Introdução aos Modelos Lineares em Ecologia

Prof. Adriano Sanches Melo - Dep. Ecologia – UFG
asm.adrimelo no gmail.com

Página do curso: www.ecologia.ufrgs.br/~adrimelo/lm/

Livro-texto: Crawley, M.J. 2005. Statistics: An Introduction using R. John Wiley & Sons.

Página do livro na internet:
http://www3.imperial.ac.uk/naturalsciences/research/statisticsusingr

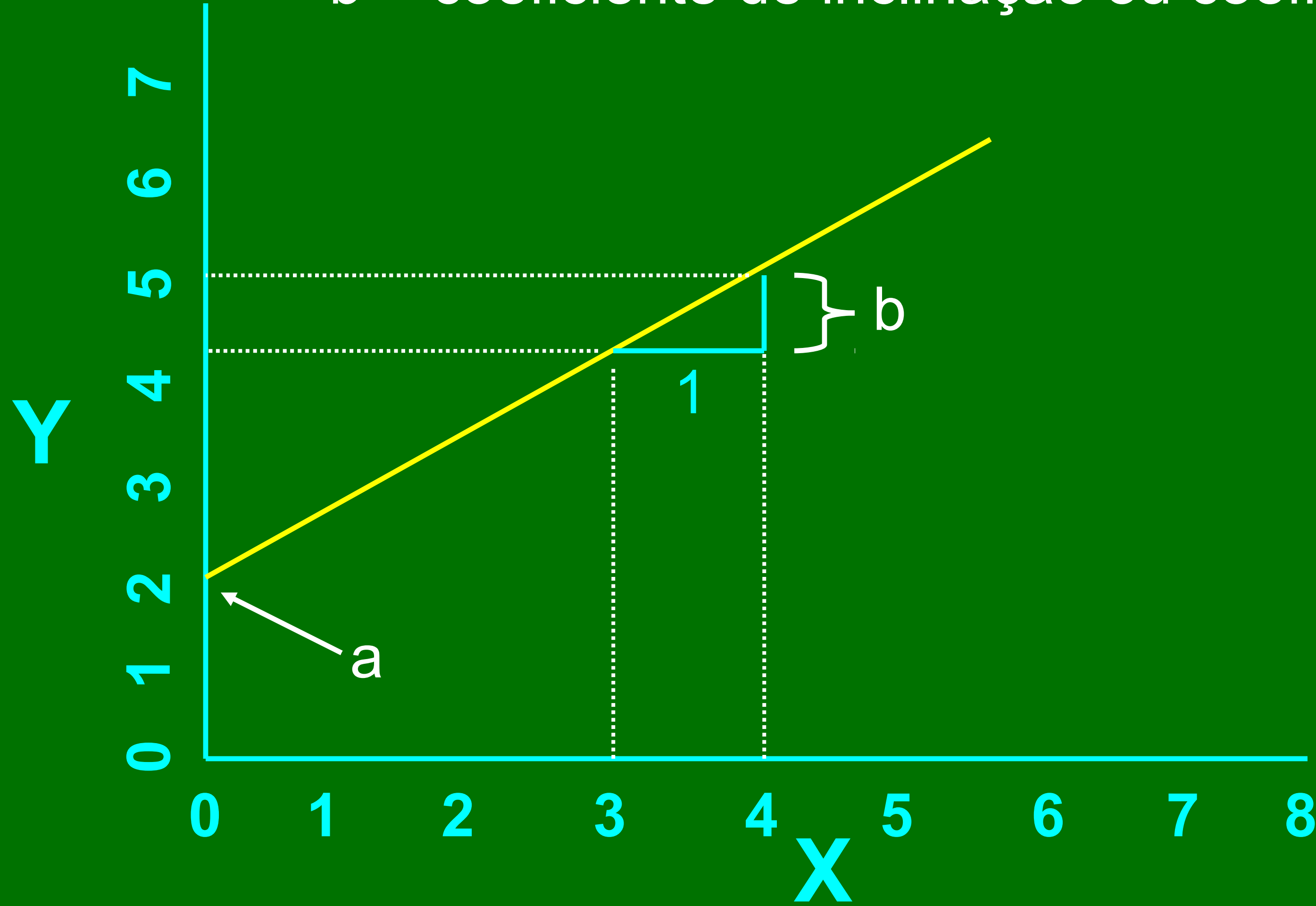
AULA 3

A equação da reta

$$Y = a + b*X$$

a = intercepto

b = coeficiente de inclinação ou coeficiente angular



Exemplos:

Valor pago (Y) = Tarifa*Número de unidades de produto (X)

Valor pago (Y) = Taxa de entrada + Tarifa* Distância em km (X)

O modelo de regressão

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

onde:

Y_i é o valor da variável resposta na observação i;

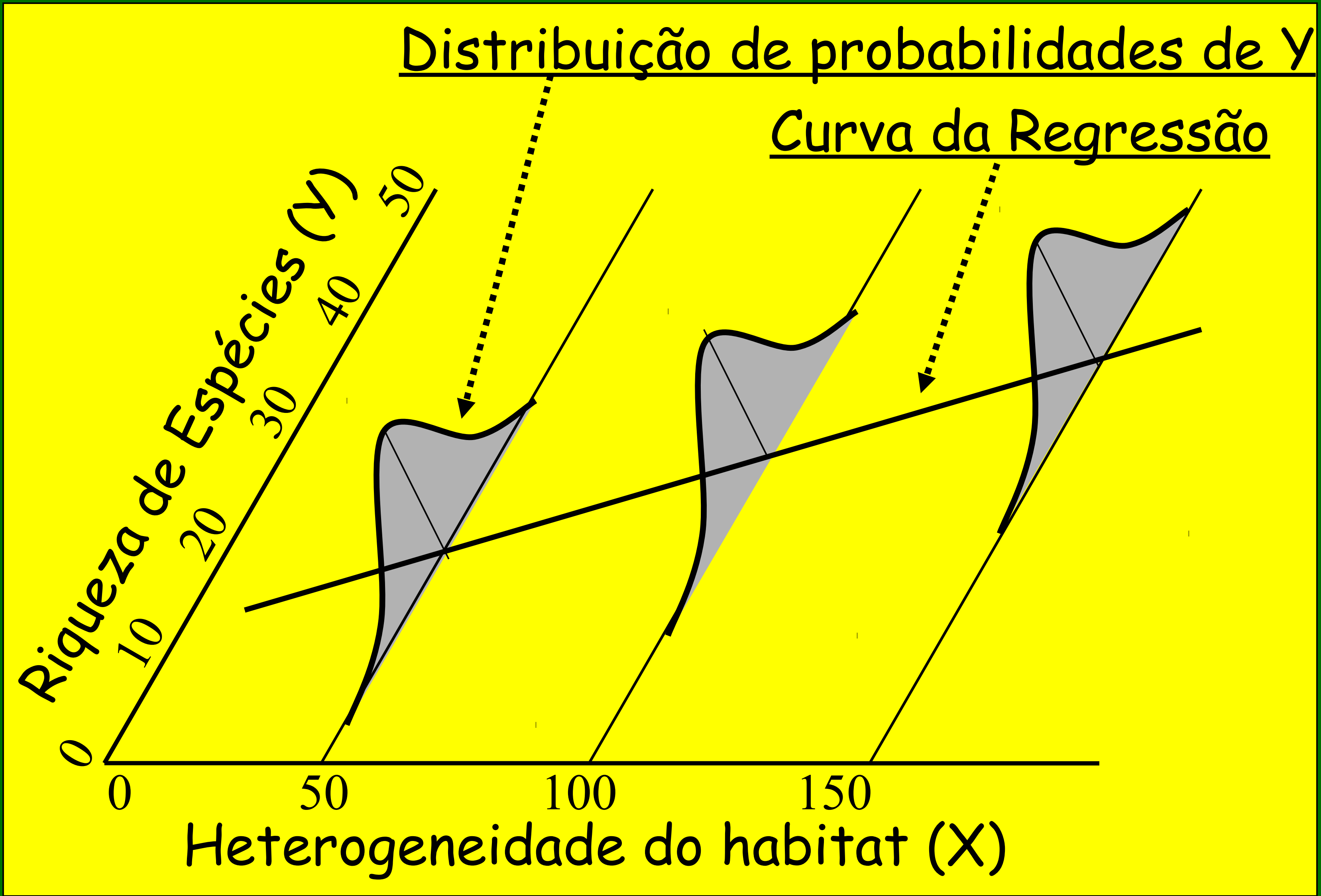
β_0 and β_1 são parâmetros;

X_i é um valor conhecido da variável explanatória;

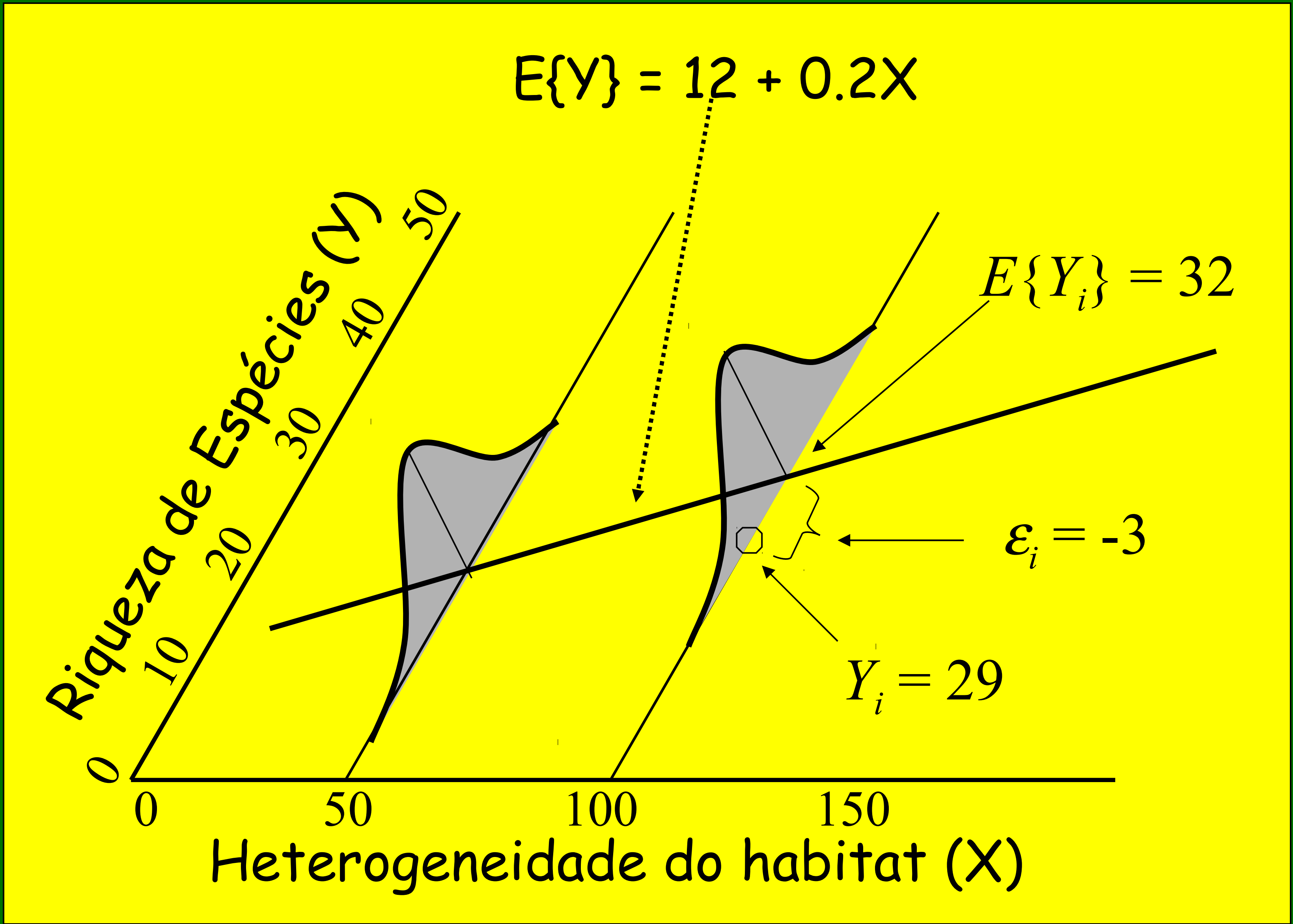
ε_i termo relacionado ao erro, e que possui média $E\{\varepsilon_i\}=0$ e variância $\sigma^2\{\varepsilon_i\} = \sigma^2$;

ε_i e ε_j não são correlacionados fazendo com que sua covariância seja zero (i.e. $\sigma\{\varepsilon_i, \varepsilon_j\}=0$ para todos i, j; $i \neq j$).

$i = 1, 2, ..., n$.



Distribuição de probabilidades de Y
Curva da Regressão
Riqueza de Espécies (Y)
0
10
20
30
40
50
0
50
100
150
Heterogeneidade do habitat (X)



E{Y} = 12 + 0.2X
Riqueza de Espécies (Y)
0
10
20
30
40
50
$E\{Y_i\} = 32$
$\varepsilon_i = -3$
$Y_i = 29$
0
50
100
150
Heterogeneidade do habitat (X)

--O modelo é *simples*, pois possui apenas 1 variável explanatória.

--O modelo é *linear nos parâmetros*, pois nenhum parâmetro aparece como expoente, nem é multiplicado ou dividido por outro parâmetro.

--O modelo é *linear na variável explanatória*, pois esta variável aparece apenas na primeira potência.

O critério de Quadrados Mínimos (Least Squares)

Definição de SS $= \Sigma (Y_i - [\beta_0 + \beta_1 X_i])^2$

Exemplo com <u>média</u> de uma população (aula anterior)

Exemplo "regressão" com apenas a constante (= média; idem exemplo acima)

Exemplo com regressão com apenas um coeficiente de regressão

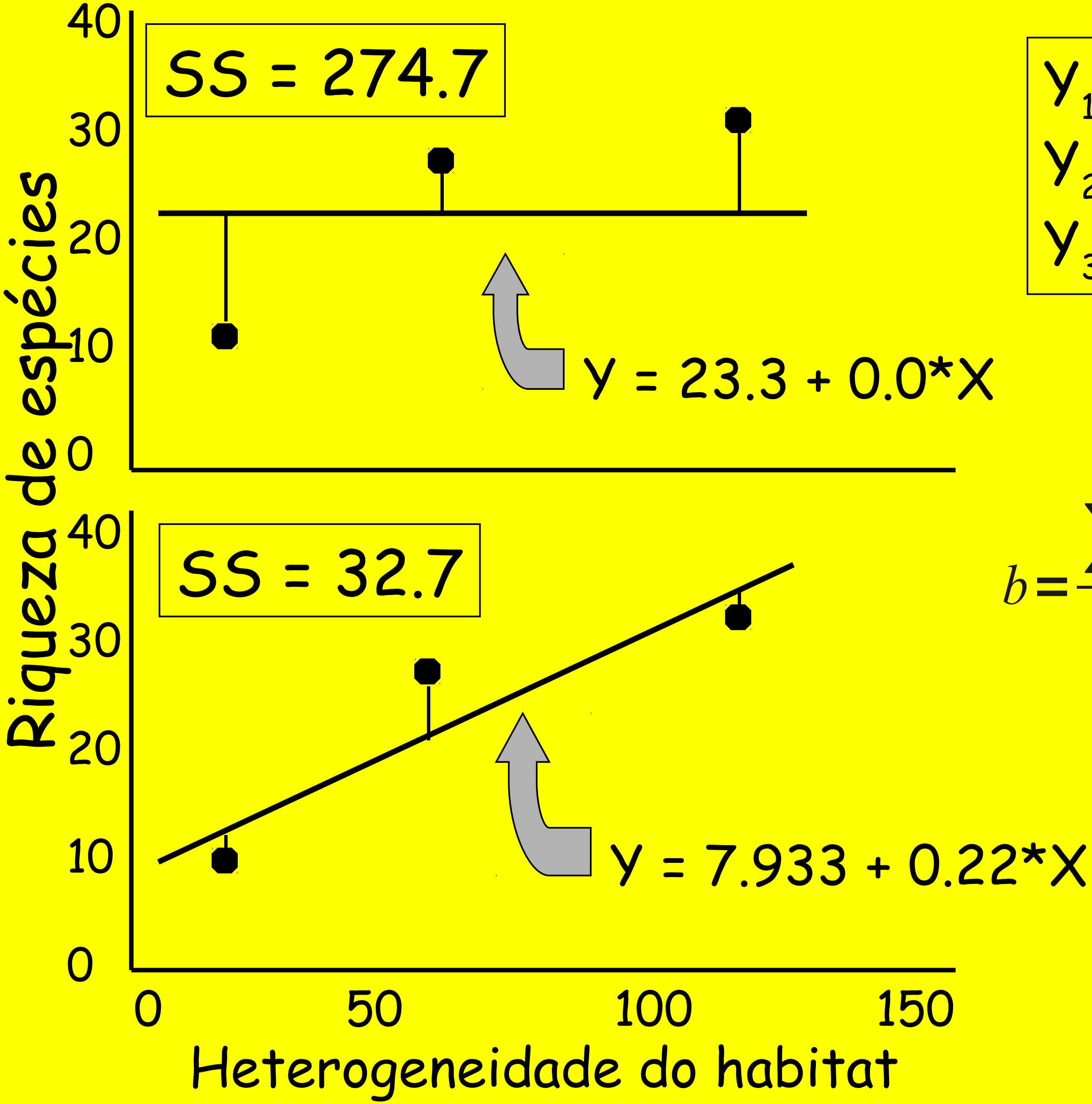
Como encontrar a e b ?

1. Por iteração (tentativa e erro)

2. De forma analítica

$$b = \frac{\sum \left(X_i - \bar{X} \right)\left(Y_i - \bar{Y} \right)}{\sum \left(X_i - \bar{X} \right)^2}$$

$$a = \frac{1}{n}\left(\sum Y_i - b \sum X_i \right) = \bar{Y} - b\,\bar{X}$$



$Y_1=10$	$X_1=20$
$Y_2=28$	$X_2=70$
$Y_3=32$	$X_3=120$

$$b = \frac{\sum \left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sum \left(X_i - \bar{X}\right)^2}$$

$$a = \bar{Y} - b\,\bar{X}$$

SS Total (SST) = SS Regressão (SSR) + SS Resíduo (SSE)

$$Y_i - \bar{Y} = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

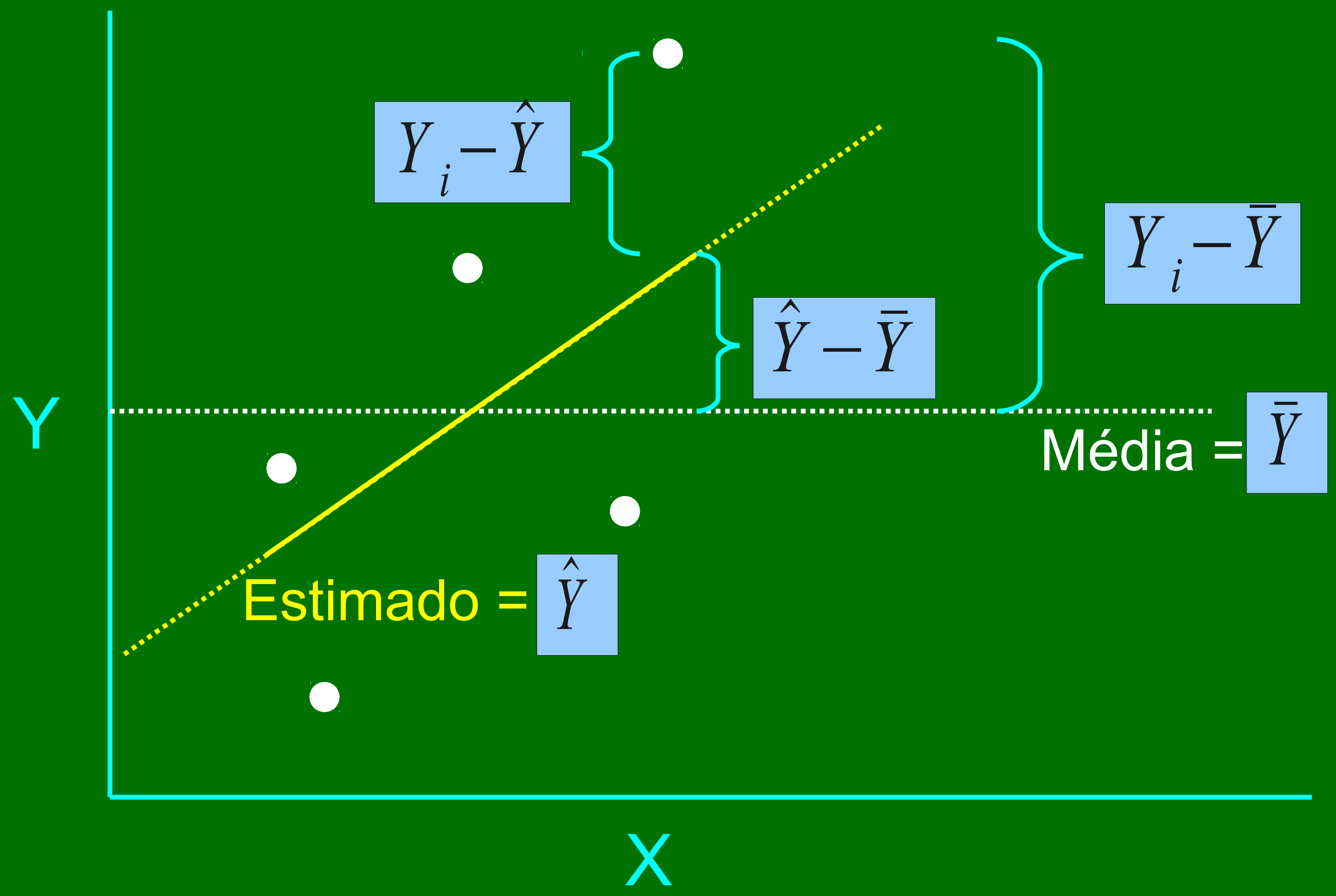
$$SST = \sum \left(Y_i - \bar{Y}\right)^2$$

$$SSR = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

$$SSE = \sum \left(Y_i - \hat{Y}\right)^2$$

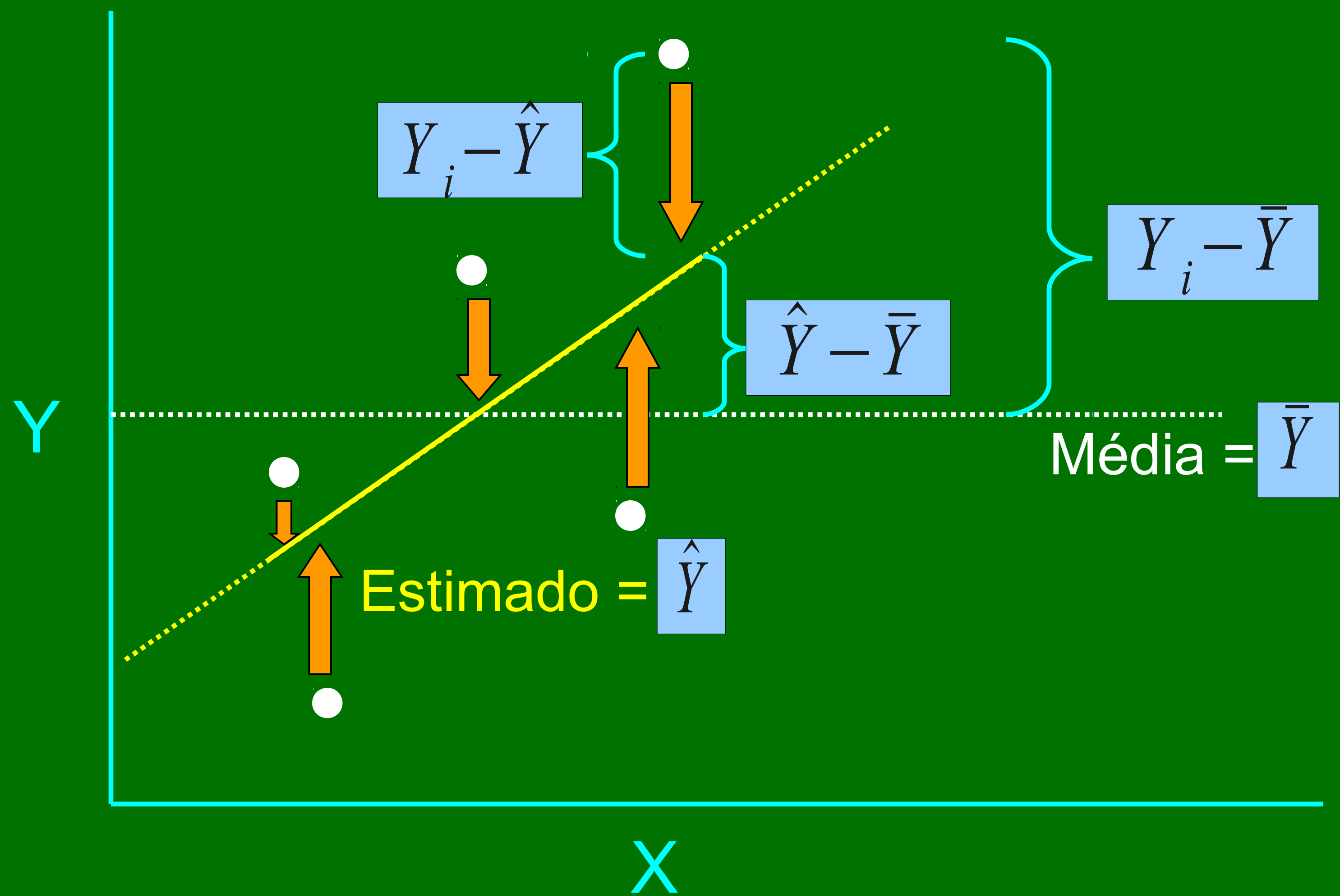
SSTotal (SST) = SSRegressão (SSR) + SSResíduo (SSE)

$$\underbrace{Y_i - \bar{Y}}_{\text{SST}} = \underbrace{\hat{Y}_i - \bar{Y}}_{\text{SSR}} + \underbrace{Y_i - \hat{Y}_i}_{\text{SSE}}$$

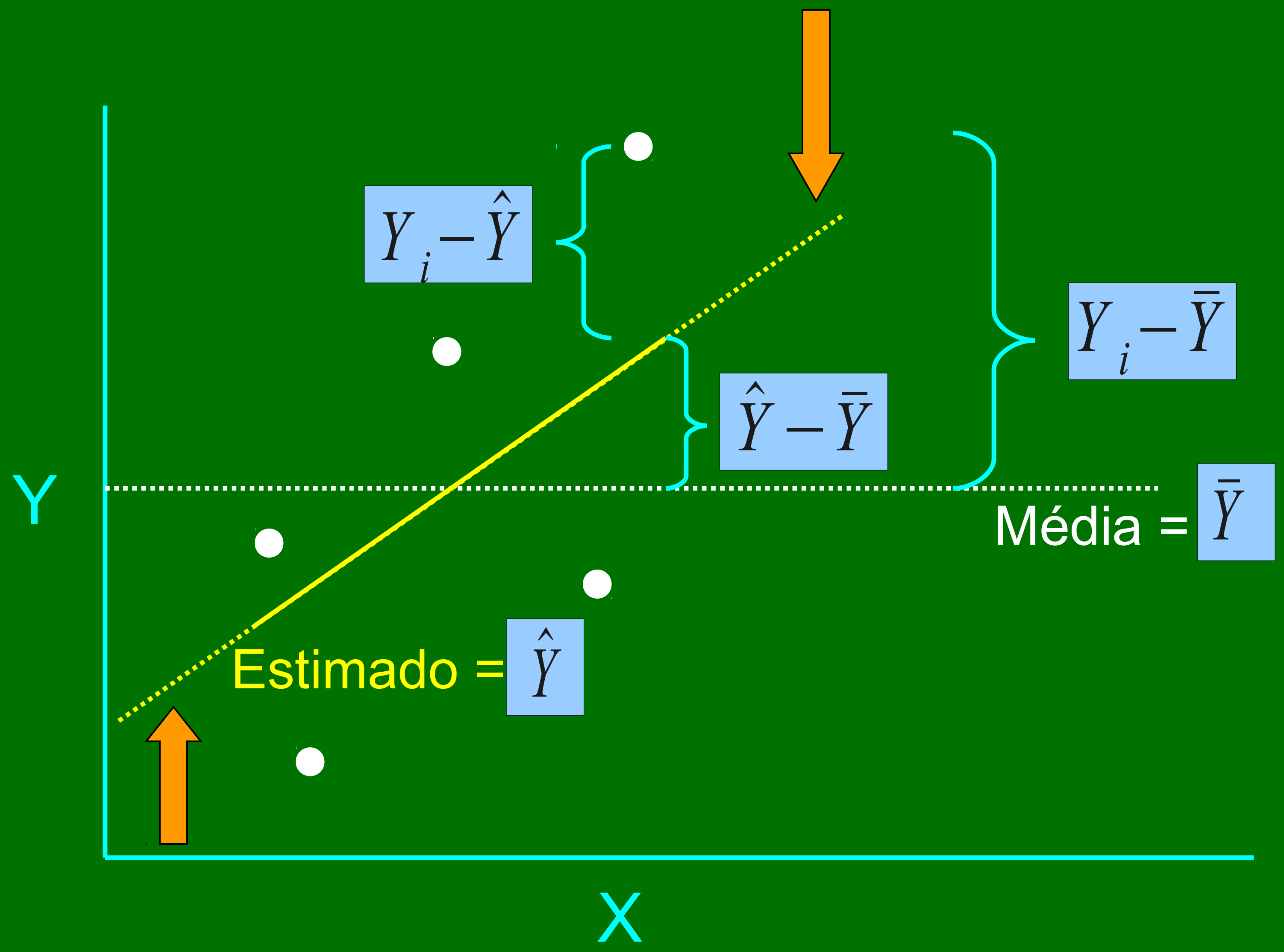


$Y_i - \hat{Y}$
$Y_i - \bar{Y}$
$\hat{Y} - \bar{Y}$
Y
Média = $\bar{Y}$
Estimado = $\hat{Y}$
X

O que acontece com $Y_i - \hat{Y}$ quando os pontos, mantendo a mesma média, se aproximam da reta?



O que acontece com $Y_i - \hat{Y}$ quando a reta de ajuste é igual à média?



Coeficiente de determinação

$$R^2 = \frac{SSR}{SST}$$

Coeficiente de Correlação

$$r = \pm\sqrt{R^2}$$

Um sinal de + ou – é anexado ao valor de acordo com a inclinação da reta

Relação entre b e r: o coeficiente angular é o coeficiente de correlação ‘re-escalado’ em relação aos desvios padrão de Y e X

$$b = r\left(\frac{s_y}{s_x}\right)$$

SSTotal (SST) = SSRegressão (SSR) + SSResíduo (SSE)

$$\underbrace{Y_i - \bar{Y}}_{SST} = \underbrace{\hat{Y}_i - \bar{Y}}_{SSR} + \underbrace{Y_i - \hat{Y}_i}_{SSE}$$

$$SST = \sum \left(Y_i - \bar{Y}\right)^2 \longrightarrow MST = Var = \frac{\sum \left(Y_i - \bar{Y}\right)^2}{n-1}$$

$$SSR = \sum \left(\hat{Y}_i - \bar{Y}\right)^2 \longrightarrow MSR = \frac{SSR}{1}$$

$$SSE = \sum \left(Y_i - \hat{Y}\right)^2 \longrightarrow MSE = \frac{SSE}{n-2}$$

SS Total (SST) = SS Regressão (SSR) + SS Resíduo (SSE)

$$Y_i - \bar{Y} = \left(\hat{Y}_i - \bar{Y}\right) + \left(Y_i - \hat{Y}_i\right)$$

$$SST = \sum \left(Y_i - \bar{Y}\right)^2$$

$$SSR = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$$

$$SSE = \sum \left(Y_i - \hat{Y}\right)^2$$

Quadrados Médios (MS) = variância = SS/df , onde df = n-1

Soma de Quadrados da Regressão (SSR)

Quadrados Médios da Regressão (MSR) = SSR/df , onde df = 1

Soma de Quadrados dos Resíduos (SSE)

Quadrados Médios dos Resíduos (MSE) = SSE/df, onde df = n-2

Inferências

Inferências (usando distribuição *t*)

"Quando uma estatística é padronizada (dividida pelo desvio padrão), mas o denominador é uma estimativa e não o valor verdadeiro, ela é chamada de uma *estatística estudentizada*.

Um importante teorema em estatística diz que:

$(b-\beta)/s\{b\}$ é distribuída como $t(n-2)$ para um modelo de regressão linear."

$$s^2\{b\} = \frac{MSE}{\sum\left(X_i - \bar{X}\right)^2}$$

Intervalo de Confiança para β : $b \pm t(1-(\alpha/2); n-2)s\{b\}$

Para saber t(1-(α/2); n-2) no R: `qt(.975,23)` para alfa=0.05 e n-2=23

Teste sobre β

Teste bicaudal:

$H_0: \beta = 0$

$H_1: \beta \neq 0$

Erro Tipo I em $\alpha = 0.05$

A conclusão pode ser obtida observando o intervalo de confiança 95% acima. Se o intervalo não incluir o 0, conclui-se H_1.

Um teste direto das duas alternativas é baseado na estatística:

$$t^* = b/s\{b\}$$

A decisão neste teste para o nível de significância em α é:

Se $|t^*| \leq t(1-(\alpha/2); n-2)$, conclui-se H_0 [Notar que é modulo da estatística t*]

Se $|t^*| > t(1-(\alpha/2); n-2)$, conclui-se H_a

Para achar *t* crítico no R: `qt((1-(0.05/2)), n-2)`

Podemos também achar a probabilidade exata:

`pt(t*,df)*2`, onde t^* é o valor de t observado.

No caso de teste bicaudal deve-se multiplicar a probabilidade encontrada por 2

Teste Unicaudal:

H_0: $\beta \leq 0$

H_1: $\beta > 0$

Se $|t^*| \leq t(1-\alpha; n-2)$, conclui-se H_0

Se $|t^*| > t(1-\alpha; n-2)$, conclui-se H_a

Para α=0.05, obtemos t(0.95,23) = 1.714 {no R: `qt(0.95,23)` }.

Se, por exemplo t^* =10.29 > 1.714, concluímos por H_a, que $\beta > 0$.

Teste F para $\beta = 0$ versus $\beta \neq 0$

$$F^* = MSR / MSE$$

Regras de decisão: visto que o teste é feito apenas para a cauda superior (*upper-tail*) e F^* é distribuído como $F(1,n-2)$ quando H_0 é verdadeiro, temos:

Se $F^* \leq F(1-\alpha; 1; n-2)$, conclui-se H_0
Se $F^* > F(1-\alpha; 1; n-2)$, conclui-se H_a
onde $F(1-\alpha; 1; n-2)$ é o $(1-\alpha)100$ percentil da distribuição F apropriada.

Exemplo:
Para $\alpha = 0.05$ e $n = 25$, obtemos $F(0.95; 1; 23) = 4.28$

No R: `qf(0.95,1,23)`

Equivalência do teste F *e do teste* t

$$F^* = (t^*)^2$$

$[t(1-\alpha/2;\ n-2)]^2 = F(1-\alpha;\ 1,\ n-2)$

NO R:

```
qf(0.95, 1, 23)
qt(0.975, 23)^2
```

O teste *t*, entretanto, é mais flexível pois pode ser usado para testes unicaudais

Tabela de Análise de Variância

Fonte	d.f.	SS	MS	F	p
Regressão	1	$SSR = \sum \left(\hat{Y}_i - \bar{Y}\right)^2$	SSR/1	$\frac{MSR}{MSE}$	1-pf(F,1,n-2)
Resíduo	n-2	$SSE = \sum \left(Y_i - \hat{Y}\right)^2$	SSE/n-2		
Total	n-1	$SST = \sum \left(Y_i - \bar{Y}\right)^2$	SST/n-1		

Testes Generalizados de Modelos Lineares

Modelo Completo (Full Model) (ou não restrito)

$$SSE(F) = \sum \left[Y_i - (a + b X_i) \right]^2 = \sum \left(Y_i - \hat{Y}_i \right)^2 = SSE$$

Modelo Reduzido (Reduced Model) (ou modelo restrito)

$$SSE(R) = \sum \left(Y_i - a \right)^2 = \sum \left(Y_i - \bar{Y} \right)^2 = SSTO$$

A lógica é comparar SSE(F) e SSE(R).
Observe que SSE(F) nunca é maior que SSE(R):

$SSE(F) \leq SSE(R)$

A razão é que conforme temos mais parâmetros no modelo, mais flexível ele é e portanto melhor se ajustará às observações. Consequentemente, os desvios em relação à reta serão menores. Quando SSE(F) não é muito menor que SSE(R), o uso do Modelo Completo não reduz muito mais a variabilidade de Y_i em relação ao Modelo Reduzido. Neste caso, o Modelo Reduzido (mais simples) é tido como adequado (i.e. que H_0 é adequada).

A estatística do teste é:

$$F = \frac{SSE(R) - SSE(F)}{df_R - df_F} \div \frac{SSE(F)}{df_F}$$

No caso de testar se β = 0, temos para o caso do modelo com 1 variável independente:

SSE(R) = SST(F)

Note que o modelo *Full* é o modelo COM β

Note que o modelo *Reduced* é o modelo SEM β

Substituindo na formula, fica MSR/MSE, que é idêntico ao teste por análise de variância. Em análises a serem vistas em aulas posteriores usaremos intensivamente este teste.

$$F = \frac{SST - SSE}{df_T - df_E} \div \frac{SSE}{df_E}$$

Exercícios e estudo individual:

-- Lista em sala de aula

-- Crawley: Cap. 8

-- Gotelli & Ellison: Cap. 9 (pp. 239-257)