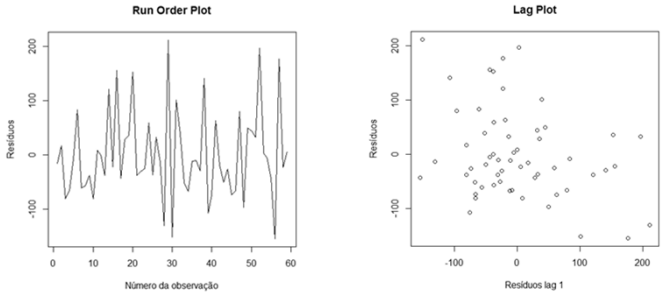
126

- Histograma:



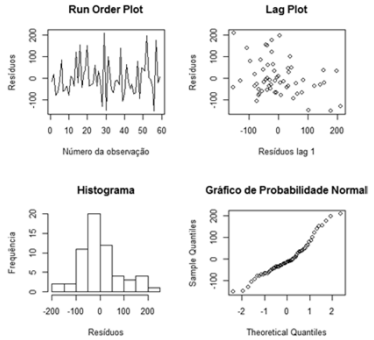
√ Assimetria a direita
  – Alguns resíduos muito grandes

- Run Plot:
  √ Considerando que a ordem foi dada:



√ Não há indicação de correlação serial

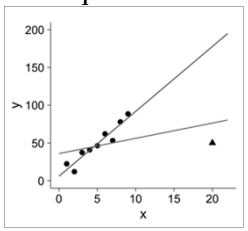
- Gráficos combinados:
  √ Repetindo gráficos anteriores



- Conclusão:
  √ O modelo linear não se ajusta bem aos dados
    – R^2 baixo
    – Desvio padrão dos resíduos relativamente alto
  √ Há violações das suposições do modelo que podem afetar as conclusões e a inferência

Valores Atípicos em X

- Podem ter grande influência na posição da reta de mínimos quadrados:



√ Pontos influentes:

– À medida que outlier se afasta, arrasta com ele a reta ajustada de regressão de mínimos quadrados

- Quando forem percebidos possíveis pontos influentes:
  - √ Analisar a relação sem incluir os pontos influentes
  - √ Reportar sua existência e influência com os resultados da análise

Referências

Bibliografia Recomendada

- AGRESTI, A.; FINLAY, B. *Métodos estatísticos para as ciências sociais.* Penso, 2012.
- MOORE, D. S.; MCCABE, G. P.; DUCKWORTH, W. M.; SLOVE, S. L. *A prática da estatística empresarial: como usar dados para tomar decisões*. LTC, 2006.
- WILD, J. W.; SEBER, G. A. F. *Encontros com o acaso: um primeiro curso de análise de dados e inferência*. LTC, 2004.