Modelo Regressivo Linear Simples

Parte 2

Análise Regressão Linear

Maria Teresa Leão Costa

Formulação do Modelo

Problema 1

- A ***taxa de metabolismo***, *taxa de consumo de energia do corpo*, é importante em estudos sobre aumento de peso, dieta e exercício.
- Em um estudo foram coletados dados sobre a ***massa do corpo sem gordura*** e a ***taxa metabólica em repouso***, para 19 indivíduos selecionados aleatoriamente entre os submetidos a um estudo de dieta.
  - A massa do corpo sem gordura, dada em quilogramas, é o peso da pessoa, eliminada toda a gordura.
  - A taxa de metabolismo é medida em calorias queimadas a cada 24 horas, as mesmas calorias usadas para descrever o conteúdo energético dos alimentos.
- Os pesquisadores acham que a massa do corpo sem gordura tem grande influência sobre a taxa de metabolismo.

Indivíduo	massa	Taxa
1	62,0	1792
2	62,9	1666
3	36,1	995
4	54,6	1425
5	48,5	1396
6	42,0	1418
7	47,4	1362
8	50,6	1502
9	42,0	1256
10	48,7	1614
11	40,3	1189
12	33,1	913
13	51,9	1460
14	42,4	1124
15	34,5	1052
16	51,1	1347
17	41,2	1204
18	51,9	1867
19	46,9	1439

- Se considerarmos apenas a taxa de metabolismo em repouso qual seria a avaliação de vocês sobre a taxa média de metabolismo em repouso ?

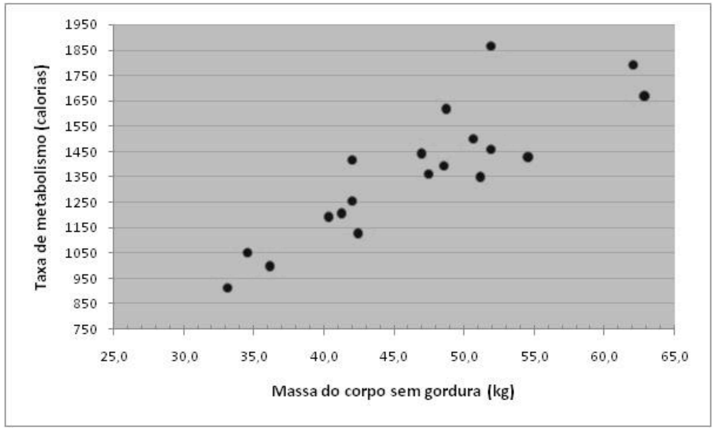
$\bar{y} = 1369,524$ *calorias*

Taxa
1792
1666
995
1425
1396
1418
1362
1502
1256
1614
1189
913
1460
1124
1052
1347
1204
1867
1439

Entretanto ...

se considerarmos
a massa do corpo sem gordura ....

UnB - IE Departamento de Estatística



O que vocês podem constatar através do gráfico?

UnB - IE Departamento de Estatística

- Existe relação entre a taxa de metabolismo e a massa do corpo sem gordura ?
- Que medida poderíamos usar para confirmar esta hipótese?
- Porque no caso em que os indivíduos pesquisados apresentam a mesma massa do corpo sem gordura de 42 kg, por exemplo, a taxa de metabolismo não foi a mesma?
- Como vocês explicariam esta variação na taxa de metabolismo ?

UnB - IE Departamento de Estatística

- Medida do grau de relacionamento LINEAR entre duas variáveis quantitativas:

$$\rho = \rho(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sqrt{V(X)}\sqrt{V(Y)}} \longrightarrow$$ ***coeficiente de correlação populacional***

- Em uma amostra de tamanho *n*, (X_i, Y_i) $i = 1, 2, \ldots, n$ temos:

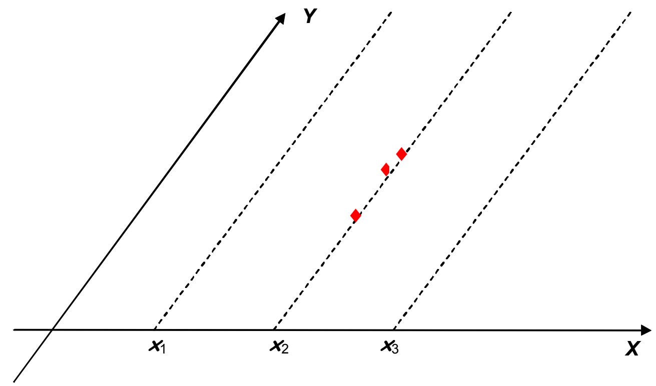
$$r = \frac{\sum_{i=1}^{n} X_i Y_i - n\bar{X}\bar{Y}}{\sqrt{\sum_{i=1}^{n} X_i^2 - n\bar{X}^2}\sqrt{\sum_{i=1}^{n} Y_i^2 - n\bar{Y}^2}} \rightarrow$$ *coeficiente de correlação da amostra*

Teste para $\rho = 0$

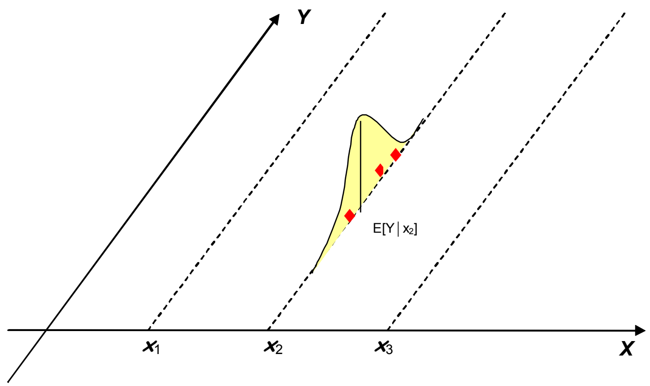
$$H_0)\ \rho = 0$$
$$H_1)\ \rho \neq 0 \quad (ou\ \ \rho < 0\ \ ou\ \ \rho > 0)$$

```
                          The CORR Procedure

                   2  Variables:    massa_    Taxa


                          Simple Statistics

Variable          N         Mean      Std Dev          Sum      Minimum      Maximum    Label

massa_           19     46.74211      8.28441    888.10000     33.10000     62.90000    massa
Taxa             19         1370    257.50412        26021    913.00000         1867    Taxa


                Pearson Correlation Coefficients, N = 19
                       Prob > |r| under H0: Rho=0

                              massa_          Taxa

                 massa_      1.00000       0.86474
                 massa                      <.0001

                 Taxa        0.86474       1.00000
                 Taxa         <.0001
```

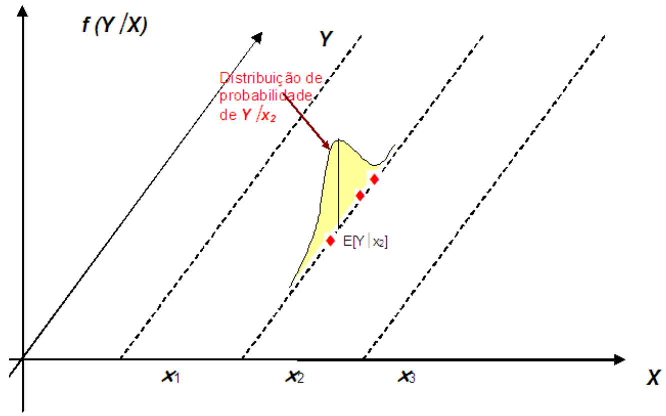

UnB - IE
Departamento de
Estatística

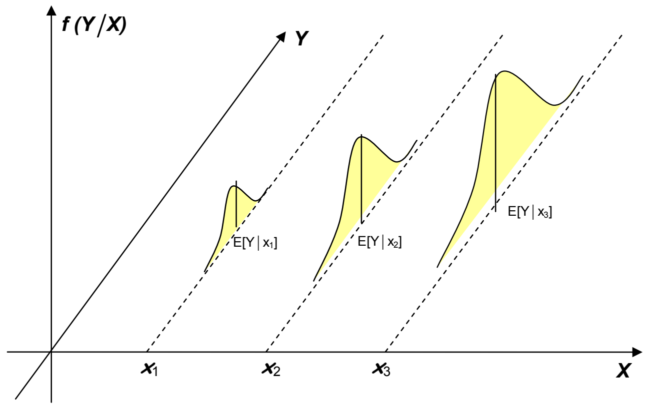


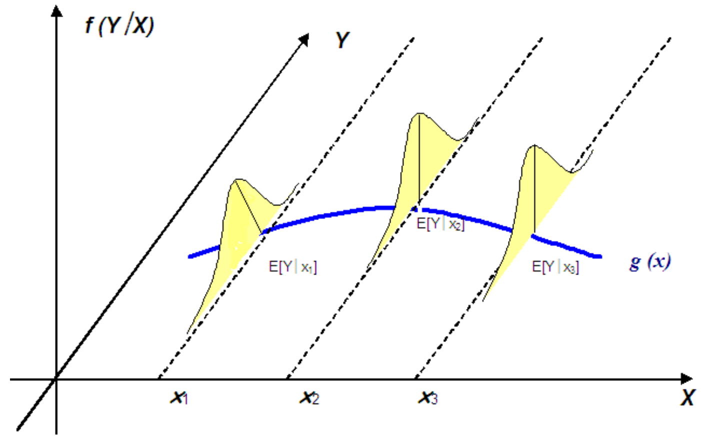
UnB - IE
Departamento de
Estatística

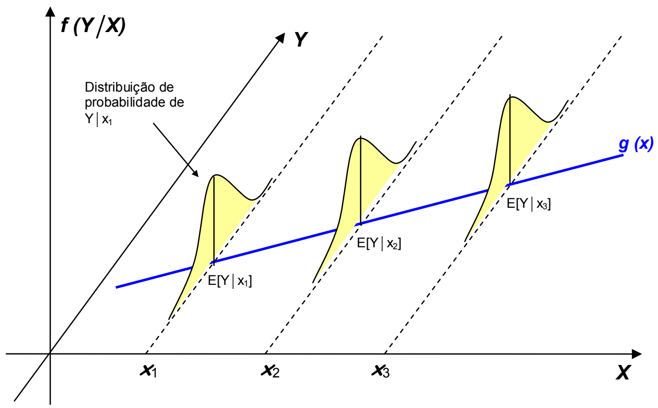


UnB - IE
Departamento de
Estatística









CONCEITOS BÁSICOS

Um modelo regressivo é um meio formal de expressar os dois ingredientes essenciais de uma relação estatística:

- *A tendência da variável resposta para variar com a variável ou variáveis explicativas de uma forma sistemática.*
- *A dispersão das observações em torno da curva da relação estatística.*

Estas duas características são incluídas no modelo regressivo postulando que:

- *Existe uma distribuição de probabilidade de para cada nível de X .*
- *A média destas distribuições de probabilidade variam de algum modo sistemático com X .*

Modelo de Regressão Linear Simples

◈ Relação entre variáveis é uma função linear

Erro Aleatório — v.a.

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Variável resposta — v.a.

Variável explicativa ⬇ **fixado**

- O modelo é dito ser:
  - ☐ **Simples** – *existe apenas uma variável independente*
  - ☐ **Linear nos parâmetros** – *nenhum parâmetro aparece como expoente ou multiplicado ou dividido por outro parâmetro*
  - ☐ **Linear na variável independente** – a variável independente aparece apenas na primeira potência.

Suposições do Modelo de Regressão Linear

Os erros aleatórios com:

- $E(\varepsilon_i) = 0$
- $\operatorname{cov}\left(\varepsilon_i, \varepsilon_j\right) = \begin{cases} 0 & se\ i \neq j \\ \sigma^2 = V\left(\varepsilon_i\right) & se\ i = j \end{cases}$
- Linearidade

Erros não correlacionados

HOMOCEDASTICIDADE (Variância constante)

Suposições do Modelo de Regressão Linear

- Normalidade
  - ☐ Distribuição de probabilidade dos erros é normal com

$$E(\varepsilon_i) = 0 \quad e \quad V(\varepsilon_i) = \sigma^2$$

OBS: Normalidade e $\operatorname{cov}\left(\varepsilon_i, \varepsilon_j\right) = 0 \Rightarrow$ **Independência dos ERROS**

⬇

Os *valores de Y* são ***normalmente*** distribuídos para cada valor x de X com média $E(Y_i) = \beta_0 + \beta_1\, x_i$
e variância $V(Y_i) = \sigma^2$.
As respostas $Y_i \;\; e \;\; Y_j$ são não correlacionadas.

Estimação dos Parâmetros do Modelo

População **Amostra Aleatória**

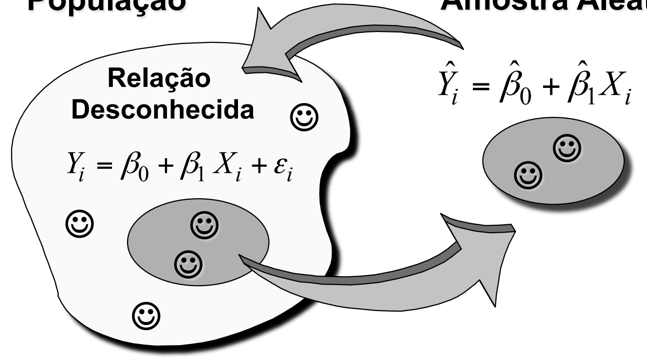
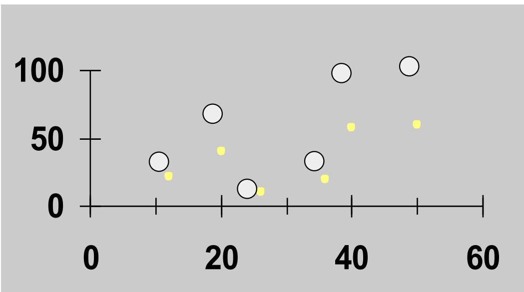


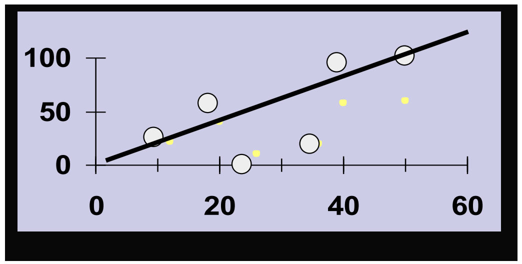
Diagrama de Dispersão

- "Plotar" todos os pares (X_i, Y_i)



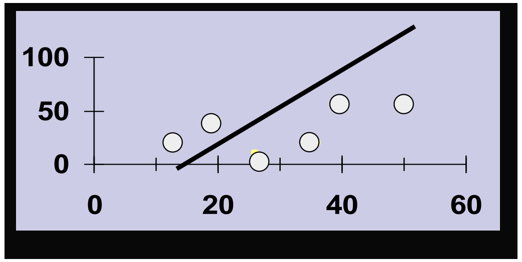
Como se pode desenhar uma reta passando por estes pontos?

Como determinar qual reta se "*ajusta melhor*"?



Como se pode desenhar uma reta passando por estes pontos?

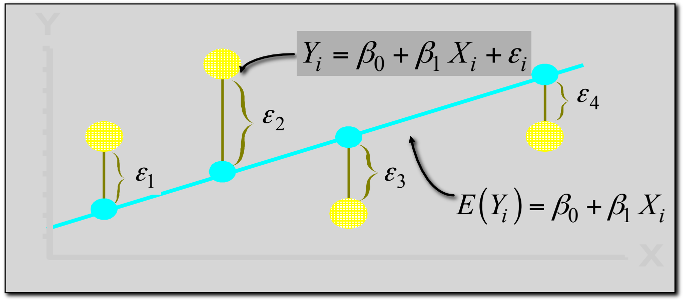
Como determinar qual reta se "*ajusta melhor*"?



UnB - IE
Departamento de Estatística

O Método dos Mínimos Quadrados

"*Melhor se ajusta*" significa diferença entre o valor real de Y e os valores preditos de Y é *mínima*.



$$\varepsilon_i = Y_i - E(Y_i)$$

UnB - IE
Departamento de Estatística

O Método dos Mínimos Quadrados

- Considerando uma amostra de tamanho ***n*** de pares temos: **(X_i, Y_i)**

Observe que:

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(Y_i - E(Y_i)\right)^2$$

e os estimadores $\hat{\beta}_0$ e $\hat{\beta}_1$ de β_0 e β_1, respectivamente, são aqueles que minimizam a soma de quadrados do erro

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 = SQ(\beta_0, \beta_1)$$

Veja Exercício 1

UnB - IE
Departamento de Estatística

Estimadores dos Parâmetros do Modelo

Coeficiente angular ou declividade

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

Intercepto

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