UNIVERSIDADE DE SÃO PAULO
Faculdade de Zootecnia e Engenharia de Alimentos

ZAB1111 – ESTATÍSTICA BÁSICA

Aula 17

CORRELAÇÃO E REGRESSÃO LINEAR SIMPLES

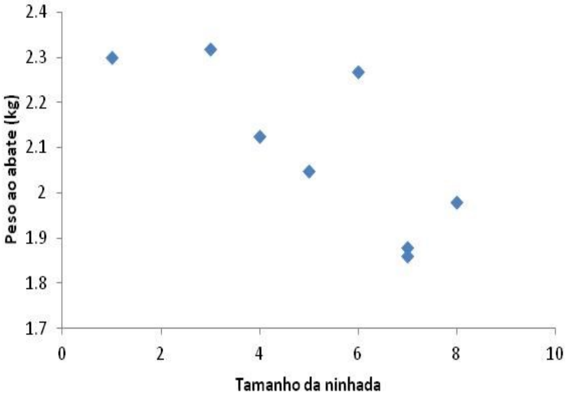
Material preparado pelo Prof. Dr. César Gonçalves de Lima – FZEA/USP

A **análise de correlação** visa quantificar **o grau** de relacionamento entre duas variáveis quantitativas, como por exemplo:

- Qual o grau de relacionamento entre o nível energético da ração consumida e o peso dos frangos de corte?
- E entre o número de horas de estudos e a nota de uma avaliação?
- E entre os pesos de frangos de corte avaliados aos 7, 14, 21, 28 e 35 dias de idade nas mesmas aves?
- E entre a o pH do solo e a produção de cana de açúcar?
- E entre a idade de um imóvel e o valor do seu aluguel?
- E entre a estatura dos pais e a estatura dos filhos?

Ferramenta útil: gráfico de dispersão!

Exemplo 1. Existe alguma relação entre o peso médio de coelhos ao abate (Y) e o tamanho da ninhada (X)?



O aumento no tamanho da ninhada tende a provocar uma diminuição no peso dos coelhos ao abate.

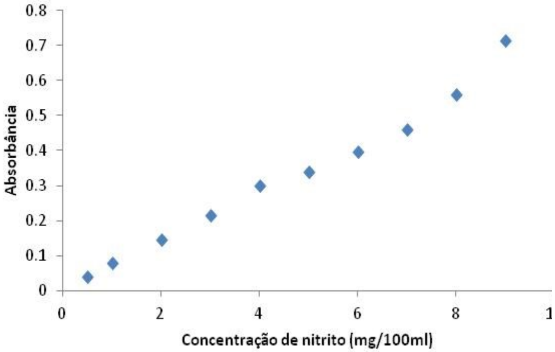
Coelhos de ninhadas maiores tendem a alcançar pesos menores ao abate.

Correlação linear simples

A **análise de regressão** é um método simples e muito utilizado no estabelecimento de fórmulas empíricas envolvendo duas ou mais variáveis.

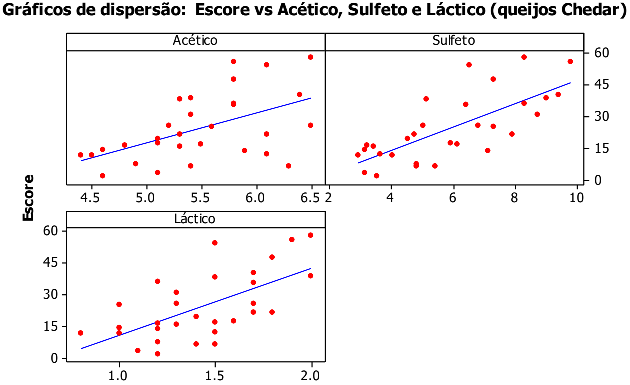
- Como expressar o relacionamento entre duas características através de uma fórmula matemática?
- Consigo determinar o valor de uma grandeza, partindo do conhecimento dos valores de outras grandezas?
- Consigo fazer previsões conhecendo a relação funcional entre as variáveis?
- Ferramenta importante: **gráfico de dispersão**.

Exemplo 2 (Regressão linear simples) Como explicar a relação entre a absorbância (Y) e a concentração de nitrito (X, em mg/100ml) em amostras de mortadela?

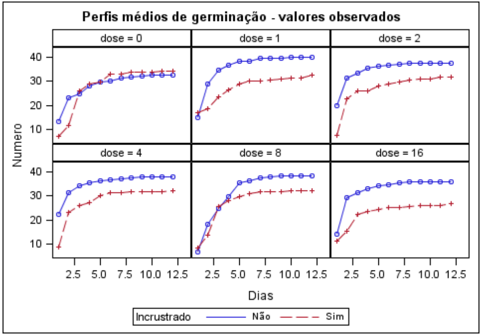


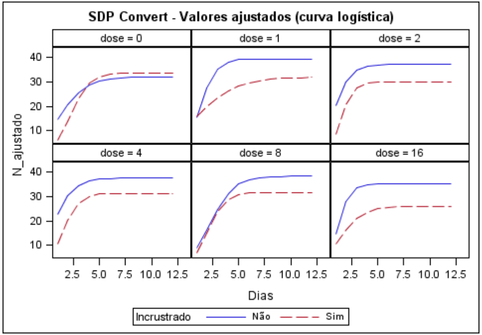
Uma reta parece explicar bem a relação entre a absorbância e o aumento na concentração de nitrito.

Exemplo 3. (Regressão linear múltipla) Podemos relacionar o escore obtido na análise sensorial (y) do queijo Cheddar com a concentração de ácido acético (x_1), de sulfeto de hidrogênio (x_2) e de ácido láctico (x_3) na sua composição química?



Exemplo 4. (Regressão não linear) Como explicar a porcentagem de germinação de sementes do capim Marandú ao longo do tempo?





A germinação de sementes do capim Marandú ao longo do tempo tem um comportamento similar nos diferentes tratamentos.

Curvas logísticas podem explicar bem o processo de germinação de sementes de capim Marandú.

6.1. CORRELAÇÃO LINEAR

Queremos saber se existe alguma relação de dependência entre um par de variáveis quantitativas e, ao invés de procurarmos um modelo que as relacionem, buscamos somente **quantificar** o grau do possível relacionamento linear existente entre elas.

Exemplo. Quantificar o grau de relacionamento entre:

i) O consumo de fumo e a incidência de doenças do coração.

ii) O peso ao nascer e o peso ao abate de coelhos.

iii) A pressão arterial e o índice de massa corpórea (IMC).

iv) O número de horas de estudo e a nota de prova etc.

Ferramenta importante: **gráfico de dispersão**

O <u>coeficiente de correlação linear de Pearson</u> é uma medida do grau de relacionamento linear entre duas <u>variáveis quantitativas</u> X e Y e é definido como:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)\,var(Y)}} \qquad -1 \le \rho(X,Y) \le 1$$

O sinal de $\rho(X,Y)$ indica o sentido da dependência entre X e Y:

- O <u>sinal positivo</u> indica que os valores de X e Y crescem no mesmo sentido ou que são grandezas <u>diretamente</u> proporcionais.
- O <u>sinal negativo</u> indica que os valores de X e Y crescem em sentidos opostos ou que são grandezas <u>inversamente</u> proporcionais.
- Um valor $\rho(X,Y)$= 0 indica que não existe qualquer relação de dependência linear entre estas variáveis.

Situações comuns podem ser visualizadas nos gráficos apresentados a seguir:

$\rho(X,Y) \cong 1$ $\rho(X,Y) \cong -1$ $\rho(X,Y) \cong 0$

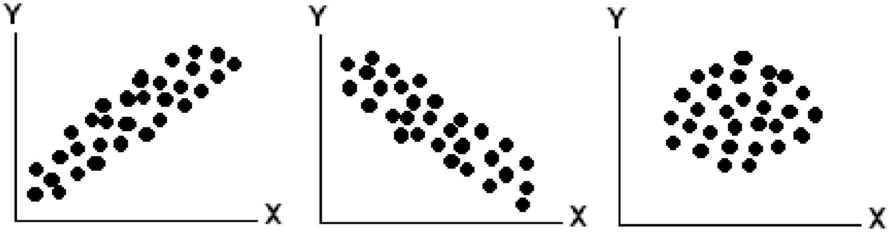


Figura 11. Gráficos de dispersão e coeficientes de correlação.

Nas inferências sobre o parâmetro $\rho(X,Y)$ nós usaremos o valor do <u>coeficiente de correlação linear amostral de Pearson</u>, $r(X,Y)$, para obter sua melhor estimativa.

O coeficiente de correlação linear amostral de Pearson é calculado pela expressão:

$$\boldsymbol{r(X,Y)} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

$$= \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}\left(\sum x_i\right)^2\right]\left[\sum y_i^2 - \frac{1}{n}\left(\sum y_i\right)^2\right]}}$$

Não se preocupem!
A calculadora pode fazer esses cálculos de graça!

Nas pesquisas em que se busca quantificar o grau de relacionamento entre variáveis utilizam-se dados amostrais.

Para inferir sobre esse relacionamento nas populações de onde foram extraídos os dados, podem ser usados testes de hipóteses para o coeficiente de correlação, como os apresentados a seguir.

1) Teste de independência das variáveis X e Y

- Hipóteses: $H_0: \rho(X,Y) = 0$ (as variáveis são independentes)

  $H_a: \rho(X,Y) \neq 0$ (as variáveis são dependentes)

- Estatística: $T = \frac{r(X,Y)\sqrt{n-2}}{\sqrt{1-r^2(X,Y)}}$ que sob H_0 tem distribuição $t_{(n-2)}$.

2) Teste H0: $\boldsymbol{\rho(X,Y) = \rho_0}$ (onde –1< $\boldsymbol{\rho_0}$ < 1 e $\boldsymbol{\rho_0 \neq 0}$)

H_0: $\rho(X,Y) = \rho_0$
H_a: $\rho(X,Y) \neq \rho_0$ (H_a: $\rho(X,Y) > \rho_0$ ou H_a: $\rho(X,Y) < \rho_0$)

- Estatística: Z = $\frac{z-\mu_Z}{\sigma_Z}$, que sob H_0 tem distribuição $N(0,1)$ e onde:

$$z = \frac{1}{2} ln \left[\frac{1+r(X,Y)}{1-r(X,Y)}\right], \quad \mu_z = \frac{1}{2} ln \left[\frac{1+\rho_0}{1-\rho_0}\right] \text{ e } \sigma_z = \frac{1}{\sqrt{n-3}}$$

Exemplo 6.1. Com o objetivo de estudar a relação entre o peso médio de coelhos ao abate (Y), em quilogramas, e o tamanho de ninhada (X), foram coletados na granja do Campus os seguintes dados:

x	4	8	6	1	7	3	7	5
y	2.125	1.980	2.270	2.300	1.880	2.320	1.860	2.050

a) Calcular o coeficiente de correlação e interpretar o seu valor.

b) Testar a independência entre as variáveis X e Y, ao nível de significância de 5%, ou seja, testar se o peso médio de coelhos ao abate, independe do tamanho da ninhada onde ele nasceu.

Resolução:

O coeficiente de correlação amostral é calculado como:

$$r(X,Y) = \frac{83{,}650 - \frac{(41)(16{,}785)}{8}}{\sqrt{\left[249 - \frac{(41)^2}{8}\right]\left[35{,}458 - \frac{(16{,}785)^2}{8}\right]}} = \frac{-2{,}373}{3{,}061} = -0{,}775$$

O coeficiente de correlação $(-0{,}775)$ mostra a existência da <u>dependência linear negativa</u> e relativamente alta entre o peso médio de coelhos ao abate e o tamanho de ninhada, indicando que "quanto maior o tamanho da ninhada menor será o peso médio dos coelhos ao abate".

O peso médio de coelhos ao abate e o tamanho da ninhada são grandezas inversamente proporcionais!

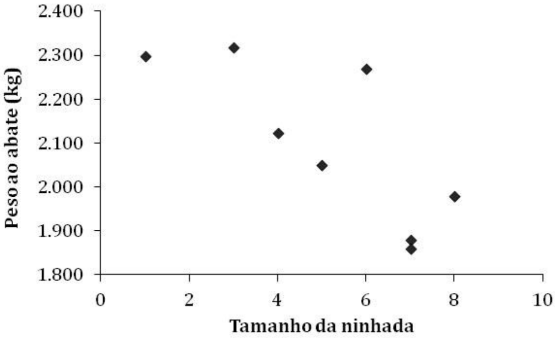


Figura 12. Gráfico de dispersão do peso médio de coelhos ao abate e tamanho de ninhada.

b) H_0: $\rho(X,Y) = 0$ (peso <u>independente</u> do peso da ninhada)

H_a: $\rho(X,Y) \neq 0$ (peso <u>dependente</u> do peso da ninhada)

- Estatística do teste: T = $\frac{r(X,Y)\sqrt{6}}{\sqrt{1-r^2(X,Y)}} \sim t_{(6)}$.
- Da Tábua III, α = 5%, obtemos t_c = 2,45 $\Rightarrow$ $RC(5\%) = \{|t| > 2{,}45\}$
- Da amostra: $T_{calc} = \frac{-0{,}775\sqrt{6}}{\sqrt{1-(-0{,}775)^2}}$ = –3,00 $\in RC(5\%)$

  Rejeitamos a hipótese H_0 e concluímos que existe uma <u>dependência linear negativa e significativa</u> entre o peso médio de coelhos ao abate e o tamanho da ninhada.

Exemplo 6.2. Com o intuito de testar a hipótese de que a correlação entre o ganho de peso e a quantidade de matéria seca ingerida por bovinos da raça Nelore é superior a 0,70, foram utilizados os dados de um experimento com 18 desses animais, resultando em $r(X, Y)$ = 0,81. O que podemos concluir sobre a hipótese, ao nível de significância de α = 1%?

Resolução:

- H_0: $\rho(X, Y) = 0{,}70$
  H_a: $\rho(X, Y) > 0{,}70$
- Estatística do teste: Z = $\frac{z-\mu_z}{\sigma_z} \sim N(0{,}1)$
- Da Tábua I, para α = 1%, z_{tab} = 2,33 $\Rightarrow RC(1\%)$ = {z > 2,33}

- Da amostra de n = 18 animais temos $r(X,Y)$ = 0,81. Então:

$$z = \frac{1}{2} ln \left[\frac{1+0,81}{1-0,81}\right] = 1,1270, \qquad \mu_z = \frac{1}{2} ln \left[\frac{1+0,70}{1-0,70}\right] = 0,8673 \text{ e}$$

$$\sigma_z = \frac{1}{\sqrt{18-3}} = 0,2582$$

$$\Rightarrow z_{calc} = \frac{1,1270-(0,8673)}{0,2582} = 1,01 \notin RC(1\%)$$

Não rejeitamos H_0 e concluímos que podemos admitir que a correlação entre o ganho de peso e a quantidade de matéria seca ingerida por bovinos da raça Nelore <u>não é superior</u> a 0,70.

6.2. REGRESSÃO LINEAR SIMPLES

Desejamos estudar o comportamento conjunto de duas ou mais variáveis, como por exemplo:

i) O peso do animal com sua idade.

ii) A quantidade de adubo com a produção de matéria seca.

iii) A digestibilidade de um alimento com o tempo.

iv) O índice de inflação e os preços de diversos itens da cesta básica, etc.

Quando o interesse está em procurar expressar essa relação sob a forma de uma equação matemática estamos fazendo uma **Análise de Regressão**.

A equação de regressão pode ser um polinômio (reta, parábola ou um polinômio de grau mais elevado), uma função do tipo exponencial (curva logística, de Gompertz, von Bertalanfy *etc*.), uma função sigmoidal *etc*.

Estudaremos <u>somente o ajuste de uma reta</u> em problemas envolvendo duas variáveis:

Y: variável resposta ou variável dependente

X: variável regressora, covariada ou variável independente.

O Gráfico de dispersão facilita a <u>visualização</u> da relação funcional entre as variáveis. A distribuição dos pontos no gráfico pode sugerir qual função explica melhor o comportamento dos dados.

Exemplo 6.3 Determinar a reta que relaciona a Absorbância (Y) com a concentração de nitrito (X, em mg/100ml) em amostras de mortadela. Os dados experimentais são:

x_i	y_i
0.5	0.040
1.0	0.078
2.0	0.145
3.0	0.215
4.0	0.300
5.0	0.340
6.0	0.395
7.0	0.460
8.0	0.560
9.0	0.715

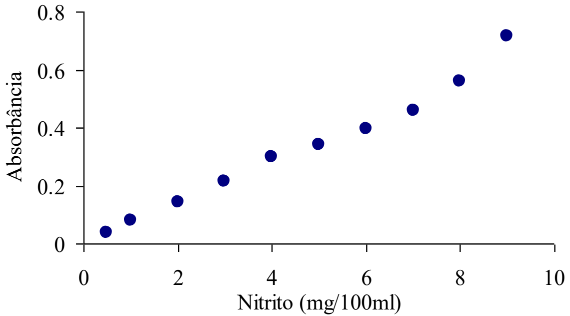


Figura 13. Gráfico de dispersão dos dados de absorbância e quantidade de nitrito.

- Percebe-se que a relação entre X e Y pode ser bem explicada por uma reta (função linear), cuja equação é $y = a + bx$.
- Teoricamente, esta reta deve passar pela origem, pois para uma solução sem nitrito espera-se uma absorbância nula.
- Inicialmente, precisamos estimar o valor do intercepto e do coeficiente angular da reta de regressão.

  Isso pode ser feito à mão livre, traçando-se uma reta que "passe pelo meio dos pontos" e pela origem (a = 0). O coeficiente angular pode ser calculado como b = $\Delta y/\Delta x$.

  Inconveniente desta solução: Observadores diferentes podem obter valores diferentes para os coeficientes a e b da reta.

  Indicado: Usar o Método dos Mínimos Quadrados Ordinários - MQO

6.2.1. O MODELO PARA REGRESSÃO LINEAR SIMPLES

A partir de uma amostra de n pares de valores (x_1, y_1),…,(x_n, y_n) nós podemos estabelecer o modelo de regressão linear simples:

$$y_i = a + bx_i + e_i$$

onde a e b são os parâmetros da reta e e_i é um erro associado aos valores de y_i.

O **erro** (e_i) pode resultar de erros de medidas e de digitação, da heterogeneidade das matérias primas utilizadas no experimento, da imprecisão dos equipamentos utilizados nas medições etc.

Para estabelecer o modelo de regressão linear estabelecemos alguns pressupostos que são indicados a seguir:

Pressuposições do modelo de regressão linear simples:

a) A relação entre as variáveis X e Y é linear.

b) Os valores da variável X não estão sujeitos a erros (são fixos).

c) A média dos erros é nula, isto é, $E(e_i) = 0$.

d) Para um dado valor x_i, a variância do erro é constante e igual a σ^2, isto é, $var(e_i) = \sigma^2$.

e) A correlação entre os erros de duas observações quaisquer é nula, isto é, $corr(e_i, e_j) = 0$, para $i \neq j$.

f) Os erros têm distribuição normal, isto é, $e_i \sim \mathrm{N}(0, \sigma^2)$.

O **Método dos Mínimos Quadrados Ordinários** (MQO) consiste em obter estimativas dos parâmetros a e b que minimizem a soma dos quadrados dos erros, ou seja, que minimizem a função:

$$SQE = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

Para obter o mínimo desta função (de duas variáveis) derivamos parcialmente SQE em relação aos parâmetros a e b:

$$\frac{\partial SQE}{\partial a} = \sum_{i=1}^{n}(y_i - a - bx_i)(-2)$$

$$\frac{\partial SQE}{\partial b} = \sum_{i=1}^{n}(y_i - a - bx_i)(-2x_i)$$

Para obter os estimadores de a e b, denotados por $\hat{a}$ e $\hat{b}$, igualamos as derivadas parciais a zero:

$$\sum_{i=1}^{n}\left(y_i - \hat{a} - \hat{b}x_i\right) = 0$$

$$\sum_{i=1}^{n}\left(y_i - \hat{a} - \hat{b}x_i\right)(x_i) = 0$$

Este sistema é chamado **Sistema de Equações Normais** (S.E.N.) e pode ser simplificado como:

$$\begin{cases} n\hat{a} + \hat{b}\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i \\ \hat{a}\sum_{i=1}^{n} x_i + \hat{b}\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} x_i y_i \end{cases}$$

Matricialmente:

$$\begin{bmatrix} n & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} x_i^2 \end{bmatrix} \begin{bmatrix} \hat{a} \\ \hat{b} \end{bmatrix} = \begin{bmatrix} \sum_{i=1}^{n} y_i \\ \sum_{i=1}^{n} x_i y_i \end{bmatrix}$$

Resolvendo o sistema, nós obtemos:

$$\hat{b} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2} = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

Não se preocupem!

A calculadora científica pode obter $\hat{a}$ e $\hat{b}$ facilmente!

O **modelo ajustado** pode ser escrito como:

$$\hat{y}_i = \hat{a} + \hat{b}x_i$$

O **resíduo** da regressão pode ser calculado como:

$$\hat{e}_i = y_i - \hat{y}_i$$

Quando a reta descreve bem a relação entre Y e X, esperamos que os valores observados (y_i) e os valores estimados pela reta $(\hat{y}_i)$ estejam próximos, o que produz resíduos $(\hat{e}_i)$ muito próximos de zero.

Podemos calcular os **resíduos padronizados** utilizando:

$$\hat{e}_i^* = \hat{e}_i \Big/ \sqrt{s^2_{y/x}} \tag{41}$$

onde $s^2_{y/x} = \frac{1}{(n-2)} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ é a variância da regressão.

Infelizmente a calculadora não fornece esta variância $(s^2_{y/x})$ facilmente…

Regra simples para identificar pontos atípicos:

Se $|\hat{e}_i^*| > 2 \Rightarrow$ o ponto (x_i, y_i) deve ser considerado um ponto atípico, discrepante ($outlier$) e merece ser estudado com atenção.

O gráfico de dispersão dos resíduos padronizados serve para:

i) Evidenciar a presença de pontos atípicos (resultantes de erros de medidas, de digitação etc.) que, após um estudo mais detalhado, poderão ser excluídos do conjunto de dados originais.

ii) Visualizar a possível fuga da suposição de variância constante.

Espera-se que os pontos do gráfico de dispersão dos resíduos estejam distribuídos aleatoriamente ao longo da reta $y = 0$, não mostrem nenhuma tendência nem variabilidades diferentes ao longo do eixo das abcissas.

A qualidade do ajuste de uma regressão pode ser avaliada através de gráficos de resíduos e do **coeficiente de determinação**:

$$R^2 = \left(\hat{b}\right)^2 \frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} = [r(X,Y)]^2 \qquad (41)$$

Como $0 \leq R^2 \leq 1$ o ajuste do modelo será tanto melhor quanto mais próximo de 1 estiver o valor de R^2.

O coeficiente de terminação (R^2) deve ser usado com cautela como medida de uma boa qualidade do ajuste.

As pressuposições do modelo também devem ser avaliadas após o ajuste do modelo. Por exemplo: esperamos que os pontos $(x_i; \hat{e}_i)$ ou $(x_i; \hat{e}_i^*)$ estejam distribuídos aleatoriamente em relação à reta y = 0, sem apresentar qualquer tendência ou aumento de variabilidade.

Exemplo: Se os resíduos apresentarem uma tendência quadrática, deveremos incluir no modelo um componente quadrático, do tipo cx_i^2 e estudar se esta inclusão foi importante.

Vamos ajustar a reta do Exemplo 6.3:

$n = 10$ $\sum x_i = 45{,}5$ $\sum x_i^2 = 285{,}25$ $\sum y_i = 3{,}248$

$\sum y_i^2 = 1{,}473$ $\sum x_i\, y_i = 20{,}438$

Com esses valores podemos estimar os parâmetros da reta:

$$\hat{b} = \frac{20{,}438 - (45{,}5)(3{,}248)/10}{285{,}25 - (45{,}5)^2/10} = \frac{5{,}6596}{78{,}2250} = 0{,}0724$$

$$\hat{a} = \frac{3{,}248}{10} - (0{,}0724)(4{,}55) = -0{,}0044$$

Reta ajustada: $\hat{y}_i = -0,0044 + 0,0724x_i$

Comentários:

- O intercepto (–0,0044) não tem um sentido prático (absorbância negativa para uma amostra sem nitrito ?!?!?)
- O coeficiente angular da reta (0,0724) pode ser entendido como o número de unidades que será acrescido a absorbância, quando a concentração de nitrito sofrer um acréscimo de 1mg/100ml.
- R^2=$(0,0724)^2(78,2250/0,4232) = 0,98$, indica que a relação entre a concentração de nitrito e a absorbância está muito bem explicada pela reta.

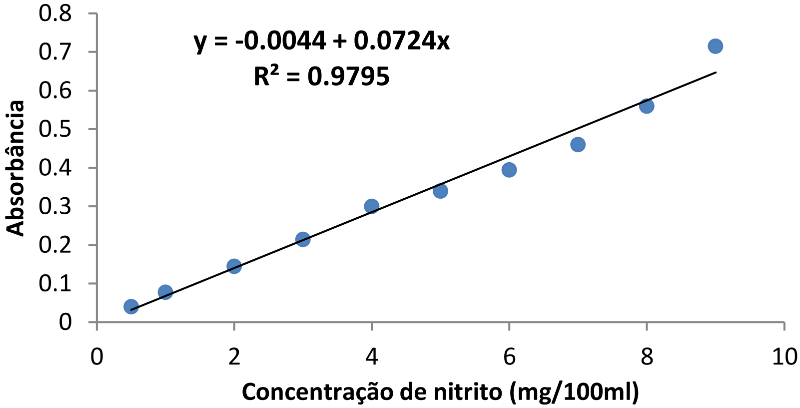


Figura. Reta de regressão ajustada aos dados do Exemplo 6.3

x_i	y_i	$\hat{y}_i$	$\hat{e}_i$	$\hat{e}_i^*$
0,5	0,040	0,032	0,008	0,241
1,0	0,078	0,068	0,010	0,301
2,0	0,145	0,140	0,005	0,151
3,0	0,215	0,213	0,002	0,060
4,0	0,300	0,285	0,015	0,452
5,0	0,340	0,357	-0,017	-0,512
6,0	0,395	0,430	-0,035	-1,054
7,0	0,460	0,502	-0,042	-1,265
8,0	0,560	0,574	-0,014	-0,422
9,0	0,715	0,647	0,068	**2,048**

Para construir o gráfico de dispersão (x_i; $\hat{e}_i^*$), calculamos os valores ajustados ($\hat{y}_i$) e os resíduos ordinários ($\hat{e}_i$).

Com esses resíduos calculamos a variância:

$$s^2_{y/x} = 0{,}0011$$

e os resíduos padronizados:

$$\hat{e}_i^* = \frac{\hat{e}_i}{\sqrt{0{,}0011}}$$

Com os resíduos padronizados podemos construir o seguinte gráfico de dispersão:

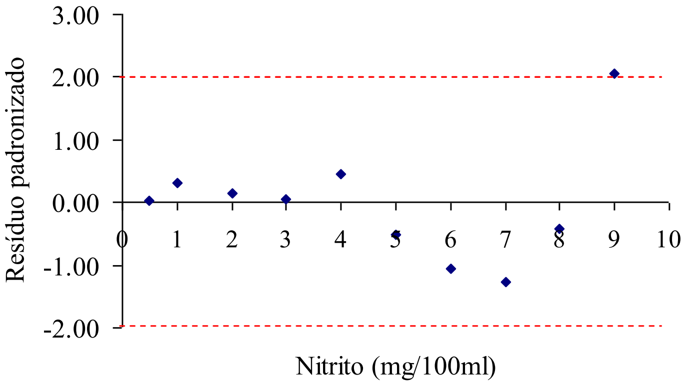


Figura 14. Gráfico de dispersão dos resíduos padronizados em função da concentração de nitrito.

Observe que:

- Os resíduos (Figura 14) não apresentam um comportamento aleatório (sequência de resíduos positivos e de resíduos negativos)
- O ponto (9,0; 0,715) tem um resíduo padronizado superior a 2 e merece atenção, podendo ser considerado um ponto atípico (!?)

Para melhorar a análise nós podemos tentar:

i) Excluir o ponto (9,0; 0,715) do conjunto de dados e ajustar uma nova reta aos dados (Fica como exercício!).

ii) Manter este ponto no conjunto e tentar ajustar outro modelo, como um polinômio de 2º grau. (Solução mais trabalhosa!).

CUIDADO: Devemos usar o R^2 com cautela!

Ajustar uma reta a cada conjunto de dados:

x_1	y_1	x_2	y_2	x_3	y_3	x_4	y_4
10	8.04	10	9.14	10	7.46	8	6.58
8	6.95	8	8.14	8	6.77	8	5.76
13	7.58	13	8.74	13	12.74	8	7.74
9	8.81	9	8.77	9	7.11	8	8.84
11	8.33	11	9.26	11	7.81	8	8.47
14	9.96	14	8.10	14	8.84	8	7.04
6	7.24	6	6.13	6	6.08	8	5.25
4	4.26	4	3.10	4	5.39	19	12.50
12	10.84	12	9.13	12	8.15	8	5.56
7	4.82	7	7.26	7	6.42	8	7.91
5	5.68	5	4.74	5	5.73	8	6.89

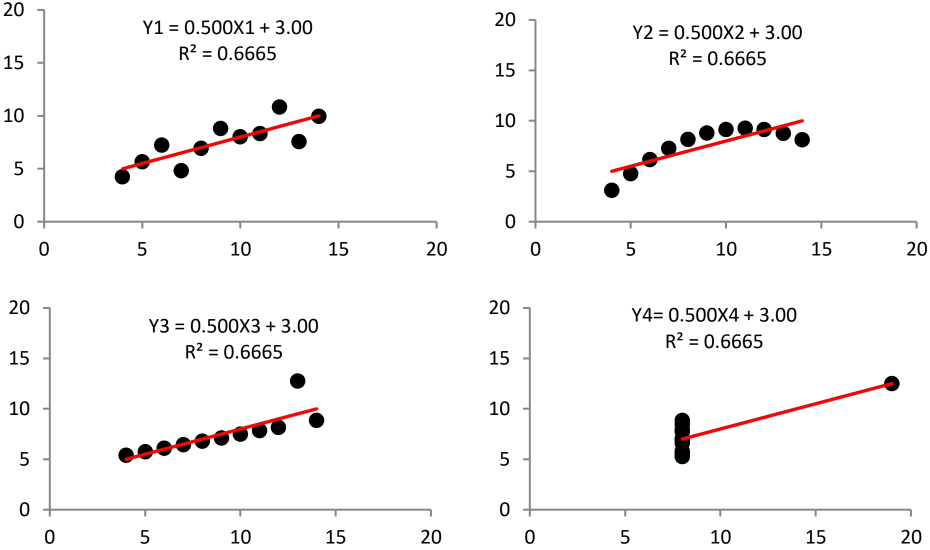
In: Chatterjee, S.; Price, B. **Regression analysis by Example**. Wiley, 1977, p. 8.

O que podemos comentar sobre os ajustes das retas?

Conjunto	Reta ajustada	R^2
1	$\hat{y}_1 = 3{,}00 + 0{,}500x_1$	0,67
2	$\hat{y}_2 = 3{,}00 + 0{,}500x_2$	0,67
3	$\hat{y}_3 = 3{,}00 + 0{,}500x_3$	0,67
4	$\hat{y}_4 = 3{,}00 + 0{,}500x_4$	0,67

Perceba que:

- Os coeficientes de determinação de todos os ajustes são iguais a 0,67, indicando uma boa qualidade do ajuste.
- Avaliando somente o valor do R^2 concluiríamos que uma reta se ajusta bem a todos os quatro conjuntos de dados. Será???


Y1 = 0.500X1 + 3.00
R² = 0.6665
Y2 = 0.500X2 + 3.00
R² = 0.6665
Y3 = 0.500X3 + 3.00
R² = 0.6665
Y4= 0.500X4 + 3.00
R² = 0.6665
0
5
10
15
20

INFERÊNCIA SOBRE OS PARÂMETROS DA RETA DE REGRESSÃO

Pode-se provar que $\boldsymbol{\hat{a}}$ e $\boldsymbol{\hat{b}}$ são estimadores não viesados de $\boldsymbol{a}$ e $\boldsymbol{b}$, respectivamente, ou seja, $\boldsymbol{E(\hat{a})} = \boldsymbol{a}$ e $\boldsymbol{E(\hat{b})} = \boldsymbol{b}$.

Os erros padrões associados a $\hat{a}$ e $\hat{b}$ podem ser calculados por:

$$ep(\hat{a}) = \sqrt{s^2_{y/x}\left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right)} \qquad ep(\hat{b}) = \sqrt{\frac{s^2_{y/x}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

onde $s^2_{y/x} = \frac{1}{(n-2)}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{(n-2)}\sum_{i=1}^{n}\hat{e}_i{}^2$.

Os ep's medem a variabilidade das estimativas dos parâmetros da reta e serão usados na construção de IC's e nos testes de hipóteses. No cálculo dos ep´s sempre aparece a variância $s^2_{y/x}$.

Intervalo de Confiança para o intercepto da reta:

$$IC(a, 100\gamma\%) = \widehat{a} \pm t_{tab} ep(\hat{a}) \tag{42}$$

onde t_{tab} é obtido da Tábua III, tal que $\gamma = P(-t_{tab} \leq T \leq t_{tab})$ e $T \sim t_{(n-2)}$.

Intervalo de Confiança para o coeficiente angular da reta:

$$IC(b, 100\gamma\%) = \hat{b} \pm t_{tab} ep(\hat{b}) \tag{43}$$

onde t_{tab} é obtido da Tábua III, tal que $\gamma = P(-t_{tab} \leq T \leq t_{tab})$ e $T \sim t_{(n-2)}$.

Intervalo de Predição para y_p, em que x_p pertence ao domínio da variável X, mas não foi usado na estimação dos parâmetros da reta.

$$IC(y_p, 100\gamma\%) = \hat{y}_p \pm t_{tab}\, ep(\hat{y}_p) \tag{44}$$

Onde $\hat{y}_p = \hat{a} + \hat{b}x_p$ e $ep(\hat{y}_p) = \sqrt{s^2_{y/x}\left(\frac{1}{n} + \frac{(x_p - \bar{x})^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)}$

Teste de hipótese para o intercepto da reta

- $H_0: a = a_0$ $vs.$ $H_a: a \neq a_0$ (ou $H_a: a < a_0$ ou $H_a: a > a_0$)
- Estatística do teste: $T = \dfrac{\hat{a} - a_0}{ep(\hat{a})} \sim t_{(n-2)}$.

Teste de hipótese para o coeficiente angular da reta

- $H_0: b = b_0 \;\; vs. \quad H_a: b \neq b_0$ (ou $H_a: b < b_0$ ou $H_a: b > b_0$)
- Estatística do teste: $T = \dfrac{\hat{b} - b_0}{ep(\hat{b})} \sim t_{(n-2)}$.

Aproveitando os dados do Exemplo 6.3 vamos calcular um IC para a inclinação da reta (γ = 95%), um IC para x_p = 9,5 mg/100ml e testar a hipótese de que o intercepto da reta é nulo, ao nível de significância α = 5%.

- $IC(b;\, 95\%) = 0{,}07235 \pm 2{,}306\sqrt{\dfrac{0{,}0011}{78{,}2250}} = 0{,}0724 \pm 0{,}0087$

  O intervalo $IC(b;\, 95\%)$ = [0,064; 0,081] contém o verdadeiro valor da inclinação da reta com uma confiança de 95%.

- O intervalo de predição para x_p = 9,5 é:

$$IC(y_p;\, 95\%) = 0{,}6832 \pm 2{,}306\sqrt{\left[\frac{1}{10} + \frac{(9{,}5 - 4{,}55)^2}{78{,}2250}\right] 0{,}0011}$$
$$= 0{,}6832 \pm 0{,}0492$$

$\Rightarrow IC(y_p;\, 95\%)$ = [0,634; 0,733] contém o verdadeiro valor previsto para a absorbância de uma amostra com 9,5mg/100 ml de nitrito, com 95% de confiança.

- Testar a hipótese que a reta passa pela origem dos eixos.

Hipóteses: H_0: $a = 0$ *versus* H_a: $a \neq 0$

Estatística do teste: T = $\frac{\hat{a}-a_0}{ep(\hat{a})} \sim t_{(8)}$

Para α = 5%, t_{tab} = 2,306 $\Rightarrow$ $RC = \{T \in \text{R}: |T| > 2{,}306\}$

Da amostra: $t_{calc} = \dfrac{-0{,}0044 - 0}{\sqrt{\left[\frac{1}{10} + \frac{4{,}55^2}{78{,}2250}\right]0{,}0011}} = \dfrac{-0{,}0044}{0{,}0200} = -0{,}22$

Como $t_{calc} \notin RC(5\%)$ não rejeitamos H_0 (α = 5%) e concluímos que o intercepto da reta pode ser considerado nulo (ou seja, a reta passa pela origem dos eixos).

Conclusão: Diante deste resultado, devemos ajustar uma nova reta que não tenha o intercepto, ou seja, uma reta $y_i = bx_i + \varepsilon_i$.

O estimador de Mínimos Quadrados do <u>novo</u> coeficiente angular da reta que passa pela origem dos eixos é dado por:

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} {x_i}^2}$$

Com os dados do Exemplo temos:

$$\hat{b} = \frac{20{,}438}{285{,}25} = 0{,}0716$$

A $nova$ reta ajustada pode ser escrita como:

$$\hat{y}_i = 0{,}0716x_i$$

Vale a pena lembrar que após o ajuste, é necessário fazer um novo estudo diagnóstico para verificar se as pressuposições do modelo estão satisfeitas.

Os softwares estatísticos (R, SAS, $Statistica\ etc$) têm ferramentas de uso simples que auxiliam nessa verificação das pressuposições.