Metodologia Ecológica

Aula 8

Estimativa de parâmetros em regressão linear simples

Mais de uma variável preditiva: regressão linear múltipla (mas nem sempre mais é melhor)

Prof. Marcus Vinícius Vieira - Instituto de Biologia UFRJ

O que é um modelo?

"Todos os modelos estão errados, mas alguns são úteis" (Box 1976)

Modelos matemáticos: y = b*x

Exemplo: Faturamento em loja de sorvete

Cada sorvete = 2 reais

Se vendeu 3 sorvetes, faturamento = 6 reais

Se vendeu 7 sorvetes, faturamento = 14 reais

Se vendeu 13 sorvetes, faturamento = 26 reais

Generalizando: Y = 2*X

O que é um modelo estatístico?

Modelos estatísticos: y = b*x + e

Cada sorvete = *em média* 2 reais (depende do freguês!)

Se vendeu 3 sorvetes, faturamento = *em média* 6 reais

Se vendeu 7 sorvetes, faturamento = *em média* 14 reais

Se vendeu 13 sorvetes, faturamento = *em média* 26 reais

Generalizando: Y = 2*X + e

Passos para construção de um modelo estatístico

- O quê se quer estimar?
  - O número de espécies em função da área
- Defina um modelo:
  - LogEspécies = z*LogÁrea +
- Defina uma função de densidade/distribuição de probabilidade para a variação residual

Relação espécies-área (Preston 1962)

Iha	Area	Nespecies	LogArea	LogEspecies
Albemarle	5824.9	325	3.765	2.512
Charles	165.8	319	2.219	2.504
Chatham	505.1	306	2.703	2.486
James	525.8	224	2.721	2.350
Indefatigable	1007.5	193	3.003	2.286
Abingdon	51.8	119	1.714	2.076
Duncan	18.4	103	1.265	2.013
Narborough	634.6	80	2.802	1.903
Hood	46.6	79	1.669	1.898
Seymour	2.6	52	0.413	1.716
Barringon	19.4	48	1.288	1.681
Gardner	0.5	48	-0.286	1.681
Bindloe	116.6	47	2.067	1.672
Jervis	4.8	42	0.685	1.623
Tower	11.4	22	1.057	1.342
Wenman	4.7	14	0.669	1.146
Culpepper	2.3	7	0.368	0.845

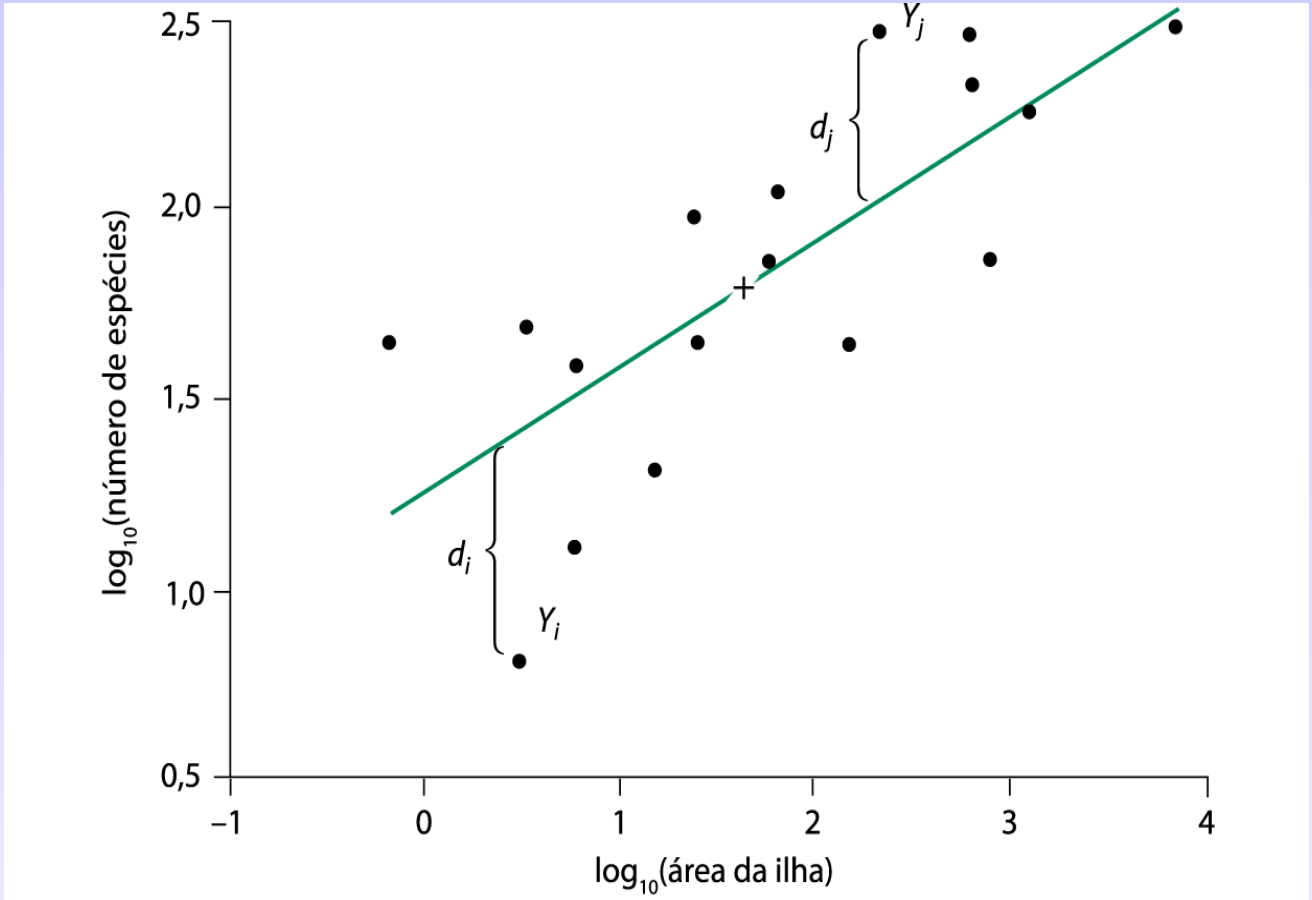


Figura 9.3 A soma dos quadrados dos resíduos é obtida somando os desvios quadrados (d_i) de cada observação da linha de regressão ajustada. A estimativa do parâmetro dos mínimos quadrados garante que essa linha de regressão minimize a soma dos quadrados dos resíduos. O + marca o ponto central dos dados ($\bar{X}$, $\bar{Y}$). Essa linha de regressão descreve, ainda, a relação entre o logaritmo da área das ilhas e o do número de espécies de plantas das Ilhas de Galápagos, dados da Tabela 8.2. A equação da regressão é $\log_{10}(\text{espécies}) = 1{,}320 + \log_{10}(\text{área}) \times 0{,}331$; $r^2 = 0{,}584$.

Definindo uma linha reta e seus dois parâmetros:

inclinação e intercepto

Fig. 9.1

Interpolação e Extrapolação: Fig. 9.2

Ajustando dados a um modelo linear:

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

$$d_i = (Y_i - \hat{Y}_i)^2$$

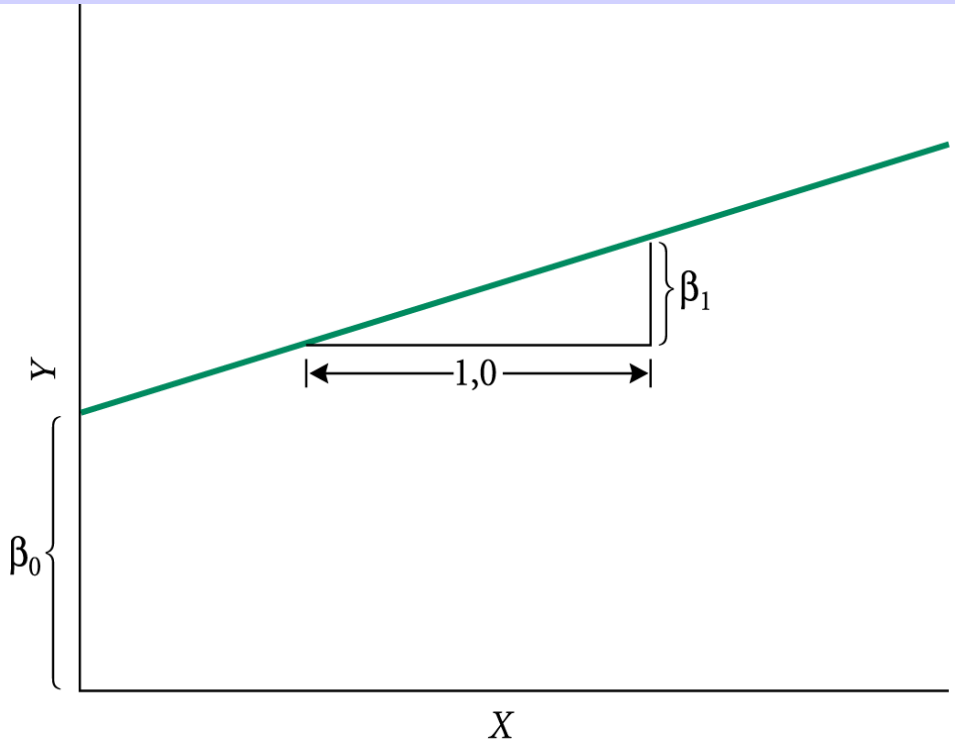


Figura 9.1 Relação linear entre as variáveis *X* e *Y*. A linha é descrita pela equação $Y = \beta_0 + \beta_1 X$, onde β_0 é o intercepto e β_1 é a inclinação da linha. O intercepto β_0 é o valor predito da equação quando $X = 0$. A inclinação da linha β_1 é o aumento na variável *Y* associado com o de uma unidade da variável *X* ($\Delta Y/\Delta X$). Se o valor de *X* é conhecido, o valor predito de *Y* pode ser calculado multiplicando *X* pela inclinação e somando o intercepto (β_0).

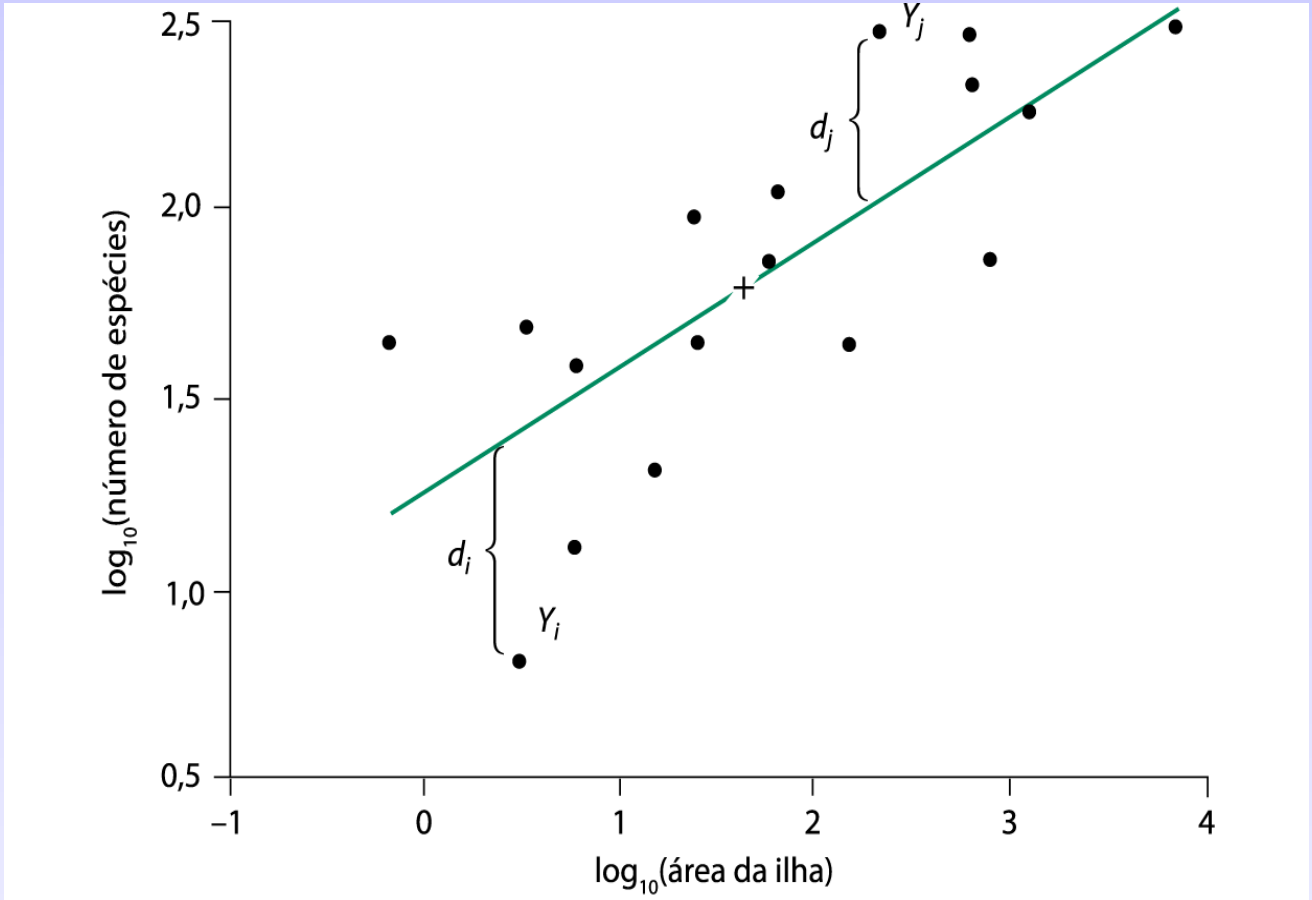


Figura 9.3 A soma dos quadrados dos resíduos é obtida somando os desvios quadrados (d_i) de cada observação da linha de regressão ajustada. A estimativa do parâmetro dos mínimos quadrados garante que essa linha de regressão minimize a soma dos quadrados dos resíduos. O + marca o ponto central dos dados ($\bar{X}$, $\bar{Y}$). Essa linha de regressão descreve, ainda, a relação entre o logaritmo da área das ilhas e o do número de espécies de plantas das Ilhas de Galápagos, dados da Tabela 8.2. A equação da regressão é $\log_{10}(\text{espécies}) = 1{,}320 + \log_{10}(\text{área}) \times 0{,}331$; $r^2 = 0{,}584$.

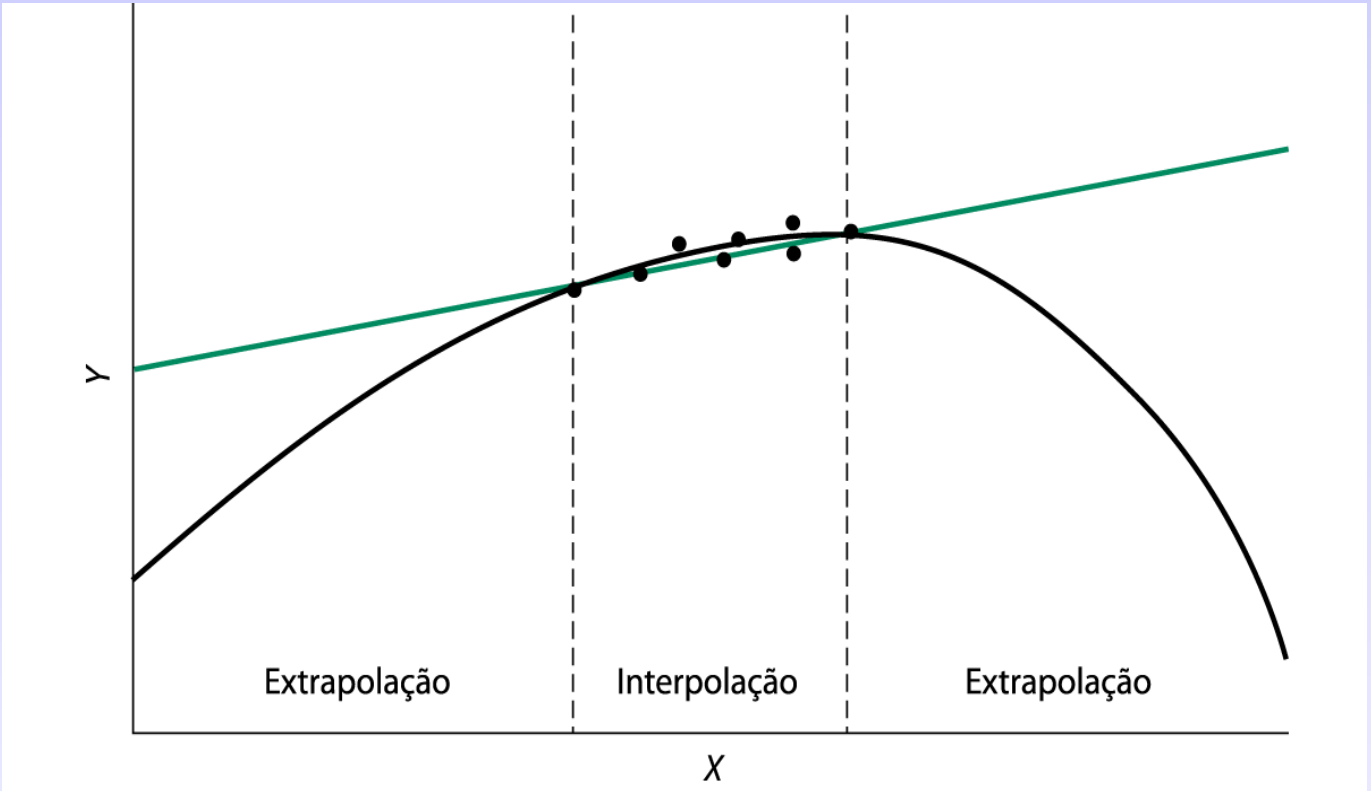


Figura 9.2 Modelos lineares podem aproximar funções não lineares sobre um domínio limitado da variável *X*. A interpolação dentro desses limites pode ser aceitável e acurada, embora o modelo linear (linha verde) não descreva a verdadeira relação funcional entre *Y* e *X* (curva preta). A extrapolação se tornará crescentemente menos acurada conforme as previsões se movem para além da amplitude dos dados coletados. Uma premissa muito importante da regressão linear é que a relação entre *X* e *Y* (ou transformações dessas variáveis) é linear.

Soma dos quadrados dos resíduos (=Residual Sum of Squares, RSS):

$$SQR = \sum_{i-1}^{n} \left(Y_i - \hat{Y}_i \right)^2$$

$$SQ_Y = \sum_{i-1}^{n} \left(Y_i - \bar{Y} \right)^2$$

$$s^{2_Y} = \frac{1}{n-1} \sum_{i-1}^{n} \left(Y_i - \bar{Y} \right)^2$$

Somatório dos produtos:

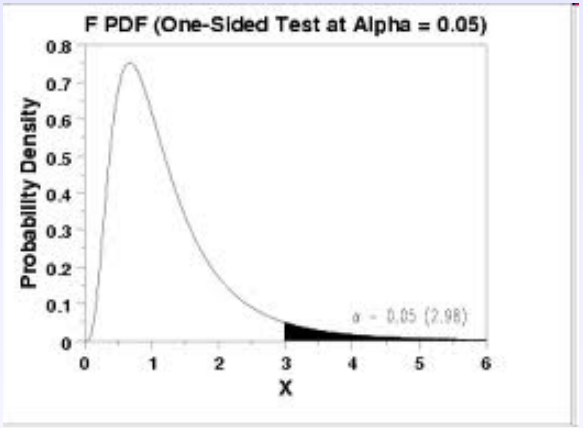
$$SQ_{XY} = \sum_{i-1}^{n} \left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right)$$

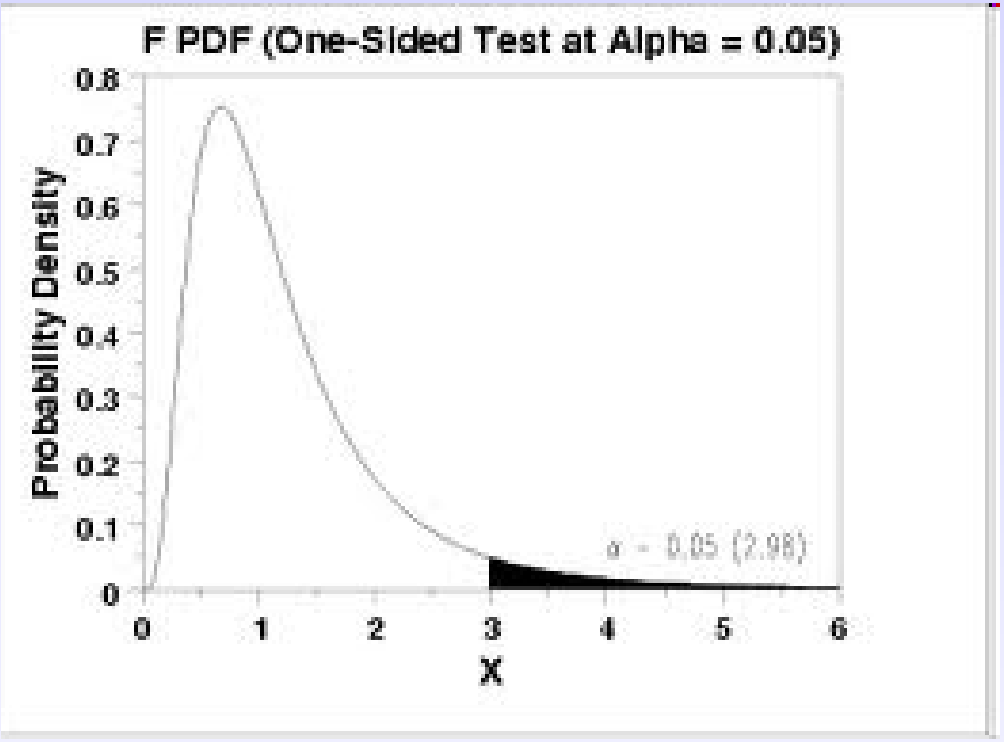
Covariância amostral:

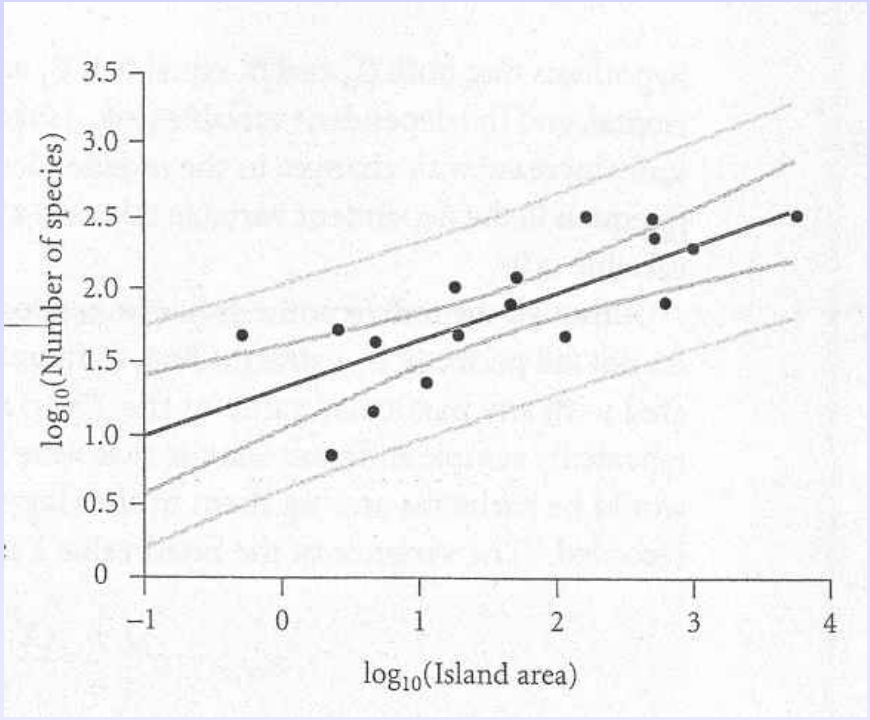
$$s_{XY} = \frac{1}{n-1} \sum_{i-1}^{n} \left(X_i - \bar{X} \right) \left(Y_i - \bar{Y} \right)$$

TABLE 9.1 Complete ANOVA table for single factor linear regression

Source	Degrees of freedom (df)	Sum of squares (*SS*)	Mean square (*MS*)	Expected mean square	F-ratio	*P*-value
Regression	1	$SS_{reg} = \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$	$\frac{SS_{reg}}{1}$	$\sigma^2 + \beta_1^2 \sum_{i=1}^{n} X^2$	$\frac{SS_{reg}/1}{RSS/(n-2)}$	Tail of the F distribution with $1, n-2$ degrees of freedom
Residual	$n-2$	$RSS = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$	$\frac{RSS}{(n-2)}$	σ^2		
Total	$n-1$	$SS_Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2$	$\frac{SS_Y}{(n-1)}$	σ_Y^2		

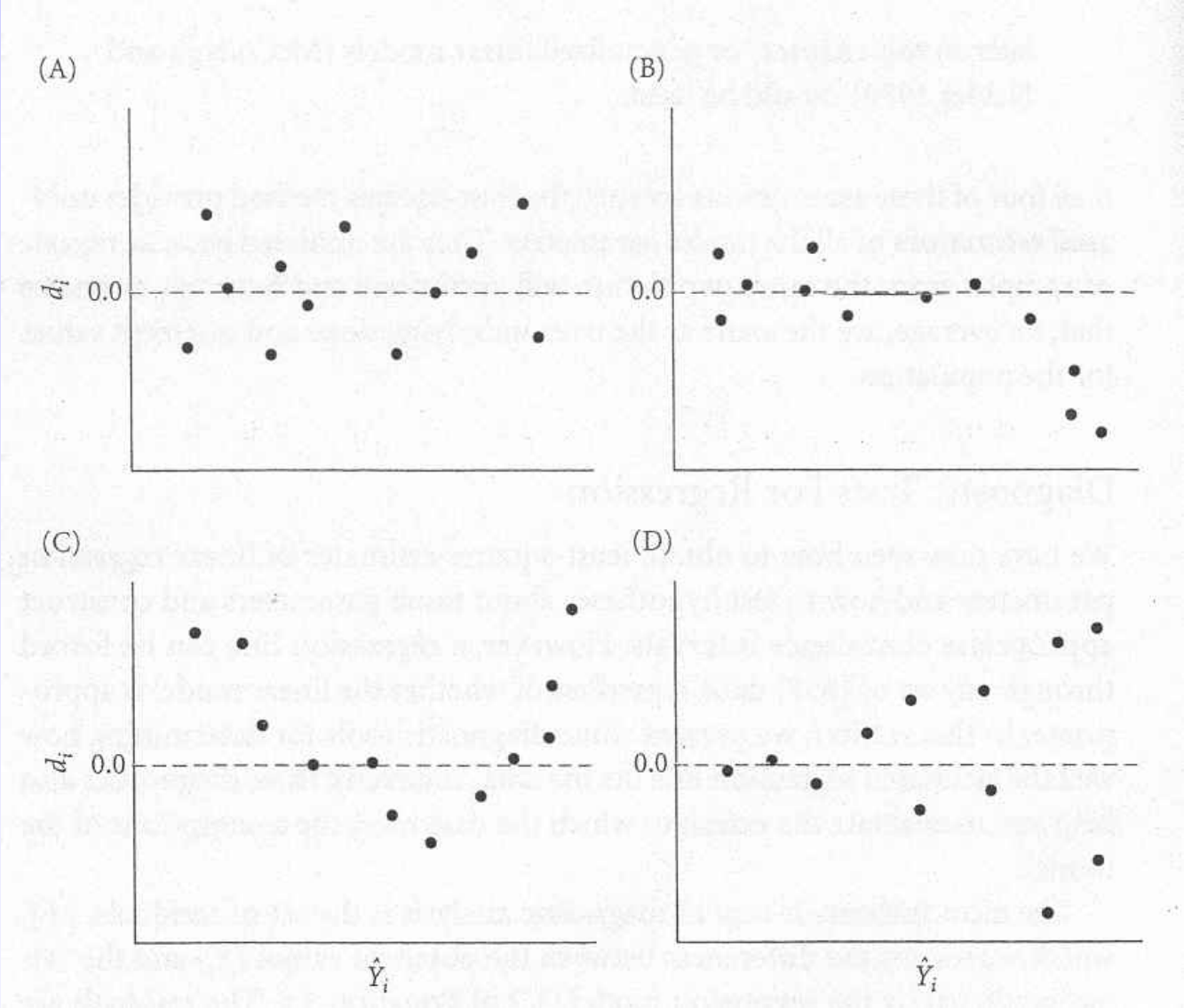


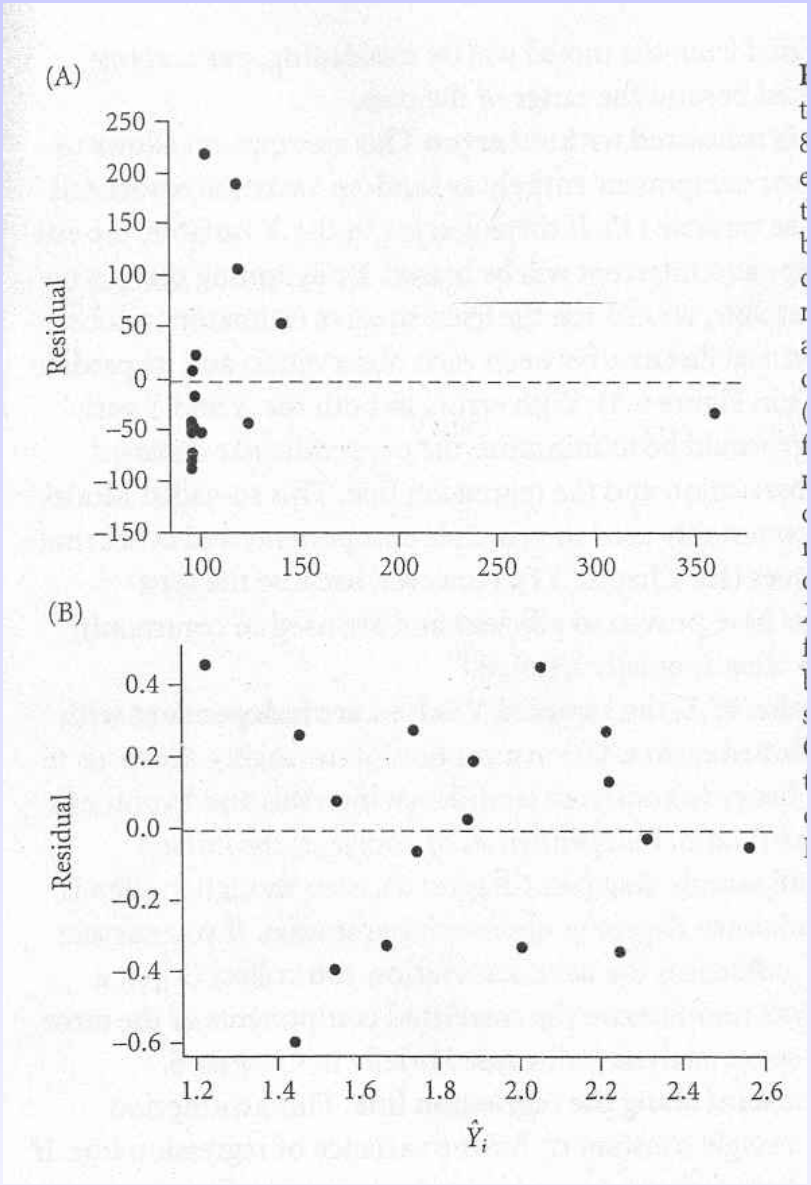

F PDF (One-Sided Test at Alpha = 0.05)
Probability Density
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
1
2
3
4
5
6
X
α = 0.05 (2.98)

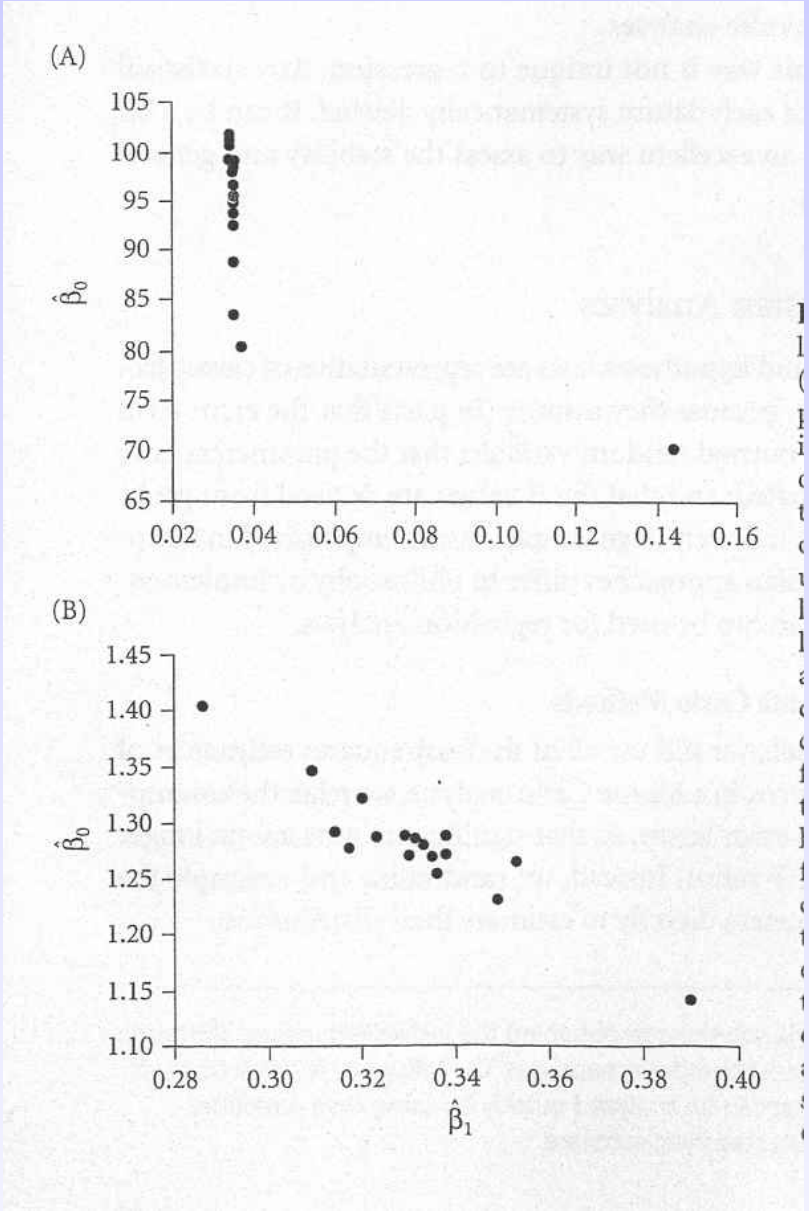

3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
log10(Number of species)
−1
0
1
2
3
4
log10(Island area)

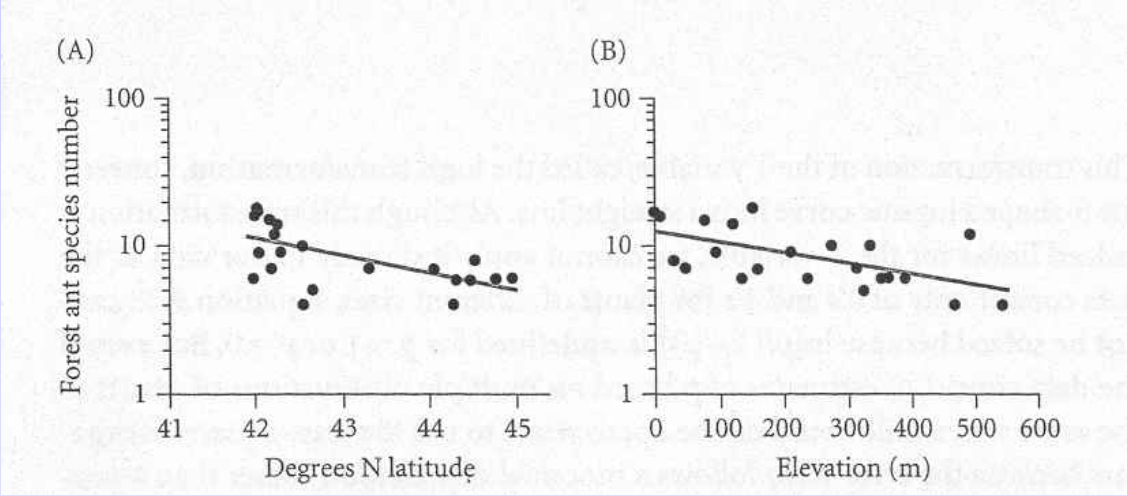
Considere os seguintes dados:
Variável dependente (Y) = riqueza: 12, 14, 17, 20, 19
Variável independente (X) = hetero: 20, 30, 40, 50, 60

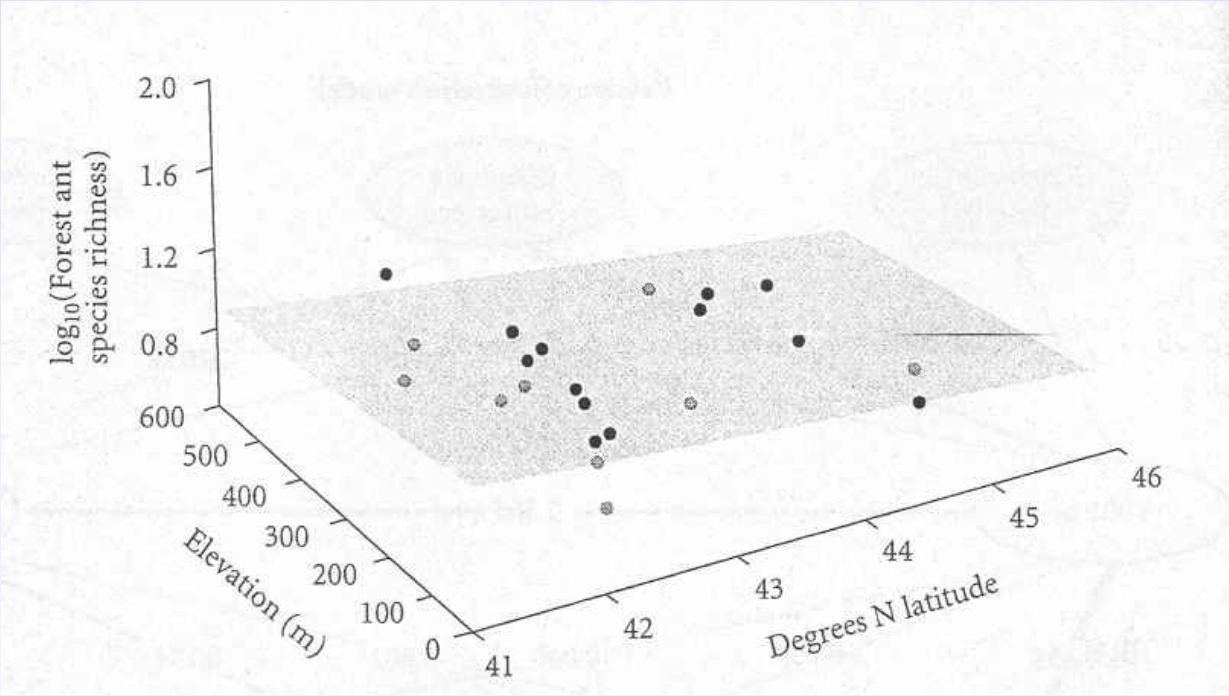
a) Obtenha a Soma de Quadrados Total
c) Obtenha os coeficientes da regressão usando as formulas
d) Obtenha os resíduos
c) Obtenha a Soma de Quadrados dos Resíduos
f) Obtenha a Soma de Quadrados da Regressão
g) Faça um gráfico aproximado contendo
    i) os valores observados,
    ii) a reta ajustada
    iii) indicação do resíduo.
h) Encontre os valores médios de X e de Y. O ponto definido por estas duas médias esta sobre a reta ajustada?

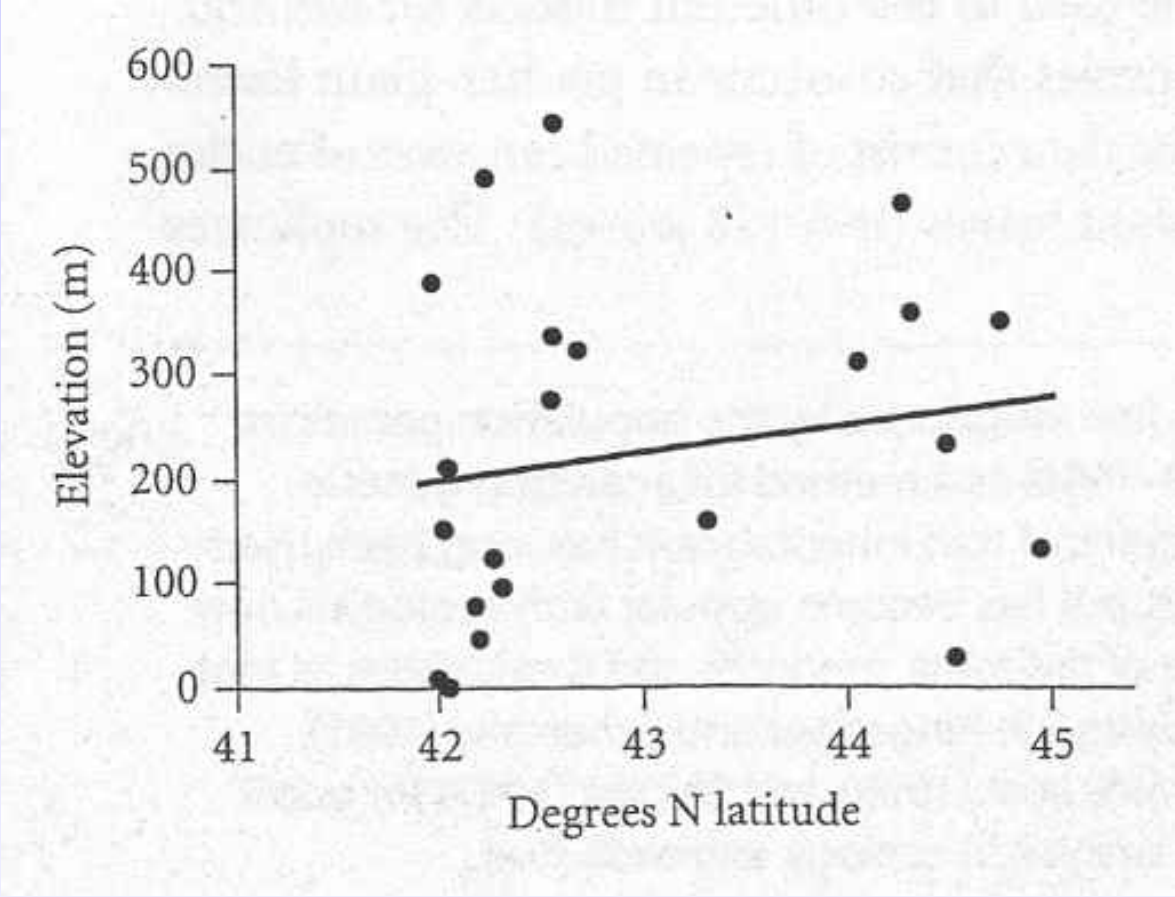

(A)
(B)
(C)
(D)
d_i
0.0
$\hat{Y}_i$


(A)
Residual
250
200
150
100
50
0
−50
−100
−150
100
150
200
250
300
350
(B)
Residual
0.4
0.2
0.0
−0.2
−0.4
−0.6
1.2
1.4
1.6
1.8
2.0
2.2
2.4
2.6
$\hat{Y}_i$


(A)
$\hat{\beta}_0$
105
100
95
90
85
80
75
70
65
0.02
0.04
0.06
0.08
0.10
0.12
0.14
0.16
(B)
$\hat{\beta}_0$
1.45
1.40
1.35
1.30
1.25
1.20
1.15
1.10
0.28
0.30
0.32
0.34
0.36
0.38
0.40
$\hat{\beta}_1$


(A)
(B)
Forest ant species number
100
10
1
41
42
43
44
45
Degrees N latitude
0
100
200
300
400
500
600
Elevation (m)


log10(Forest ant species richness)
2.0
1.6
1.2
0.8
600
500
400
300
200
100
0
Elevation (m)
41
42
43
44
45
46
Degrees N latitude


600
500
400
300
200
100
0
Elevation (m)
41
42
43
44
45
Degrees N latitude