UENF
Universidade Estadual do Norte Fluminense Darcy Ribeiro

CCT
LCMAT

Regressão e Correlação

Prof. Fermín Alfredo Tang Montané

Análise de Regressão

Introdução

- O objetivo da análise de regressão é explorar a relação entre duas (ou mais) variáveis, de modo que possamos obter informações sobre uma delas, por meio dos valores conhecidos da(s) outra(s).
- O objetivo de grande parte dos cálculos é investigar as variáveis que estão relacionadas deterministicamente. Afirmar que x e y estão relacionadas dessa maneira significa dizer que o conhecimento do valor de x implica o conhecimento exato do valor de y.
- Por exemplo:
  - Suponha que decidamos alugar uma van por um dia e que o valor do aluguel seja $ 25,00 mais $ 0,30 por quilômetro rodado. Sendo:

    $x = o\ número\ de\ quilômetros\ rodados;\ e$

    $y = a\ despesa\ de\ aluguel;$
  - Então: $y = 25 + 0{,}3x$
  - Se percorrermos 100 quilômetros com a van ($x = 100$), então
  - $y = 25 + 0{,}3x = 25 + 0{,}3(100) = 55$.

Análise de Regressão

Introdução

- Por exemplo:
  - Se a velocidade inicial de uma partícula for v_0 e ela sofrer uma aceleração constante, sendo que:

    $x = o\ tempo; e$

    $y = a\ distância\ percorrida;$
  - Então temos que: $y = v_0 x + \frac{1}{2} a x^2$
  - Trata-se de um relação determinística não-linear.

Análise de Regressão

Introdução

- Muitas variáveis x e y podem até parecer relacionadas uma com a outra, mas não de maneira determinística.
- Um exemplo é dado pelas variáveis:

  $x =$ média total geral de notas escolares do ensino médio; e

  $y =$ coeficiente de rendimento da faculdade.
- O valor de y não pode ser determinado apenas com base no valor conhecido de x e dois diferentes alunos podem ter o mesmo valor de x, mas com valores de y bem diferentes.
- Existe uma tendência de alunos médias altos (baixos) no ensino médio terem também coeficientes de rendimentos altos (baixos) na faculdade. Conhecer a média do ensino médio de um aluno pode ser muito útil para nos ajudar a prever seu desempenho na faculdade.

Análise de Regressão

Introdução

- Temos exemplos adicionais de relação não-determinística:
- Exemplo 1:

  $x =$ idade de uma criança; e

  $y =$ tamanho do vocabulário dessa criança.
- Exemplo 2:

  $x =$ tamanho de um motor em centímetros cúbicos; e

  $y =$ rendimento do combustível de um automóvel equipado com esse motor;
- Exemplo 3:

  $x =$ força de tração aplicada; e

  $y =$ quantidade de alongamento de uma tira de metal.

Análise de Regressão

Definição

- A análise de regressão é a parte da estatística que investiga a relação entre duas ou mais variáveis relacionadas de maneira não-determinística.
- Estendemos (generalizamos) a relação linear determinística $y = \beta_0 + \beta_1 x$ para uma relação linear probabilística;
- Desenvolvemos procedimentos para fazer inferências sobre os parâmetros do modelo e obter uma medida quantitativa (o coeficiente de correlação) sobre até que ponto as duas variáveis estão relacionadas.

Regressão Linear Simples

Definição

- A relação matemática determinística mais simples entre duas variáveis x e y é uma relação linear $y = \beta_0 + \beta_1 x$.
- O conjunto de pares (x, y) para o qual $y = \beta_0 + \beta_1 x$ determina uma reta com coeficiente angular β_1 e termo constante β_0 .
- Nosso objetivo é desenvolver um modelo probabilístico linear.

- A relação matemática determinística mais simples entre duas variáveis x e y é uma relação linear $y = \beta_0 + \beta_1 x$.
- O conjunto de pares (x, y) para o qual $y = \beta_0 + \beta_1 x$ determina uma reta com coeficiente angular β_1 e termo constante β_0 .
- Nosso objetivo é desenvolver um modelo probabilístico linear.

Regressão Linear Simples

Definição

- Se as duas variáveis não estiverem relacionadas deterministicamente, então, para um valor fixo de x, o valor da segunda variável será aleatório.
- Por exemplo:
- Se estivermos investigando a relação entre a idade de uma criança e o tamanho do seu vocabulário e decidirmos selecionar uma criança de idade $x = 5{,}0$ anos, então, antes que a seleção seja feita, o tamanho do vocabulário será uma variável aleatória Y. Depois que uma determinada criança de 5 anos tiver sido selecionada e testada, o resultado será, por exemplo, um vocabulário de 2000 palavras.
- Portanto, podemos dizer que o valor observado de Y, associado com a fixação de $x = 5{,}0$, foi $y = 2.000$.
- Na maioria das vezes, a variável cujo valor é fixado pelo pesquisador será representada por x e será chamada de variável independente, previsão ou explicativa. Para x fixo, a segunda variável será aleatória; representamos essa variável aleatória e seu valor observado por Y e y, respectivamente, e as chamamos de variável dependente ou resposta.

Regressão Linear Simples

Definição

- Em geral, as observações serão feitas para inúmeros conjuntos da variável independente.
- Sejam x_1, x_2, ..., x_n os valores da variável independente para as quais são feitas as observações e sejam Y e y, respectivamente, a variável aleatória e o valor observado associados a x_i.
- Portanto, os dados bivariados fornecidos consistem nos n pares $(x_1,\ y_1)$, $(x_2,\ y_2)$, ... , $(x_n,\ y_n)$. O primeiro passo na análise de regressão com duas variáveis é elaborar um gráfico de dispersão dos dados observados.
- Em um gráfico desse tipo, cada $(x_i,\ y_i)$ é representado como um ponto representado graficamente em um sistema de coordenadas bidimensional.

Regressão Linear Simples

Modelo Probabilístico Linear

- Para o modelo determinístico $y = \beta_0 + \beta_1 x$, o valor real observado de y é uma função linear de x.
- A generalização apropriada dessa característica para um modelo probabilístico pressupõe que o valor esperado de Y é uma função linear de x, mas que, para um x fixo, a variável Y difere de seu valor esperado de uma quantidade aleatória.
- Existem parâmetros β_0, β_1, e σ^2 tais que, para qualquer valor fixo da variável independente x, a variável dependente está relacionada a x por meio da equação:

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

- A quantidade ϵ na equação do modelo é uma variável aleatória, considerada normalmente distribuída com $\mathrm{E}(\epsilon) = 0$ e $\mathrm{V}(\epsilon) = \sigma^2$.
- A variável ϵ normalmente é chamada de **desvio aleatório** ou **erro aleatório** do modelo. Sem ϵ, qualquer par observado (x, y) corresponderia a um ponto disposto exatamente na reta $y = \beta_0 + \beta_1 x$, denominada **reta de regressão real (**ou da **população).**

Regressão Linear Simples

Modelo Probabilístico Linear

- A inclusão do termo "erro aleatório" permite que (x, y) fique acima da reta de regressão real (quando $\epsilon > 0$) ou abaixo da reta (quando $\epsilon < 0$).
- Os pontos (x_2, y_2) , ... , (x_n, y_n) resultantes de n observações independentes serão então dispersos próximos da reta de regressão:

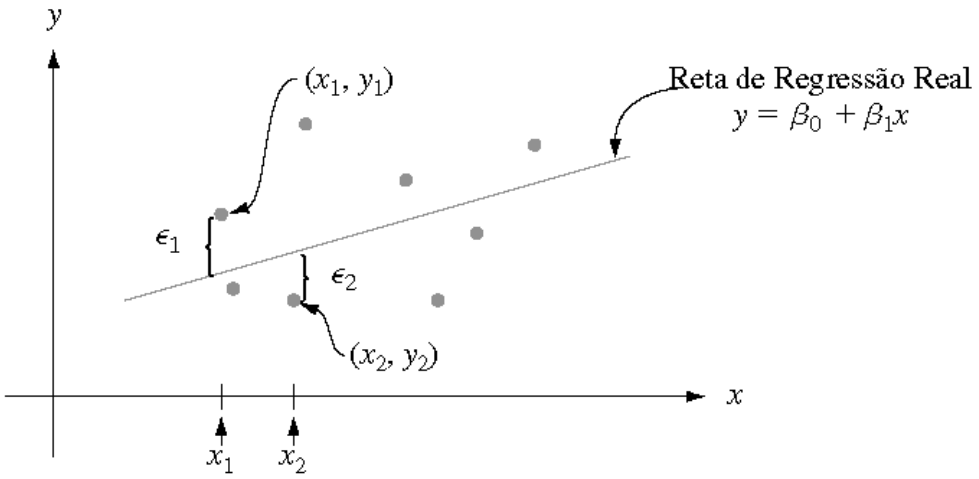


Figura 12.3 Pontos correspondentes a observações do modelo de regressão linear simples

Regressão Linear Simples

Modelo Probabilístico Linear

- O parâmetro de variância σ^2 determina até que ponto cada curva normal se dispersa ao redor de seu valor médio (a altura da reta). Quando σ^2 é pequeno, um ponto observado (x, y) quase sempre ficará bem próximo da reta de regressão real, ao passo que as observações podem desviar consideravelmente de seus valores esperados quando σ^2 for grande.

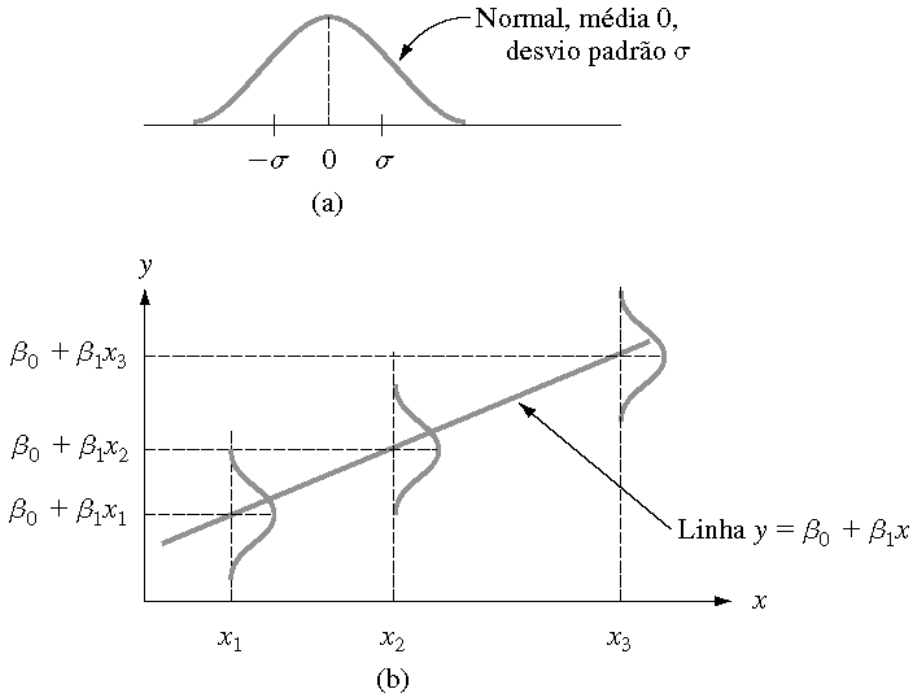


Figura 12.4 (a) Distribuição de ϵ; (b) distribuição de Y para diferentes valores de x

Estimando Parâmetros do Modelo

Definição

- Assumiremos que as variáveis x e y estão relacionadas de acordo com o modelo de regressão linear simples. Os valores de β_0, β_1 e σ^2 praticamente nunca serão conhecidos pelo investigador.
- Em vez disso, serão fornecidos dados amostrais, compreendendo n pares observados (x_1, y_1), ... , (x_n, y_n), com base nos quais os parâmetros de modelo e a própria reta de regressão real podem ser estimados.
- Supõe-se que essas observações tenham sido obtidas independentemente uma da outra. Ou seja, y_i é o valor observado de uma va Y_i, onde $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ e os n desvios ϵ_1, ..., ϵ_n são vas independentes. A independência de Y_1, ..., Y_n depende da independência dos ϵ_is.

Estimando Parâmetros do Modelo

Definição

- De acordo com o modelo, os pontos observados serão distribuídos nas imediações da reta de regressão real de maneira aleatória.
- A Figura mostra pares observados com duas candidatas à reta de regressão estimada, $y = a_0 + a_1 x$ e $y = b_0 + b_1 x$.
- A primeira reta não é uma estimativa razoável da reta real $y = \beta_0 + \beta_1 x$ porque, os pontos observados deveriam ficar mais próximos dessa reta.
- A primeira reta é uma estimativa mais plausível porque, os pontos observados estão dispersos em torno dessa reta, e não próximos.

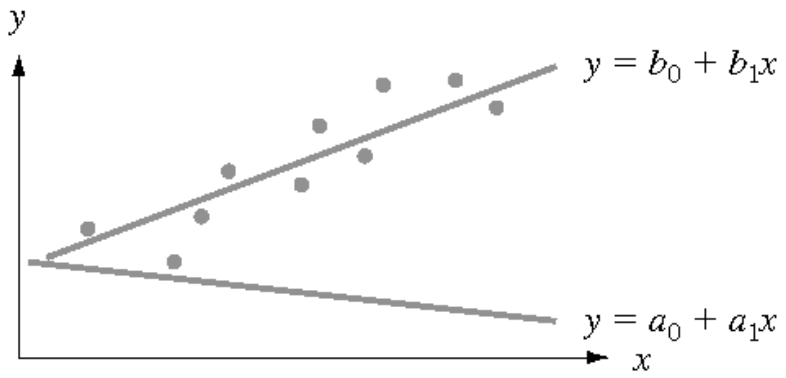


Figura 12.6 Duas diferentes estimativas da reta de regressão real

Estimando Parâmetros do Modelo

Princípio dos Mínimos Quadrados

- O princípio dos mínimos quadrados (Gauss,1777–1855), estabelece que, uma reta oferece uma boa aderência aos dados, se as distâncias verticais (desvios) dos pontos observados em relação à reta são pequenos.
- A medida da aderência é a soma dos quadrados desses desvios. A reta de melhor aderência é, portanto, aquela que tem a menor soma possível de desvios ao quadrado.

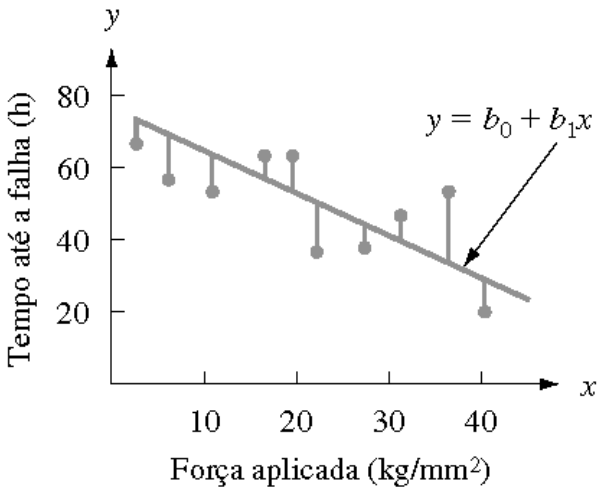


Figura 12.7 Desvios dos dados observados da reta $y = b_0 + b_1x$

Estimando Parâmetros do Modelo

Princípio dos Mínimos Quadrados

- O princípio dos mínimos quadrados:
- O desvio vertical do ponto $(x_i,\ y_i)$ da reta $y = b_0 + b_1 x$

  altura do ponto – altura da reta $= y_i - (b_0 + b_1 x_i)$
- A soma dos desvios quadrados verticais dos pontos $(x_1,\ y_1), \ldots, (x_n,\ y_n)$ à reta é o seguinte:

$$f(b_0, b_1) = \sum_{i=1}^{n} [y_i - (b_0 + b_1 x_i)]^2$$

- As estimativas pontuais de β_0 e β_1, representadas por $\hat{\beta}_0$ e $\hat{\beta}_1$ são denominadas estimativas dos mínimos quadrados, e são aqueles valores que minimizam $f(b_0, b_1)$.
- Ou seja, $\hat{\beta}_0$ e $\hat{\beta}_1$ são tais que $f(\hat{\beta}_0, \hat{\beta}_1) \leq f(b_0, b_1)$ para qualquer b_0 e b_1.
- A reta de regressão estimada ou a reta dos mínimos quadrados é, portanto, a reta cuja equação é $y = \hat{\beta}_0 + \hat{\beta}_1 x$.

Estimando Parâmetros do Modelo

Princípio dos Mínimos Quadrados

- Os valores de minimização de b_0 e b_1 são identificados tomando-se derivadas parciais de $f(b_0, b_1)$ em relação a b_0 e b_1, igualando-as a zero e resolvendo as equações. Temos assim:

$$\frac{\partial f(b_0, b_1)}{\partial b_0} = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-1) = 0$$

$$\frac{\partial f(b_0, b_1)}{\partial b_1} = \sum_{i=1}^{n} 2(y_i - b_0 - b_1 x_i)(-x_i) = 0$$

- Cancelando-se o fator –2 e reorganizando as equações, obtemos o sistema a seguir, denominado equações normais:

$$nb_0 + (\sum x_i)b_1 = \sum y_i$$

$$(\sum x_i)b_0 + (\sum {x_i}^2)b_1 = \sum x_i y_i$$

- As equações normais são lineares nas duas incógnitas b_0 e b_1.

Estimando Parâmetros do Modelo

Princípio dos Mínimos Quadrados

- Resolvendo o sistema: $nb_0 + (\sum x_i)b_1 = \sum y_i$

$$(\sum x_i)b_0 + (\sum x_i^2)b_1 = \sum x_i y_i$$

- A estimativa dos mínimos quadrados do coeficiente angular $\hat{\beta}_1$:

$$\hat{\beta}_1 = b_1 = \frac{S_{xy}}{S_{xx}} \qquad S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n$$

$$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n$$

- A estimativa dos mínimos quadrados do termo constante $\hat{\beta}_0$:

$$\hat{\beta}_0 = b_0 = \frac{\sum y_i - (\sum x_i)b_1}{n}$$

$$\hat{\beta}_0 = b_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- As fórmulas de cálculo de S_{xy} e S_{xx} exigem apenas as estatísticas $\sum x_i$, $\sum y_i, \sum x_i^2, \sum x_i y_i$ e que se minimizem os efeitos do arredondamento.

Estimando Parâmetros do Modelo

Exemplo 1

- O concreto sem finos, fabricado com um agregado rústico nivelado de maneira uniforme e uma pasta de cimento-água, é benéfico em áreas propensas a muita chuva por causa de suas excelentes propriedades de drenagem.
- Utilizaremos uma análise de mínimos quadrados ao estudar como:

  $y =$ porosidade (%) está relacionada com

  $x =$ peso unitário (pcf) em amostras de concreto;
- Considere os dados representativos a seguir, exibidos em um formato tabular conveniente para calcular os valores das estatísticas:

Estimando Parâmetros do Modelo

Exemplo 1

- Considere os dados exibidos para calcular os valores das estatísticas:

Obs	x	y	x^2	xy	y^2
1	99,0	28,8	9.801,00	2.851,20	829,44
2	101,1	27,9	10.221,21	2.820,69	778,41
3	102,7	27,0	10.547,29	2.772,90	729,00
4	103,0	25,2	10.609,00	2.595,60	635,04
5	105,4	22,8	11.109,16	2.403,12	519,84
6	107,0	21,5	11.449,00	2.300,50	462,25
7	108,7	20,9	11.815,69	2.271,83	436,81
8	110,8	19,6	12.276,64	2.171,68	384,16
9	112,1	17,1	12.566,41	1.916,91	292,41
10	112,4	18,9	12.633,76	2.124,36	357,21
11	113,6	16,0	12.904,96	1.817,60	256,00
12	113,8	16,7	12.950,44	1.900,46	278,89
13	115,1	13,0	13.248,01	1.496,30	169,00
14	115,4	13,6	13.317,16	1.569,44	184,96
15	120,0	10,8	14.400,00	1.296,00	116,64
Soma	1640,1	299,8	179.849,73	32.308,59	6.430,06

- Temos:

$$\bar{x} = \frac{1640,1}{15} = 109,34$$

$$\bar{y} = \frac{299,8}{15} = 19,986$$

$$\hat{\beta}_1 = \frac{\sum x_i y_i - (\sum x_i)(\sum y_i)/n}{\sum x_i^2 - (\sum x_i)^2/n}$$

$$= \frac{32308,59 - (1640,1)(299,8)/15}{179849,73 - (1640,1)^2/15} = \frac{-471,572}{521,196} = -0,9047$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} = 19,986 - (-0,9047)109,34 = 118,909$$

Estimando Parâmetros do Modelo

Exemplo 1

- A equação da reta de regressão estimada (reta dos mínimos quadrados) é, portanto:

$$y = \hat{\beta}_0 + \hat{\beta}_1 x$$
$$y = 118{,}909 - 0{,}9047x$$

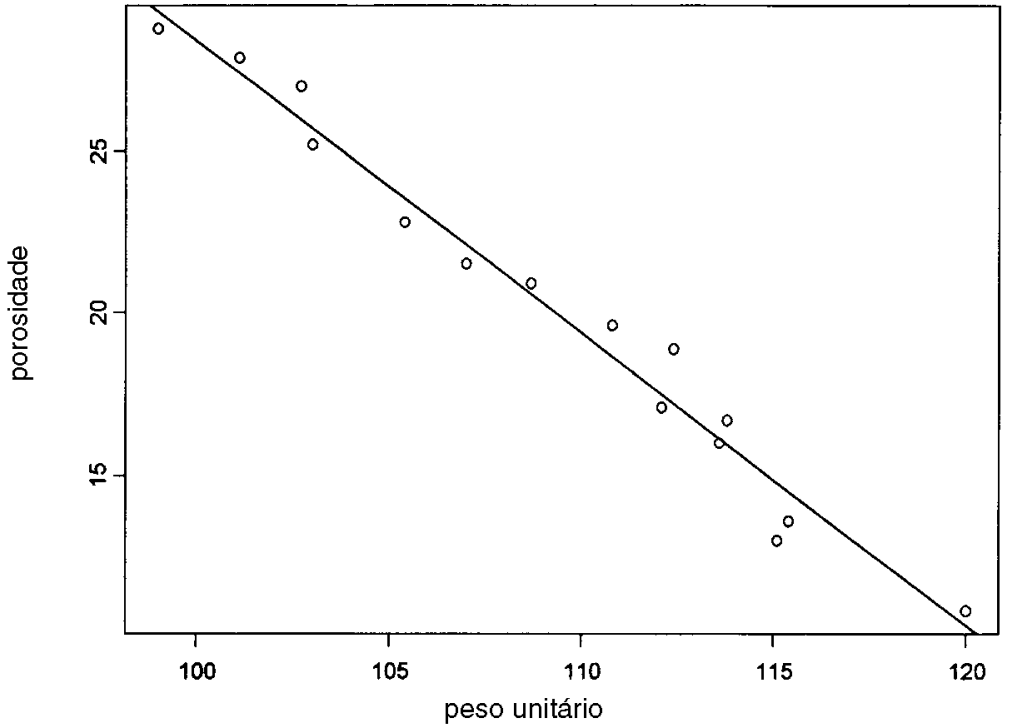


Figura 12.8 Gráfico de dispersão dos dados do Exemplo 12.4 com a reta dos mínimos quadrados sobreposta, feito do S-Plus

- A mudança esperada na porosidade associada com um aumento de 1 pcf no peso unitário é –0,905% (um decréscimo de 0,905%).

Estimando Parâmetros do Modelo

Exemplo 2

- A reta de regressão estimada pode ser usada imediatamente para dois diferentes propósitos. Para um valor fixo x^*, $\hat{\beta}_0 + \hat{\beta}_1 x^*$ fornece tanto:
  - (1) uma estimativa pontual do valor esperado de Y quando $x = x^*$
  - (2) uma previsão pontual do valor Y que resultará de uma única nova observação feita em $x = x^*$.
- Refira-se aos dados sobre peso unitário–porosidade no exemplo anterior. Uma estimativa pontual da porosidade média real de todas as amostras cujo peso unitário é 110 é:

$$\hat{\mu}_{Y=110} = \hat{\beta}_0 + \hat{\beta}_1 x = 118{,}909 - 0{,}9047(110) = 19{,}36\%$$

- Se for selecionada uma única amostra cujo peso unitário é 110 pcf, 19,4% será também uma previsão pontual da porosidade dessa amostra.
- A reta dos mínimos quadrados não deve ser usada para estimar um valor x que esteja muito além da amplitude dos dados, como $x = 90$ ou $x = 135$ no Exemplo. O risco de extrapolação é o de que a relação ajustada possa não ser válida para tais valores x.

Estimando Parâmetros do Modelo

Estimando σ^2 e σ

- O parâmetro σ^2 determina a variabilidade inerente no modelo de regressão.
- Quando σ^2 for um valor grande, significa que os (x_i, y_i) observados se encontram muito dispersos em relação à reta de regressão real;
- Quando σ^2 for um valor pequeno, os pontos observados tenderão a ficar bem próximos da reta real.

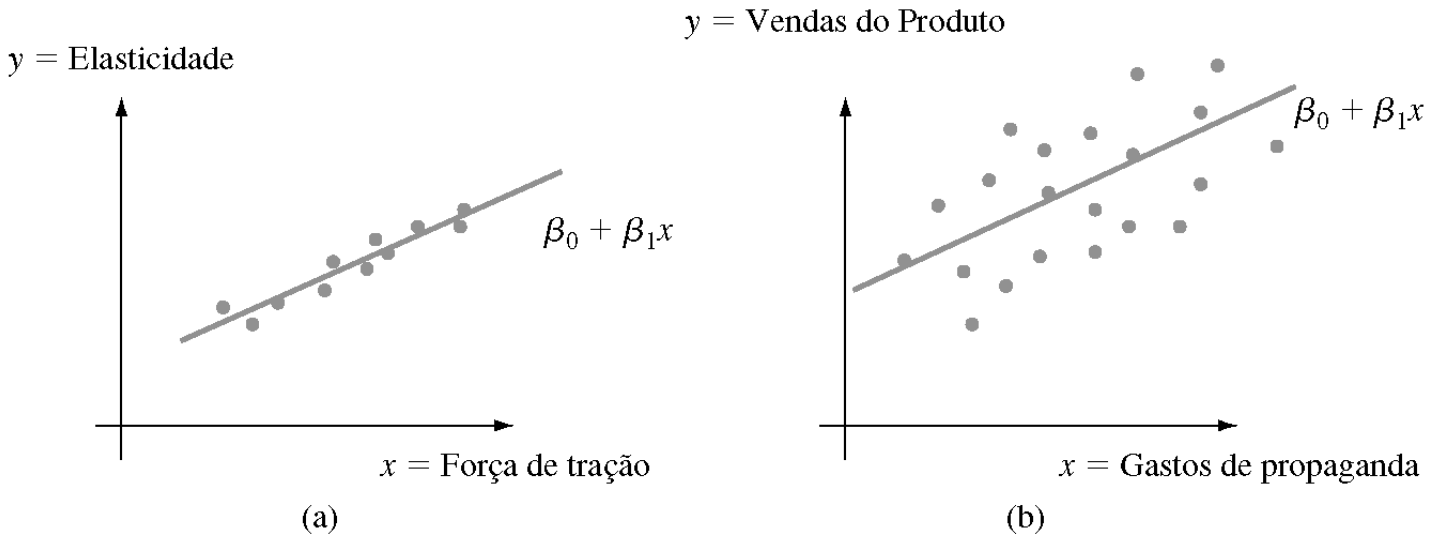


Figura 12.9 Amostra típica de σ^2: (a) pequena; (b) grande

Estimando Parâmetros do Modelo

Estimando σ^2 e σ

- Uma estimativa de σ^2 poderá ser usada nas fórmulas do Intervalo de Confiança (IC) e procedimentos de teste de hipótese.
- Pelo fato de não se conhecer a equação da reta real, a estimativa se baseia em até que ponto as observações amostrais se desviam da reta estimada.
- Muitos desvios grandes (residuais) sugerem um valor grande de σ^2, ao passo que todos os desvios de pequena magnitude sugerem que σ^2 é pequeno.

- Os **valores previstos (ou ajustes)** $\hat{y}_1, \hat{y}_2, ..., \hat{y}_n$ são obtidos substituindo-se sucessivamente x_1, x_2, ..., x_n na equação da reta de regressão estimada: $\hat{y}_1 = \hat{\beta}_0 + \hat{\beta}_1 x_1, ..., \hat{y}_n = \hat{\beta}_0 + \hat{\beta}_1 x_n$.
- Os **resíduos** são os desvios verticais $y_1 - \hat{y}_1$, $y_2 - \hat{y}_2$, ..., $y_n - \hat{y}_n$ dos pontos observados até a reta estimada.

Estimando Parâmetros do Modelo

Estimando σ^2 e σ

- Se todos forem resíduos pequenos, a variabilidade nos valores y observados devem decorrer da relação linear entre x e y;
- Já, muitos resíduos grandes sugerem grande variabilidade inerente em y, em relação à quantidade decorrente da relação linear.

- Teoricamente a soma dos resíduos deve ser zero. Na prática, a soma pode desviar um pouco de zero em decorrência do arredondamento.

Estimando Parâmetros do Modelo

Exemplo 3

- A alta densidade populacional do Japão provocou problemas relacionados à remoção de lixo. No desenvolvimento de uma máquina de compressão para processamento do lodo de esgoto, foi necessário relacionar dados sobre a umidade de grânulos comprimidos (y, em %) com a taxa de filtragem da máquina (x, em kg–DS/m/h).

x	125,3	98,2	201,4	147,3	145,9	124,7	112,2	120,2	161,2	178,9
y	77,9	76,8	81,5	79,8	78,2	78,3	77,5	77,0	80,1	80,2

x	159,5	145,8	75,1	151,4	144,2	125,0	198,8	132,5	159,6	110,7
y	79,9	79,0	76,7	78,2	79,5	78,1	81,5	77,0	79,0	78,6

$$\sum x_i = 2817{,}9 \quad \sum y_i = 1574{,}8 \quad \sum x_i^2 = 415.949{,}85 \quad \sum x_i y_i = 222.657{,}88$$

$$S_{xy} = \sum x_i y_i - (\sum x_i)(\sum y_i)/n = 776{,}434$$

$$S_{xx} = \sum x_i^2 - (\sum x_i)^2/n = 18921{,}8295$$

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{776{,}434}{18921{,}8295} = 0{,}041$$

$$\bar{x} = 140{,}895 \quad \bar{y} = 78{,}74$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$
$$= 78{,}74 - (0{,}041)140{,}895 = 72{,}958$$

Estimando Parâmetros do Modelo

Exemplo 3

- Assim, a reta dos mínimos quadrados é (para obter precisão numérica):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{y}_i = 72{,}958547 + 0{,}04103377 x_i$$

- Para $x_1 = 125{,}3$ e $y_1 = 77{,}9$ temos:

$$\hat{y}_1 = 72{,}958547 + 0{,}04103377(125{,}3) = 78{,}1$$

- Onde o resíduo é:

$$y_1 - \hat{y}_1 = 77{,}9 - 78{,}1 = -0{,}20$$

Estimando Parâmetros do Modelo

Exemplo 3

- Assim, a reta dos mínimos quadrados é (para obter precisão numérica):

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

$$\hat{y}_i = 72{,}958547 + 0{,}04103377 x_i$$

- Para $x_1 = 125{,}3$ e $y_1 = 77{,}9$ temos:

$$\hat{y}_1 = 72{,}958547 + 0{,}04103377(125{,}3) = 78{,}1$$

- Onde o resíduo é:

$$y_1 - \hat{y}_1 = 77{,}9 - 78{,}1 = -0{,}20$$

- Um resíduo positivo corresponde a um ponto acima do gráfico da reta dos mínimos quadrados;
- Um resíduo negativo resulta de um ponto disposto abaixo da reta.

Estimando Parâmetros do Modelo

Exemplo 3

- Todos os valores previstos e resíduos são exibidos na tabela a seguir.

Obs	Líquido Filtrado	Concentração de Umidade	Ajuste	Resíduo
1	125,3	77,9	78,100	−0,200
2	98,2	76,8	76,988	−0,188
3	201,4	81,5	81,223	0,277
4	147,3	79,8	79,003	0,797
5	145,9	78,2	78,945	−0,745
6	124,7	78,3	78,075	0,225
7	112,2	77,5	77,563	−0,063
8	120,2	77,0	77,891	−0,891
9	161,2	80,1	79,573	0,527
10	178,9	80,2	80,299	−0,099
11	159,5	79,9	79,503	0,397
12	145,8	79,0	78,941	0,059
13	75,1	76,7	76,040	0,660
14	151,4	78,2	79,171	−0,971
15	144,2	79,5	78,876	0,624
16	125,0	78,1	78,088	0,012
17	198,8	81,5	81,116	0,384
18	132,5	77,0	78,396	−1,396
19	159,6	79,0	79,508	−0,508
20	110,7	78,6	77,501	1,099

Estimando Parâmetros do Modelo

Definição

- Da mesma forma que os desvios da média no caso de uma única amostra foram somados para obter a estimativa $S^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$, a estimativa de σ^2 na análise de regressão tem por base elevar ao quadrado e somar os resíduos.

- A soma dos quadrados dos erros, SQE, (que equivale à soma dos quadrados dos resíduos):

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

- A estimativa de σ^2 é:

$$\hat{\sigma}^2 = S^2 = \frac{SQE}{n-2} = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}$$

- O divisor $n-2$ em S^2 é o número de graus de liberdade (gl) associados com a estimativa. Já que 2 parâmetros $\hat{\beta}_0$ e $\hat{\beta}_1$ devem ser estimados primeiro.

Estimando Parâmetros do Modelo

Exemplo 4

- No exemplo anterior, a soma dos quadrados dos erros é:

Resíduo
−0,200
−0,188
0,277
0,797
−0,745
0,225
−0,063
−0,891
0,527
−0,099
0,397
0,059
0,660
−0,971
0,624
0,012
0,384
−1,396
−0,508
1,099

$$SQE = \sum_{i=1}^{20} (y_i - \hat{y}_i)^2 = (0{,}200)^2 + \ldots + (1{,}099)^2 = 7{,}968$$

- A estimativa de σ^2 é:

$$\hat{\sigma}^2 = S^2 = \frac{SQE}{n-2} = \frac{7{,}968}{20-2} = 0{,}4427$$

- O desvio padrão estimado é:

$$\hat{\sigma} = S = \sqrt{0{,}4427} = 0{,}665$$

Estimando Parâmetros do Modelo

Exemplo 4

- Uma formula alternativa para o cálculo de SQE é a seguinte:

$$SQE = \sum y_i^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

- Que resulta da substituição de $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ em:

$$SQE = \sum (y_i - \hat{y}_i)^2$$

- A formula alternativa é particularmente sensível aos efeitos de arrendondamento de $\hat{\beta}_0$ e $\hat{\beta}_1$, de modo que usar o maior número de dígitos possível nos cálculos é recomendado.

Estimando Parâmetros do Modelo

Coeficiente de Determinação

- Nos gráficos da Figura as alturas dos diferentes pontos variam substancialmente, indicando que há muita variabilidade nos valores y observados.

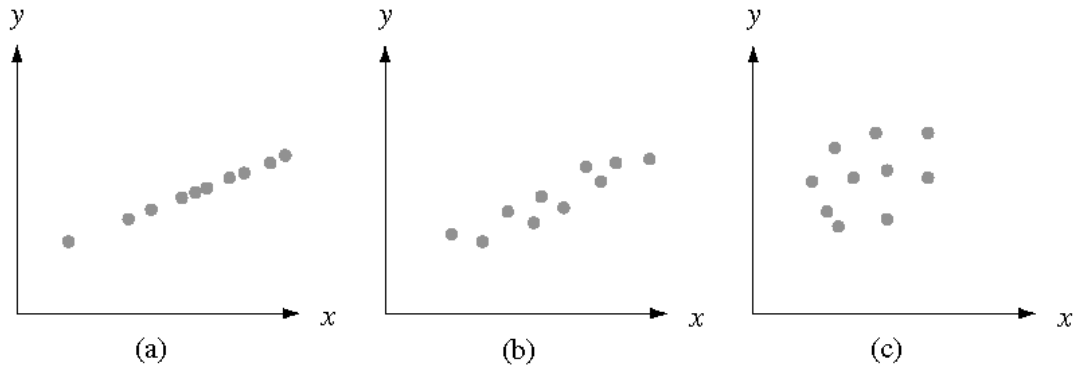
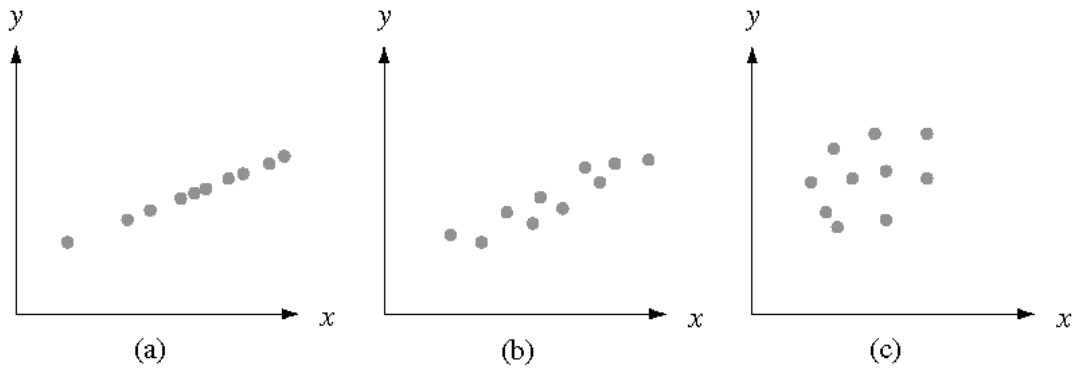


Figura 12.10 Usando o modelo para explicar a variação de *y*: (a) dados para os quais toda a variação é justificada; (b) dados para os quais grande parte da variação é justificada; (c) dados para os quais pouca variação é justificada

- No entanto, no primeiro os pontos se dispõem exatamente numa reta;
- No segundo os desvios com relação a reta são pequenos;
- No terceiro há uma variação significativa em torno da reta.

Estimando Parâmetros do Modelo

Coeficiente de Determinação



- No primeiro, a variação amostral em y pode ser atribuída 100% ao fato de x e y estarem relacionados linearmente.
- No segundo, nesse caso, que grande parte da variação observada de y pode ser atribuída à relação linear;
- No terceiro, modelo de regressão linear simples não consegue explicar a variação em y relacionando-o a x.

Estimando Parâmetros do Modelo

Coeficiente de Determinação

- A soma dos quadrados dos erros SQE pode ser interpretada como uma medida da quantidade de variação em y deixada inexplicada pelo modelo (que não pode ser atribuída a uma relação linear).

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

- Uma medida quantitativa da quantidade total de variação nos valores observados de y é dada pela soma total dos quadrados SQT

$$SQT = S_{yy} = \sum (y_i - \bar{y})^2$$

- A soma total dos quadrados é a soma dos desvios quadrados ao redor da média amostral dos valores observados de y.

Estimando Parâmetros do Modelo

Coeficiente de Determinação

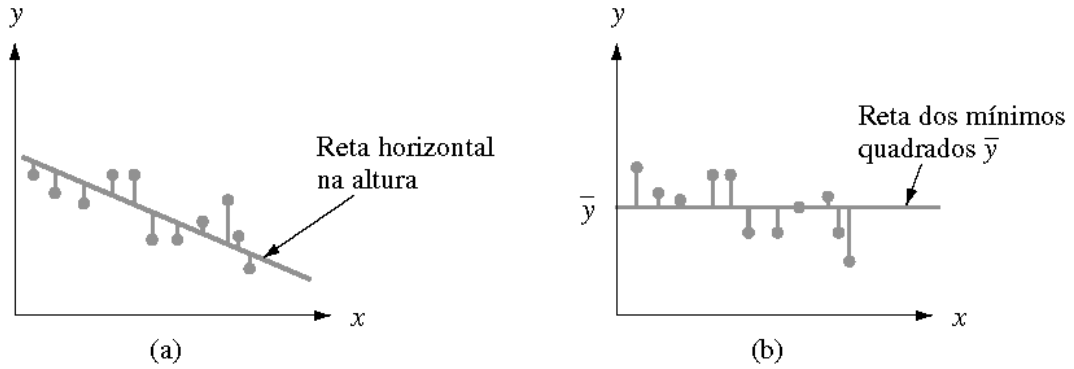


Figura 12.11 Somas dos quadrados ilustradas: (a) SQE = soma dos desvios quadrados em torno da reta dos mínimos quadrados; (b) SQT = soma dos quadrados total em torno da reta horizontal

$$SQE < SQT$$

- A razão SQE/SQT é a proporção da variação total que não pode ser explicada pelo modelo de regressão linear simples.

Estimando Parâmetros do Modelo

Coeficiente de Determinação

- O coeficiente de determinação, representado por r^2, é dado por:

$$r^2 = 1 - \frac{SQE}{SQT}$$

- Esse coeficiente é interpretado como a proporção da variação de y observada que pode ser explicada pelo modelo de regressão linear simples (considerando uma relação linear aproximada entre y e x).

- Quanto mais alto o valor de r^2, mais o modelo de regressão linear simples consegue explicar a variação y.
- Se r^2 for pequeno, em geral o analista vai querer procurar um modelo alternativo (um modelo não-linear ou um modelo de regressão múltipla com
- mais de uma variável independente) que possa explicar mais eficientemente a variação y.

Estimando Parâmetros do Modelo

Exemplo

- Considere uma expressão alternativa para o numerador da variança amostral:

$$\begin{aligned} S_{xx} = \sum(x_i - \bar{x})^2 &= \sum({x_i}^2 - 2\bar{x}x_i + \bar{x}^2) \\ &= \sum {x_i}^2 - 2\bar{x}\sum x_i + \sum \bar{x}^2 \\ &= \sum {x_i}^2 - 2\bar{x}n\bar{x} + n\bar{x}^2 \\ &= \sum {x_i}^2 - n\bar{x}^2 \\ &= \sum {x_i}^2 - (\sum x_i)^2/n \end{aligned}$$

- Voltando as medidas de desvio SQE eSQT:

$$SQE = \sum(y_i - \hat{y}_i)^2 = \sum {y_i}^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$

$$SQT = S_{yy} = \sum(y_i - \bar{y})^2 = \sum {y_i}^2 - \left(\sum y_i\right)^2 / n$$

Estimando Parâmetros do Modelo

Exemplo

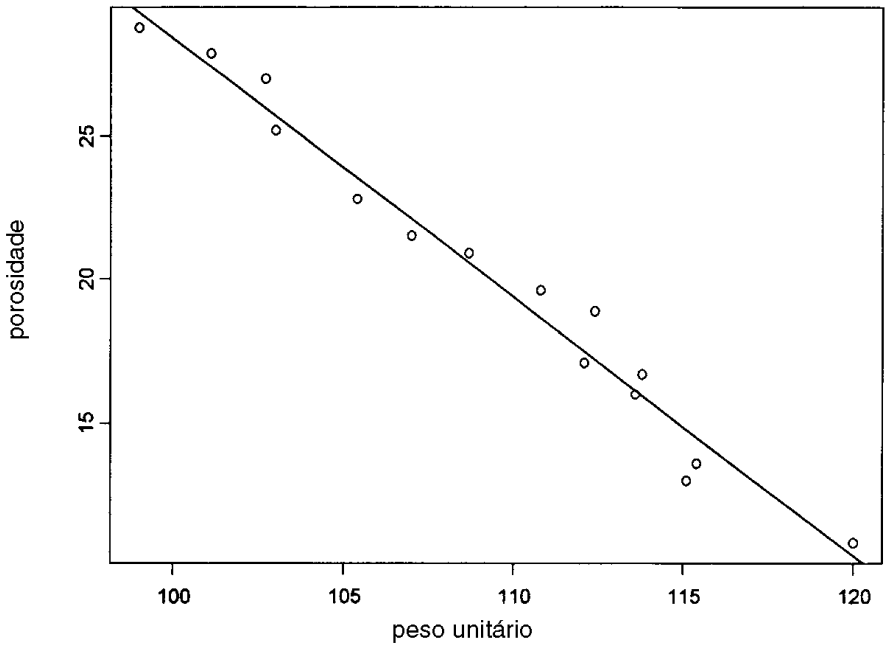
- O gráfico de dispersão dos dados sobre o concreto sem finos do Exemplo 1, prognostica um valor r2 alto r^2 :

$\hat{\beta}_0 = 118{,}909917$
$\hat{\beta}_1 = -0{,}9047x$

$\sum y_i{}^2 = 6430{,}06$

$\sum y_i = 299{,}8$

$\sum x_i y_i = 32.308{,}59$



$$SQT = \sum y_i{}^2 - (\sum y_i)^2/n$$
$$= 6430{,}06 - (299{,}8)^2/15 = 438{,}057333$$

$$r^2 = 1 - \frac{SQE}{SQT}$$

$$SQE = \sum y_i{}^2 - \hat{\beta}_0 \sum y_i - \hat{\beta}_1 \sum x_i y_i$$
$$= 6430{,}06 - (118{,}909917)(299{,}8) - (-0{,}90473066)(32.308{,}59)$$
$$= 11{,}4388$$

Estimando Parâmetros do Modelo

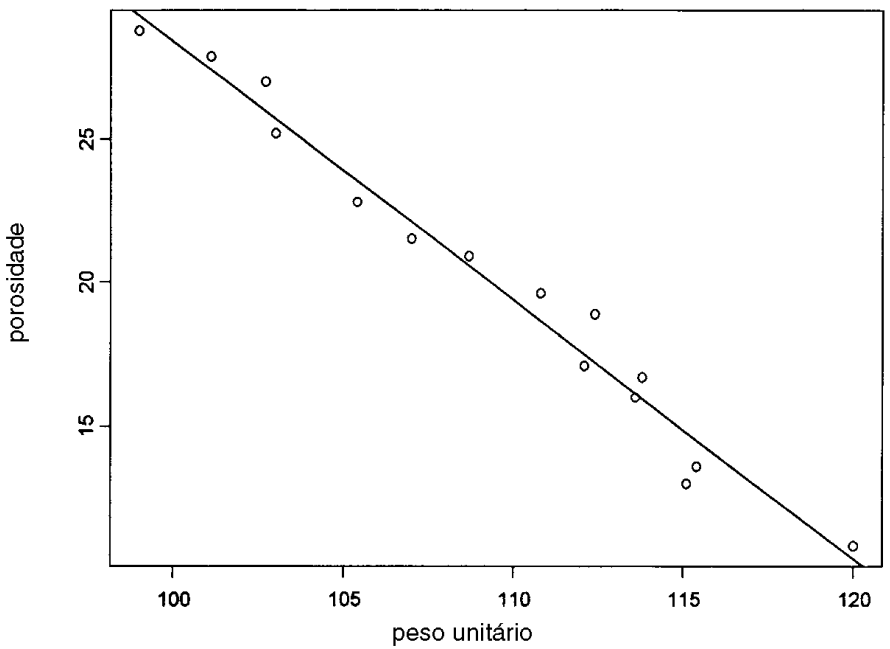
Exemplo

- O gráfico de dispersão dos dados sobre o concreto sem finos do Exemplo 1, prognostica um valor r2 alto r^2 :

$SQT = 438{,}057333$

$SQE = 11{,}4388$

$$r^2 = 1 - \frac{SQE}{SQT}$$
$$= 1 - \frac{11{,}4388}{438{,}057333} = 0{,}974$$



- Ou seja, 97,4% da variação observada na porosidade pode ser explicada pela relação linear aproximada entre porosidade e peso unitário do concreto.

Correlação

Coeficiente de Correlação Amostral r

- Dados os n pares de observações $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$, é natural que se fale de x e y como tendo uma relação positiva, se xs grandes estiverem pareados com ys grandes e xs pequenos com ys pequenos.
- De modo semelhante, se xs grandes estiverem pareados com ys pequenos e xs pequenos com ys grandes, então está implícita uma relação negativa entre as variáveis.
- Considere o seguinte termo:

$$S_{xy} = \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^{n} x_i\, y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n$$

- Se a relação for fortemente positiva, um x_i acima da média $\bar{x}$ tenderá a estar pareado com um y_i, acima da média $\bar{y}$, de modo que $(x_i - \bar{x})(y_i - \bar{y}) > 0$, e esse produto será também positivo quando ambos, x_i e y_i estiverem abaixo das respectivas médias.

Correlação

Coeficiente de Correlação Amostral r

- Portanto, uma relação positiva significa que S_{xy} será positiva.
- Analogamente, quando a relação for negativa, S_{xy} será negativa, visto que a maioria dos produtos $(x_i - \bar{x})(y_i - \bar{y})$ será negativa.

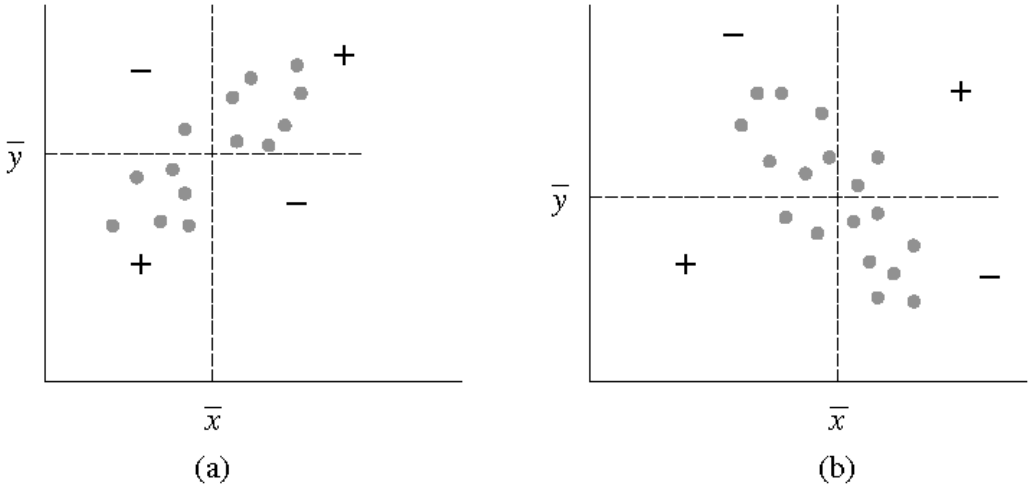


Figura 12.19 (a) Gráfico de dispersão com S_{xy} positiva; (b) gráfico de dispersão com S_{xy} negativa [+ médias $(x_i - \bar{x})(y_i - \bar{y}) > 0$, e − médias $(x_i - \bar{x})(y_i - \bar{y}) < 0$]

Correlação

Coeficiente de Correlação Amostral r

- Uma condição razoável a ser imposta a qualquer medida sobre quão forte x e y estão relacionados não deve depender das unidades particulares usadas para medi-los. Essa condição é alcançada modificando-se S_{xy} para obter o coeficiente de correlação amostral.
- O coeficiente de correlação amostral de n pares $(x_1, y_1), \ldots, (x_n, y_n)$ é:

$$r = \frac{S_{xy}}{\sqrt{\sum(x_i - \bar{x})^2}\sqrt{\sum(y_i - \bar{y})^2}} = \frac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$$

Correlação

Propriedades de r

- As propriedades mais importantes de r são as seguintes:

1. O valor de r não depende de qual das duas variáveis em estudo é chamada de x e qual é chamada de y.
2. O valor de r independe das unidades com as quais x e y são medidos.
3. $-1 \leq r \leq 1$
4. $r = 1$ se, e somente se, todos os pares $(x_i,\ y_i)$ estiverem alinhados em linha reta com um coeficiente angular positivo; e

   $r = -1$ se, e somente se, todos os pares $(x_i,\ y_i)$ estiverem alinhados com um coeficiente angular negativo.
5. O quadrado do coeficiente de correlação amostral fornece o valor do coeficiente de determinação que resultaria de um ajuste do modelo de regressão linear simples – em símbolos, $(r)^2 = r^2$.

Correlação

Propriedades de r

- A Propriedade 2 equivale a dizer que r não muda se:

  Houver mudança na escala de medida dos dados: cada x_i for substituído por cx_i e cada y_i for substituído por dx_i;

  Houver deslocamento no eixo de medida dos dados: cada x_i for substituído por $x_i - a$ e y_i por $y_i - b$.

- A Propriedade 3 expressa que o valor máximo de r, correspondente ao maior grau possível de relação positiva, é $r = 1$, ao passo que a relação mais negativa é identificada com $r = -1$.

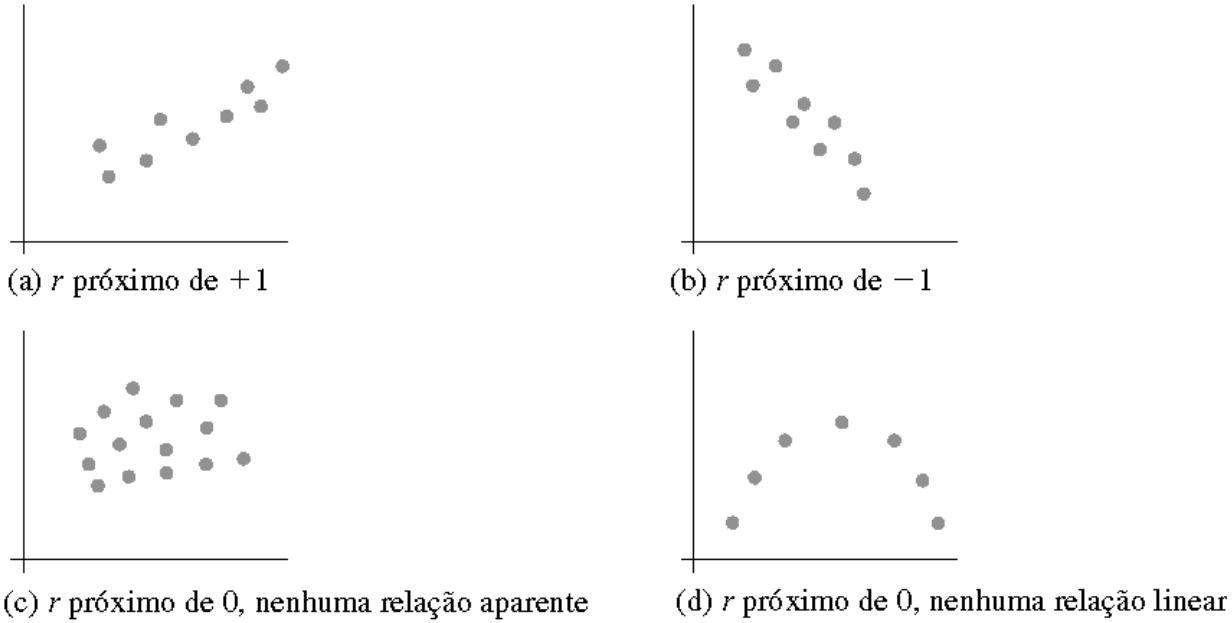
Correlação

Propriedades de r

- Segundo a Propriedade 4, as maiores correlações positivas e negativas são alcançadas somente quando todos os pontos estendem-se sobre uma linha reta. Qualquer outra configuração de pontos, mesmo se a configuração sugerir uma relação determinística entre as variáveis, produzirá um valor r menor que 1 em magnitude absoluta.
- Portanto, r mede o grau de relação linear entre variáveis. Um valor de r próximo de 0 não é uma evidência de que não existe uma relação forte, mas apenas de que falta uma relação linear, de modo que esse valor de r deve ser interpretado com cuidado.
- A Propriedade 5 mostra que a proporção de variação na variável dependente explicada pelo ajuste do modelo de regressão linear simples não depende de qual variável desempenha esse papel.

Correlação
Propriedades de r

- A Figura ilustra várias configurações de pontos para a diferentes valores de r.



(a) r próximo de $+1$

(b) r próximo de -1

(c) r próximo de 0, nenhuma relação aparente

(d) r próximo de 0, nenhuma relação linear

Figura 12.20 Gráficos de dados para diferentes valores de r

Correlação

Propriedades de r

- Uma pergunta frequente é: "Quando é possível dizer que existe uma correlação forte entre as variáveis e quando a correlação é fraca?".
- Uma regra prática razoável é afirmar que:
- a correlação é fraca se $0 \leq |r| \leq 0{,}5$;
- a correlação é forte se $0{,}8 \leq |r| \leq 1$;
- a correlação é moderada em caso contrário.

Correlação

Exemplo

- Uma avaliação precisa sobre a produtividade do solo é fundamental para o planejamento racional do uso da terra. Infelizmente, não é tão fácil de estabelecer um índice de produtividade do solo aceitável. Uma das dificuldades é que a produtividade é determinada parcialmente pela cultura plantada, e a relação entre a produção de duas diferentes culturas plantadas no mesmo solo pode não ser muito forte.
- Apresenta-se os dados a seguir sobre a produção de milho x e a produção de amendoim e y (medidas em mT/Ha) de oito diferentes tipos de solo.

x	2,4	3,4	4,6	3,7	2,2	3,3	4,0	2,1
y	1,33	2,12	1,80	1,65	2,00	1,76	2,11	1,63

Correlação

Exemplo

x	2,4	3,4	4,6	3,7	2,2	3,3	4,0	2,1
y	1,33	2,12	1,80	1,65	2,00	1,76	2,11	1,63

- Com base em: $r = \dfrac{S_{xy}}{\sqrt{S_{xx}}\sqrt{S_{yy}}}$ $\quad S_{xy} = \sum_{i=1}^{n} x_i\, y_i - \left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)/n$
- Onde: $S_{xx} = \sum_{i=1}^{n} {x_i}^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n$ $\quad S_{yy} = \sum_{i=1}^{n} {y_i}^2 - \left(\sum_{i=1}^{n} y_i\right)^2/n$
- Temos:
- $\sum x_i = 25{,}7$; $\quad \sum y_i = 14{,}40$;
- $\sum {x_i}^2 = 88{,}31$; $\sum {y_i}^2 = 26{,}4324$; $\quad \sum x_i\, y_i = 46{,}856$;

- Com isso:
- $S_{xx} = 88{,}31 - \dfrac{(25{,}7)^2}{8} = 88{,}31 - 82{,}56 = 5{,}75$
- $S_{yy} = 26{,}4324 - \dfrac{(14{,}40)^2}{8} = 0{,}5124$
- $S_{xy} = 46{,}856 - \dfrac{(25{,}7)(14{,}40)}{8} = 0{,}5960$

Logo:

$$r = \frac{0{,}5960}{\sqrt{5{,}75}\sqrt{0{,}5124}} = 0{,}347$$