Modelos Lineares I

Regressão Linear Simples (RLS):

(1ª, 2ª e 3ª Aulas)

Professor: Dr. José Rodrigo de Moraes

Universidade Federal Fluminense (UFF)

Departamento de Estatística (GET)

1

Regressão Linear

Conceito básico:

É uma ferramenta que utiliza a relação existente entre duas ou mais variáveis de modo que uma delas pode ser prevista (explicada) através da outra ou outras variáveis.

2

Relação entre duas variáveis:

- É um conceito primitivo que em alguns casos pode ser expresso por meio de relação funcional que pode ser:
  - Relação determinística (*matemática*)
  - Relação não determinística (*estatística*)
- Em geral, define-se uma relação entre duas variáveis X e Y, de forma que Y depende de X, por:

  Y = *f*(X), onde:

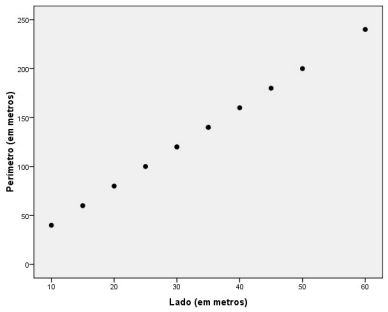
  X → variável independente (ou explicativa)

  Y → variável dependente (ou resposta)
- Conhecida a expressão analítica da função *f*, pode-se obter um valor de Y para um dado valor de X fixado.

3

Relação linear determinística:

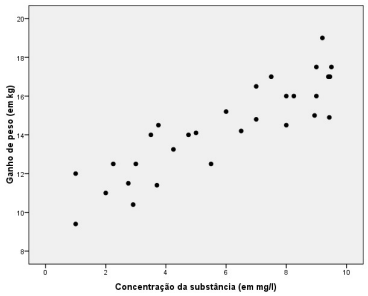
- **Gráfico de dispersão entre o lado (m) e o perímetro (m) de 10 quadrados.**



4

Relação linear estatística:

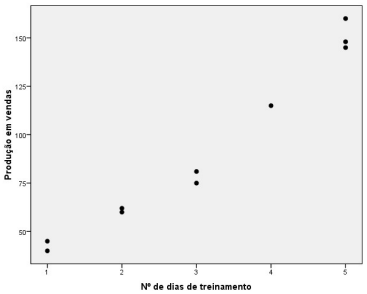
- **Gráfico de dispersão entre a concentração da substância (em mg/L) e o ganho de peso (em kg) de 30 bois (Magalhães & Lima, 2005).**



5

Que tipo de relação é essa ?

- **Gráfico de dispersão entre o número de dias de treinamento e a produção em vendas para 10 vendedores (Azevedo,2001).**



6

Modelo de Regressão Linear Simples (RLS)

- Modelo estatístico que considera uma variável explicativa (ou independente) X e uma variável resposta (ou dependente) Y.
- A construção teórica deste tipo de modelo exige o estabelecimento de um conjunto de hipóteses, as quais deverão ser testadas de modo que o modelo possa ser validado.
- A primeira hipótese consiste na identificação, por meio do **gráfico de dispersão**, de que os dados considerados (referentes às variáveis X e Y) apresentam uma relação do tipo linear (crescente ou decrescente).

7

Modelo de Regressão Linear Simples (RLS)

➢ **Características de uma relação estatística:**

- A dispersão dos pontos em torno de uma reta ou curva;
- Os valores da variável resposta tendem a variar com os valores da variável explicativa de forma sistemática.

8

Estabelecimento Formal do Modelo de RLS

(Hipóteses Básicas Gerais)

- Obtidos os pares de valores (X_i, Y_i) das variáveis X e Y para uma amostra de n elementos, **pode-se** obter um modelo linear, da forma: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
- Este modelo tem duas componentes: uma determinística e uma aleatória.
  - **Componente determinística** → representada por uma função de X, que indica a informação sobre Y obtida com base no conhecimento da variável X.
  - **Componente aleatória** → denominada "*erro aleatório*" que representa os vários outros fatores que podem interferir no comportamento da variável resposta Y.

9

Representação genérica do modelo de RLS

- Modelo estatístico: $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$
  - $Y_i \rightarrow$ valor observado da variável resposta Y referente ao i-ésimo elemento da amostra.
  - β_0 e $\beta_1 \rightarrow$ são os parâmetros desconhecidos a serem estimados.
  - $X_i \rightarrow$ valor observado da variável explicativa X referente ao i-ésimo elemento da amostra.
  - $\varepsilon_i \rightarrow$ erro aleatório do modelo, tal que:

    $E(\varepsilon_i)=0 \quad \forall \; i=1,2,...,n$

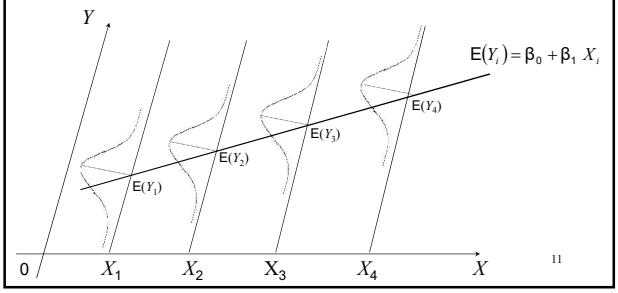
    $VAR(\varepsilon_i)=\sigma^2 \quad \forall \; i=1,2,...,n$

    $COV(\varepsilon_i,\varepsilon_j)= 0 \quad \forall \; i \neq j=1,2,..., n$

10

Observações:

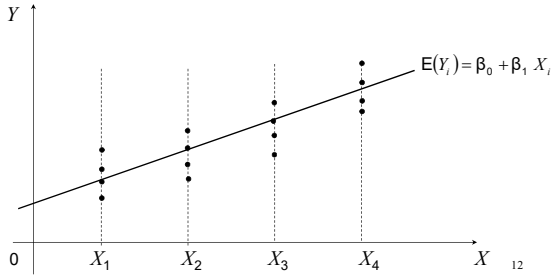
- A partir da expressão $E(Y_i)= \beta_0 + \beta_1 X_i$ é possível observar (Figuras) duas características contempladas no modelo de RLS:
  - **1ª característica:**



11

Observações:

- **2ª característica:**



12

Hipóteses decorrentes do modelo de RLS:

- O valor observado Y_i difere da verdadeira reta de regressão por uma quantidade igual a componente aleatória:
  $\varepsilon_i = Y_i - E(Y_i)$, $i=1,2, ..., n$.
- A variável resposta Y_i tem distribuição de probabilidade com variância constante:
  $VAR(Y_i)=\sigma^2, \forall\ i=1,2,..., n$.
- As variáveis respostas Y_i e Y_j são não-correlacionadas:
  $COV(Y_i, Y_j) = 0\ , \forall\ i \neq j = 1,2,..., n$.

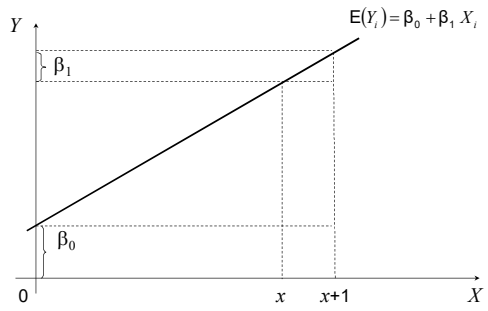
13

Parâmetros da reta de regressão:

- Os parâmetros β_0 e β_1 do modelo também são chamados de coeficientes de regressão:
  - $\beta_0 \rightarrow$ é o coeficiente linear da reta, isto é, a altura que em que a reta de regressão intercepta o eixo dos Y`s.
  - $\beta_1 \rightarrow$ é o coeficiente angular da reta, ou seja, é o acréscimo esperado na variável resposta Y quando a variável explicativa é acrescida de 1 unidade.
- A figura a seguir mostra a representação gráfica de um modelo de regressão linear simples (RLS):

14

Representação gráfica de um modelo de RLS:



15

Estimação dos parâmetros do modelo de RLS:

Em "modelos lineares" o método comumente usado para a estimação dos parâmetros β_0 e β_1 do modelo, é o **método dos mínimos quadrados (MQ).**

O método dos MQ consiste em encontrar os valores de β_0 e β_1 que minimizam a soma dos quadrados dos erros, isto é:

16

Estimação dos parâmetros do modelo de RLS:

$$S = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} [Y_i - E(Y_i)]^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

Derivando S em relação a β_0 e β_1:

$$\frac{\partial S}{\partial \beta_0} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)$$

$$\frac{\partial S}{\partial \beta_1} = -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) X_i$$

17

Estimação dos parâmetros do modelo de RLS:

As estimativas de mínimos quadrados (MQ) dos parâmetros β_0 e β_1 são os valores $\hat{\beta}_0$ e $\hat{\beta}_1$ que minimizam as equações abaixo:

$$\frac{\partial S}{\partial \beta_0} = 0 \ \rightarrow\ -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial S}{\partial \beta_1} = 0 \ \rightarrow\ -2 \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i) X_i = 0$$

18

Estimação dos parâmetros do modelo de RLS:

Desenvolvendo, temos que:

$$\frac{\partial S}{\partial \beta_0} = 0 \quad \rightarrow \quad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\frac{\partial S}{\partial \beta_1} = 0 \quad \rightarrow \quad \hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^{n} X_i^2 - n \bar{X}^2}$$

19

Estimação dos parâmetros do modelo de RLS

Fórmula ramificada

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^{n} X_i^2 - n \bar{X}^2}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n} (X_i - \bar{X})^2}$$

20

Estimadores de MQ de β_0 e β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n \bar{X} \bar{Y}}{\sum_{i=1}^{n} X_i^2 - n \bar{X}^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

Teorema de Gauss - Markov: Os estimadores de mínimos quadrados dos parâmetros do modelo β_0 e β_1 são não-tendenciosos e tem variância mínima dentre todos os estimadores lineares não tendenciosos.

21

Estimação da média da variável resposta (reta de regressão estimada):

Calculadas as estimativas de MQ dos parâmetros β_0 e β_1 estima-se a reta de regressão $E(Y_i)=\beta_0+\beta_1 X_i$ por:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \quad i = 1,2,...,n$$

onde:

$\hat{Y}_i \rightarrow$ Estimador pontual da média da variável resposta Y do modelo no nível X_i.

OBS: $\hat{Y}_i$ é um estimador não tendencioso da média da variável resposta $E(Y_i)$, tendo variância mínima dentre todos os estimadores lineares não tendenciosos.

22

Ajustamento do modelo de Regressão Linear Simples (RLS):

Observados n pares com os valores das variáveis X e Y de interesse, (X_1,Y_1), (X_2,Y_2), ..., (X_n,Y_n), é possível ajustar um modelo de regressão linear para representar a relação existente entre essas duas variáveis:

$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$, onde:

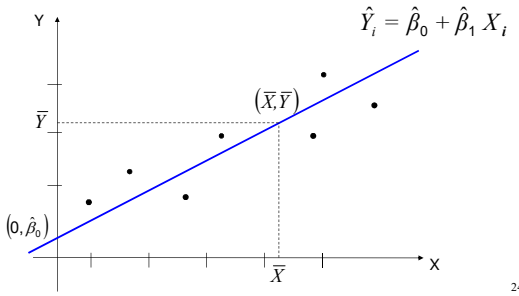
$\hat{Y}_i \rightarrow$ é o valor estimado (ou ajustado) da variável resposta Y do modelo referente ao i-ésimo elemento.

$X_i \rightarrow$ é o valor da variável explicativa X do modelo referente ao i-ésimo elemento.

$\hat{\beta}_0$ e $\hat{\beta}_1$ $\rightarrow$ são as estimativas de mínimos quadrados dos parâmetros do modelo.

23

Representação do modelo de RLS ajustado:



24

Interpretação das estimativas dos parâmetros do modelo ajustado $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ **:**

$\hat{\beta}_0 \rightarrow$ é o valor estimado para a variável resposta Y_i quando X_i é igual a zero (X_i=0).

$\hat{\beta}_1 \rightarrow$ é o quanto varia o valor estimado da variável resposta Y ao aumentar em 1 unidade o valor X_i da variável explicativa X.

Em outras palavras:

O valor estimado de Y_i varia em $\hat{\beta}_1$ unidades para cada unidade de acréscimo de X_i.

25

❑ **Exemplo – Modelo de Regressão de Linear Simples:**

Em uma dada região acredita-se que o gado alimentado em determinado pasto tem ganho de peso maior que o normal. Estudos de laboratório detectaram uma substância no pasto e deseja-se obter evidências de que tal substância pode ser utilizada para melhorar o ganho de peso dos bovinos. Foram selecionados n=30 bois de mesma raça e idade, e cada animal recebeu uma determinada concentração da substância X (em mg/L). O ganho de peso (Y) após 30 dias, foi medido e os dados estão apresentados na tabela abaixo (em kg):

26

Tabela: Dados sobre a concentração da substância X (em mg/L) e ganho de peso (em kg) após trinta dias, de n=30 bovinos:

Boi	Conc. Subst. (mg/l)	Ganho de peso (kg)	Boi	Conc. Subst. (mg/l)	Ganho de peso (kg)
1	1,00	9,40	16	5,00	14,10
2	3,70	11,40	17	5,50	12,50
3	1,00	12,00	18	6,00	15,20
4	9,00	16,00	19	6,50	14,20
5	2,00	11,00	20	7,00	16,50
6	2,25	12,50	21	7,50	17,00
7	2,91	10,40	22	8,00	14,50
8	2,75	11,50	23	8,25	16,00
9	3,00	12,50	24	9,40	17,00
10	3,50	14,00	25	9,43	14,90
11	3,75	14,50	26	8,94	15,00
12	9,45	17,00	27	9,20	19,00
13	4,25	13,25	28	9,50	17,50
14	7,00	14,80	29	8,00	16,00
15	4,75	14,00	30	9,00	17,50

27

Cálculos para a obtenção das estimativas dos parâmetros do modelo:

boi i	X_i	Y_i	X_i^2	Y_i^2	$X_i Y_i$
1	1,00	9,40	1,0000	88,3600	9,4000
2	3,70	11,40	13,6900	129,9600	42,1800
3	1,00	12,00	1,0000	144,0000	12,0000
4	9,00	16,00	81,0000	256,0000	144,0000
5	2,00	11,00	4,0000	121,0000	22,0000
6	2,25	12,50	5,0625	156,2500	28,1250
7	2,91	10,40	8,4681	108,1600	30,2640
8	2,75	11,50	7,5625	132,2500	31,6250
9	3,00	12,50	9,0000	156,2500	37,5000
10	3,50	14,00	12,2500	196,0000	49,0000
11	3,75	14,50	14,0625	210,2500	54,3750
12	9,45	17,00	89,3025	289,0000	160,6500
13	4,25	13,25	18,0625	175,5625	56,3125
14	7,00	14,80	49,0000	219,0400	103,6000
15	4,75	14,00	22,5625	196,0000	66,5000

28

Continuação – Obtenção das estimativas dos parâmetros

boi i	X_i	Y_i	X_i^2	Y_i^2	$X_i Y_i$
16	5,00	14,10	25,0000	198,8100	70,5000
17	5,50	12,50	30,2500	156,2500	68,7500
18	6,00	15,20	36,0000	231,0400	91,2000
19	6,50	14,20	42,2500	201,6400	92,3000
20	7,00	16,50	49,0000	272,2500	115,5000
21	7,50	17,00	56,2500	289,0000	127,5000
22	8,00	14,50	64,0000	210,2500	116,0000
23	8,25	16,00	68,0625	256,0000	132,0000
24	9,40	17,00	88,3600	289,0000	159,8000
25	9,43	14,90	88,9249	222,0100	140,5070
26	8,94	15,00	79,9236	225,0000	134,1000
27	9,20	19,00	84,6400	361,0000	174,8000
28	9,50	17,50	90,2500	306,2500	166,2500
29	8,00	16,00	64,0000	256,0000	128,0000
30	9,00	17,50	81,0000	306,2500	157,5000
Total	**177,53**	**431,15**	**1.283,9341**	**6.358,8325**	**2.722,2385**

29

Cálculo das estimativas dos parâmetros do modelo β_0 e β_1:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\,\overline{X}\,\overline{Y}}{\sum_{i=1}^{n} X_i^2 - n\,\overline{X}^2} =$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} =$$

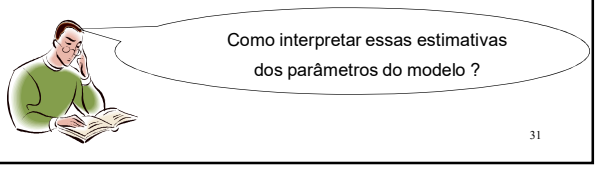
30

Modelo (ou Reta) de regressão ajustado (a)

❑ Modelo ajustado:

$$\hat{Y}_i = 10{,}040 + 0{,}732\ X_i,\ \ i = 1,2,...,30$$

- $\hat{Y}_i \rightarrow$ ganho de peso (em kg) estimado do i-ésimo boi.
- $X_i \rightarrow$ concentração de substância (em mg/L) ingerida pelo i-ésimo boi.



31

Exemplo: Dados sobre a concentração da substância X (mg/L) e ganho de peso Y (kg) de n=30 bois:

Resultados do Ajuste (n=30 bois) usando o SPSS: ***Analyse /Regression / Linear***

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	10,040	,495		20,277	,000
	X_conc.subs	,732	,076	,877	9,672	,000

a. Dependent Variable: Y_ganho_peso

Estimativas dos parâmetros por MQO

32

Resíduos do modelo de RLS:

O resíduo do modelo e_i referente ao i-ésimo elemento da amostra, i=1,2,...,n, mede a distância (ou a discrepância) entre o valor observado Y_i e o correspondente valor estimado $\hat{Y}_i$, isto é:

$$e_i = Y_i - \hat{Y}_i, \quad \forall\ i = 1,2,...,n$$

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

33

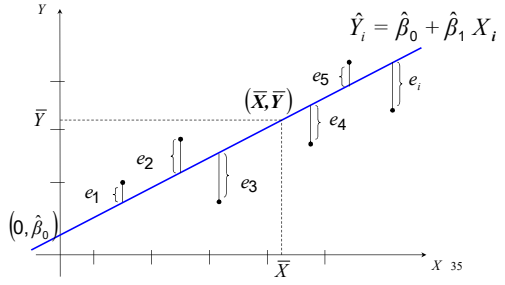
Resíduos do modelo de RLS:

- Valores bem (mal) ajustados devem apresentar pequenos (grandes) resíduos.
- O sinal do resíduo indica se o valor observado (Y_i) é menor (resíduo negativo: e_i<0) ou maior (resíduo positivo: e_i>0) que o valor estimado ($\hat{Y}_i$).
- No caso de resíduo nulo (e_i=0), tem-se que o valor observado (Y_i) é exatamente igual ao valor estimado ($\hat{Y}_i$).

34

Modelo de RLS Ajustado:

Para cada elemento i, devemos diferenciar o valor observado Y_i, i=1,2,...,n do valor estimado $\hat{Y}_i$ (ou valor ajustado) obtido com base no modelo de regressão linear ajustado:



35

Propriedades da reta de regressão:

1. A soma dos resíduos é igual a zero:

$$\sum_{i=1}^{n} e_i = 0$$

2. A soma dos valores observados é igual a soma dos valores estimados de Y:

$$\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n} \hat{Y}_i$$

36

Propriedades da reta de regressão:

3. A soma dos resíduos ponderados pelos valores da variável explicativa X é zero:

$$\sum_{i=1}^{n} X_i \, e_i = 0$$

4. A soma dos resíduos ponderados pelos valores estimados da variável resposta Y é zero:

$$\sum_{i=1}^{n} \hat{Y}_i \, e_i = 0$$

37

Propriedades da reta de regressão:

5. A reta de regressão ajustada passa pelo ponto $(\overline{X},\overline{Y})$.

$$\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \, \overline{X}$$

6. A soma dos quadrados dos resíduos é mínima segundo o método de mínimos quadrados.

38

Exemplo: Dados sobre a concentração da substância X (mg/L) e ganho de peso Y (kg) de n=30 bois:

Cálculo dos resíduos do modelo:

boi i	X_i	Y_i	$\hat{Y}_i = 10{,}040 + 0{,}732\,X_i$	$e_i = Y_i - \hat{Y}_i$
1	1,00	9,40	10,77	-1,37
2	3,70	11,40	12,75	-1,35
3	1,00	12,00	10,77	1,23
4	9,00	16,00	16,63	-0,63
5	2,00	11,00	11,50	-0,50
6	2,25	12,50	11,69	0,81
7	2,91	10,40	12,17	-1,77
8	2,75	11,50	12,05	-0,55
9	3,00	12,50	12,24	0,26
10	3,50	14,00	12,60	1,40
11	3,75	14,50	12,79	1,72
12	9,45	17,00	16,96	0,04
13	4,25	13,25	13,15	0,10
14	7,00	14,80	15,16	-0,36
15	4,75	14,00	13,52	0,48

39

Cálculo dos resíduos do modelo (Continuação):

boi i	X_i	Y_i	$\hat{Y}_i = 10{,}040 + 0{,}732\,X_i$	$e_i = Y_i - \hat{Y}_i$
16	5,00	14,10	13,70	0,40
17	5,50	12,50	14,07	-1,57
18	6,00	15,20	14,43	0,77
19	6,50	14,20	14,80	-0,60
20	7,00	16,50	15,16	1,34
21	7,50	17,00	15,53	1,47
22	8,00	14,50	15,90	-1,40
23	8,25	16,00	16,08	-0,08
24	9,40	17,00	16,92	0,08
25	9,43	14,90	16,94	-2,04
26	8,94	15,00	16,58	-1,58
27	9,20	19,00	16,77	2,23
28	9,50	17,50	16,99	0,51
29	8,00	16,00	15,90	0,10
30	9,00	17,50	16,63	0,87
Total	**177,53**	**431,15**	**431,15**	**0**

40