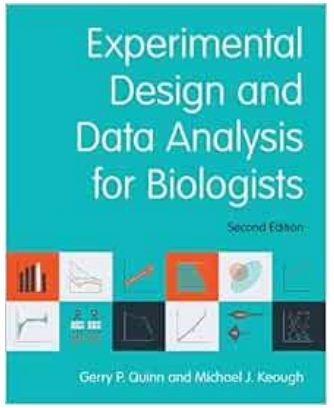
Aula 9 – Modelos Lineares: Regressão (e correlação) linear

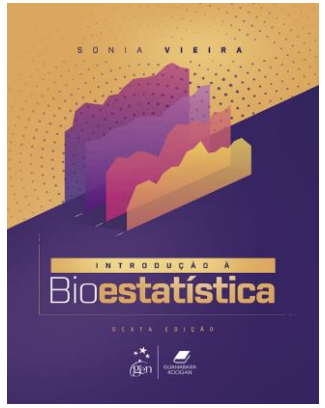
Ao final da aula você deverá saber:

- Breve histórico
- Introdução aos Modelos Lineares
- Equação da regressão linear (Ordinary Least Squares, OLS)
  - Parâmetros estimados
    - Intercepto (”a”, alpha, β_0)
    - Inclinação (slope: b, β, β_1)
- Teste de hipótese em regressão
- Método dos mínimos quadrados (least squares)
- Tabela de ANOVA
- Coeficiente de determinação (R^2)
- Pressupostos
- Diagnose dos resíduos (inspeção gráfica)

Aula baseada em

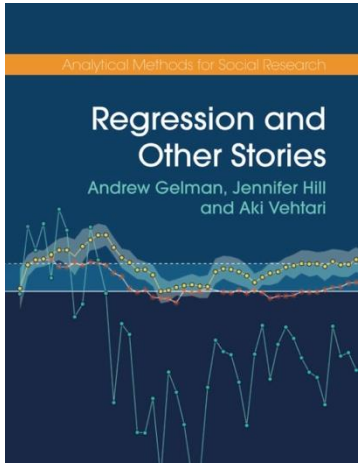


Cap 6

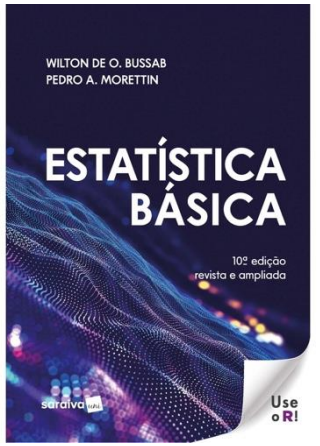


Cap 5 e 6

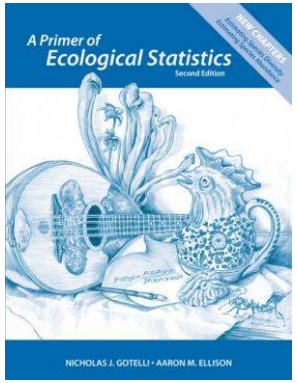
Cap 10



Cap 7 e 8

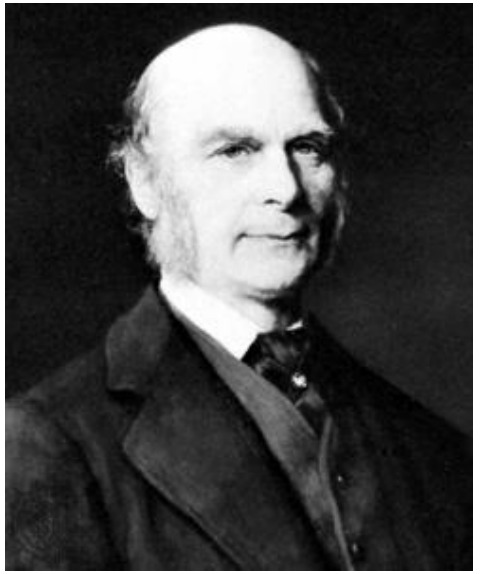


Cap 16



Cap 9

Um pouco de história



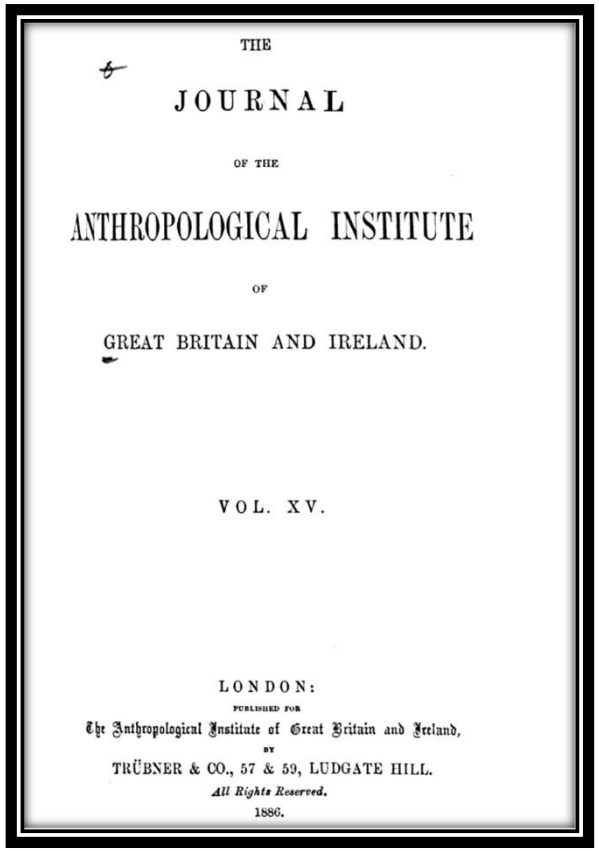
- Francis Galton (1822-1911): Antropólogo inglês, primo de Charles Darwin, defendia idéias de *eugenia*
- Cunhou o termo ”regressão” no contexto de estudos de herdabilidade de caracteres quantitativos em 1886
- Mediu **altura** de filhos e seus pais para investigar o padrão de herdabilidade desta característica

> “the average regression of the offspring is a constant fraction of their respective mid-parental deviations”.

ANTHROPOLOGICAL MISCELLANEA.

Regression *towards* Mediocrity *in* Hereditary Stature.

By Francis Galton, F.R.S., &c.



THE

JOURNAL

OF THE

ANTHROPOLOGICAL INSTITUTE

OF

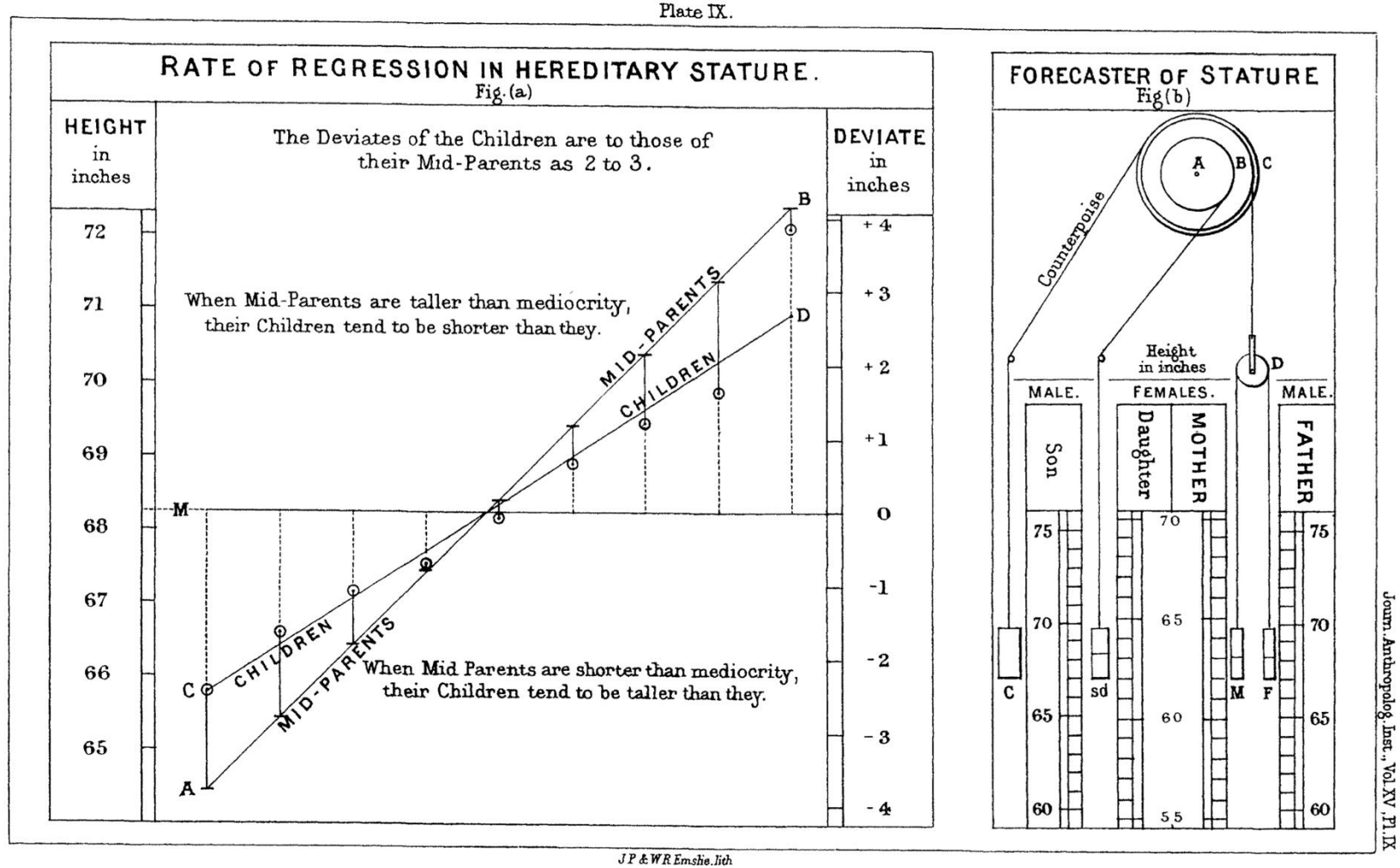
GREAT BRITAIN AND IRELAND.

VOL. XV.

LONDON:
PUBLISHED FOR
The Anthropological Institute of Great Britain and Ireland,
BY
TRÜBNER & CO., 57 & 59, LUDGATE HILL.

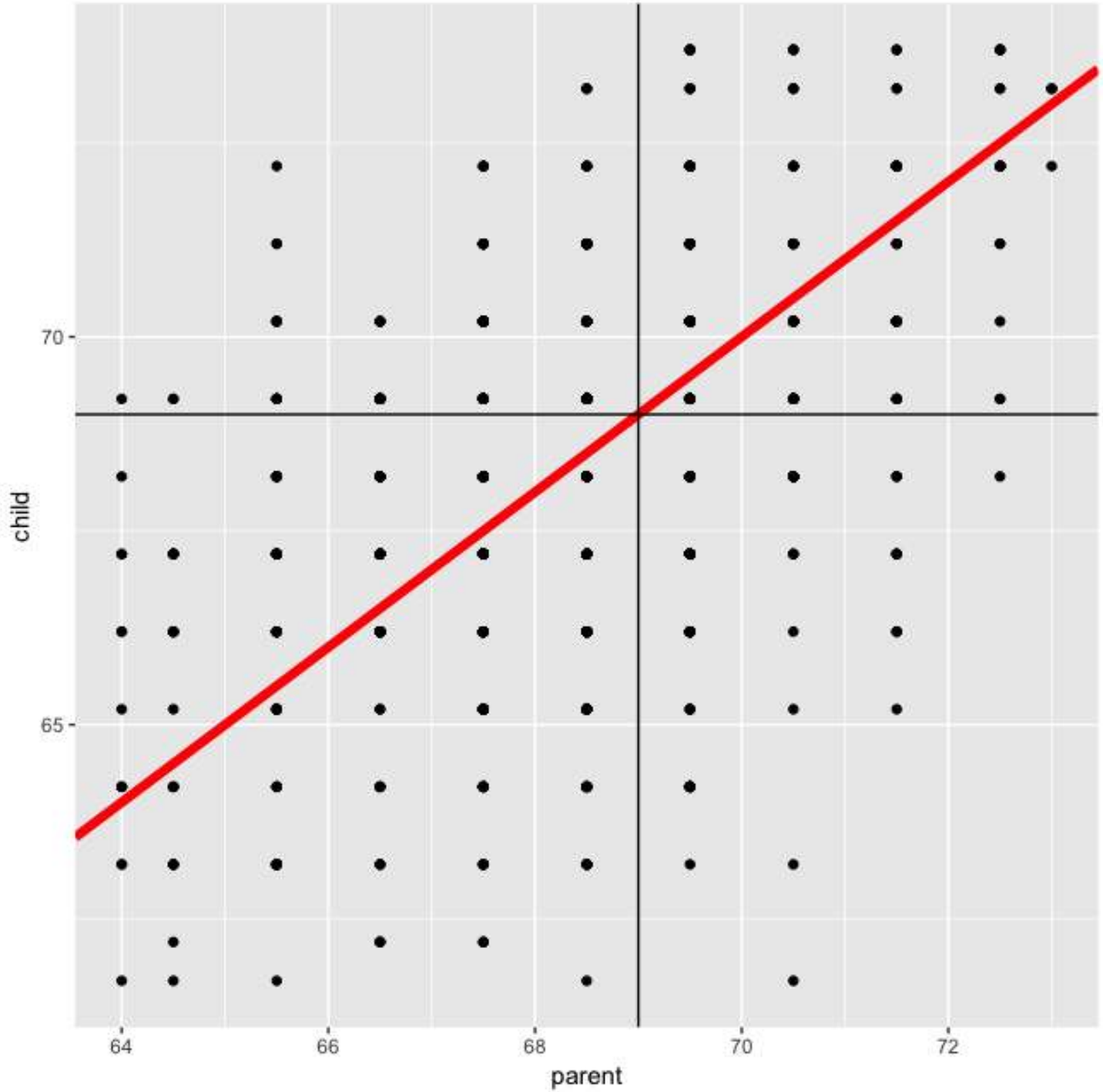
1886.

Plate IX.



J P & W R Emslie, lith.

Journ. Anthropolog. Inst., Vol. XV., Pl. IX.


70
65
child
64
66
68
70
72
parent

Breve histórico

- A diferença entre um filho e seus dois pais numa dada característica é proporcional ao desvio de seus pais da média da população.
- Por exemplo, se seus pais são cada um 2 cm mais altos que a média dos homens e mulheres, em média, o seu filho será mais baixo que seus pais por um fator $(1-R^2)$ vezes 2 cm.
- Nos dados de Galton, a altura de uma pessoa seria um ponto médio de cerca de 2/3 o desvio dos pais da média da população
- Logo, a altura dos filhos estaria *regredindo* à média
- Em seguida, K. Pearson trabalha no problema da correlação e regressão, derivando o coeficiente de regressão (r)

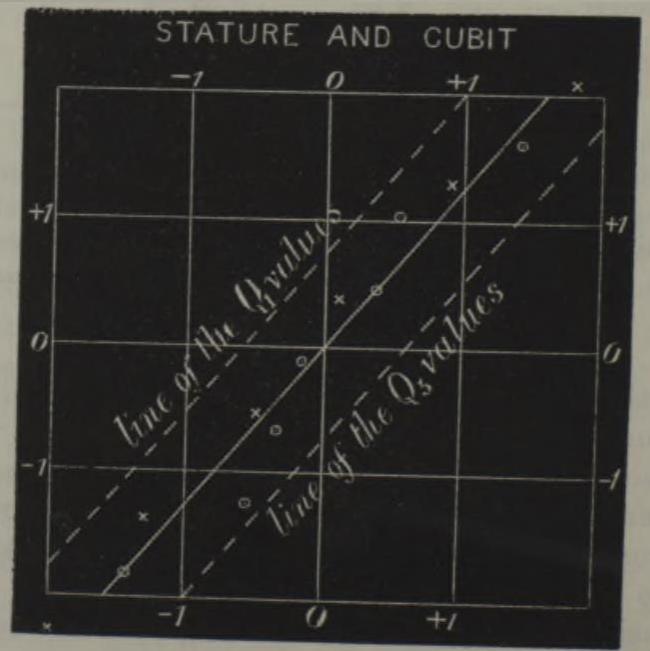
December 20, 1888.

Professor G. G. STOKES, D.C.L., President, in the Chair.

The Presents received were laid on the table, and thanks ordered for them.

The following Papers were read:—

I. "Co-relations and their Measurement, chiefly from Anthropometric Data." By FRANCIS GALTON, F.R.S. Received December 5, 1888.



PROCEEDINGS

OF THE

ROYAL SOCIETY OF LONDON.

From November 15, 1888, *to April* 11, 1889.

VII. *Mathematical Contributions to the Theory of Evolution.—III. Regression, Heredity, and Panmixia.*

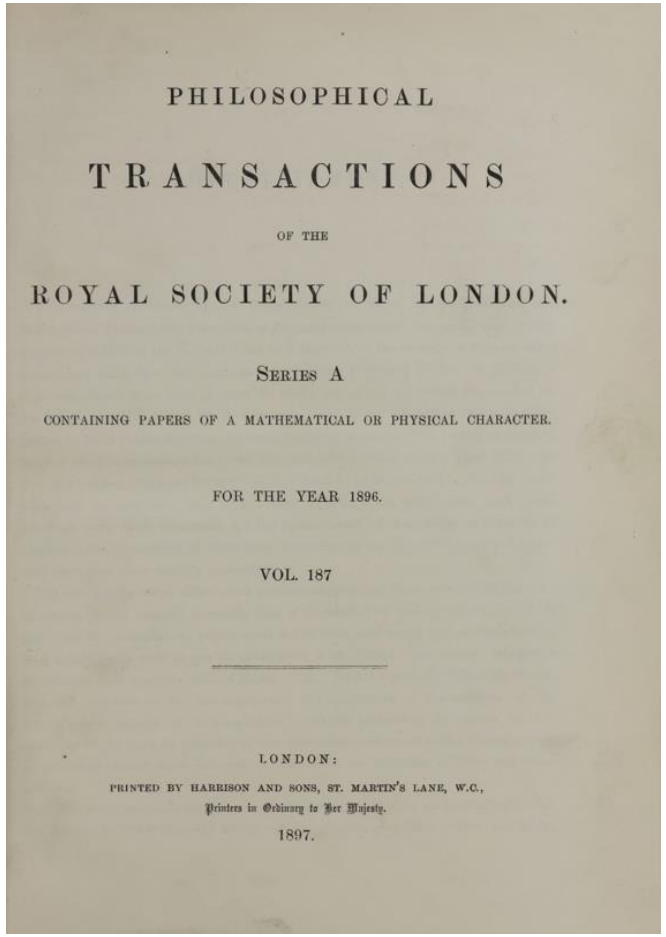
By KARL PEARSON, *University College, London.*

Communicated by Professor HENRICI, *F.R.S.*

Received September 28,—Read November 28, 1895.
Revised November 29, 1895.

Karl Pearson (1857-1936)

Define coeficiente de regressão


PHILOSOPHICAL

TRANSACTIONS

OF THE

ROYAL SOCIETY OF LONDON.

SERIES A

CONTAINING PAPERS OF A MATHEMATICAL OR PHYSICAL CHARACTER.

FOR THE YEAR 1896.

VOL. 187

LONDON:

PRINTED BY HARRISON AND SONS, ST. MARTIN'S LANE, W.C.,
Printers in Ordinary to Her Majesty.

1897.

https://www.npg.org.uk/collections/search/portrait/mw166224/Karl-Pearson-Sir-Francis-Galton

Karl Pearon e
Francis Galton
em 1909

Tabela 1. Sugestão de alguns testes estatísticos a empregar de acordo com o tipo de variável observada. Entre parênteses alguns testes não-paramétricos.

Variável Dependente	Variável Independente	Teste
Quantitativa	1 Categórica com 2 níveis	Teste t (teste U)
Quantitativa	1 Categórica com + 2 níveis	ANOVA 1-fator (Kruskall-Wallys)
Quantitativa	2 Categóricas	ANOVA 2-fatores (Friedman[1])
Quantitativa	1 Quantitativa	Regressão simples (correlação Spearman)
Quantitativa	2 ou mais quantitativas	Regressão múltipla
Quantitativa	1 categórica e 1 ou mais quantitativas	ANCOVA
Categórica	1 Categórica	Qui-quadrado[2]; Teste G[2]
Categórica	2 ou mais categóricas	Log-linear[2]

(1) No caso de amostras dependentes, (2) Esses testes eventualmente verificam não a relação de dependência entre variáveis, mas sim a associação entre elas, descaracterizando, portanto a classificação de variáveis dependentes e independentes.

Apostila Paglia & de Marco

Correlação Linear

Coeficiente de Correlação

$$r_{Y_1 Y_2} = \frac{\sum_{i=1}^{n} [(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)]}{\sqrt{\sum_{i=1}^{n} (y_{i1} - \bar{y}_1)^2 \sum_{i=1}^{n} (y_{i2} - \bar{y}_2)^2}}$$

Relação linear entre duas variáveis, mas sem assumir uma dependência funcional entre as duas

$-1 < r > 1$

Coeficiente de Correlação

$$r_{Y_1 Y_2} = \frac{\sum_{i=1}^{n} [(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)]}{\sqrt{\sum_{i=1}^{n} (y_{i1} - \bar{y}_1)^2 \sum_{i=1}^{n} (y_{i2} - \bar{y}_2)^2}}$$

Soma dos produtos cruzados (Covariância)

Desvio padrão das duas variáveis

Coeficiente de Correlação

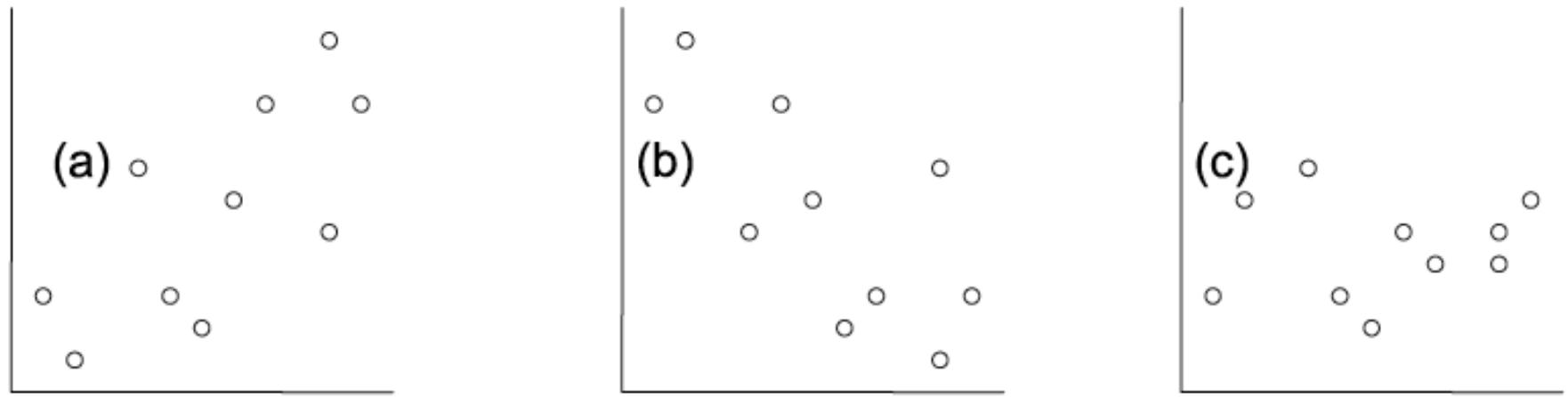
$$r_{Y_1 Y_2} = \frac{\sum_{i=1}^{n} [(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)]}{\sqrt{\sum_{i=1}^{n} (y_{i1} - \bar{y}_1)^2 \sum_{i=1}^{n} (y_{i2} - \bar{y}_2)^2}}$$

Pode ser positivo ou negativo

Sempre positivo

Se o numerador for negativo, o r será negativo e vice-versa

Figure 5.2 Scatterplots illustrating (a) a positive linear relationship ($r = 0.72$), (b) a negative linear relationship ($r = -0.72$), (c) and (d) no relationship ($r = 0.10$ and -0.17), respectively, and (e) a nonlinear relationship ($r = 0.08$).

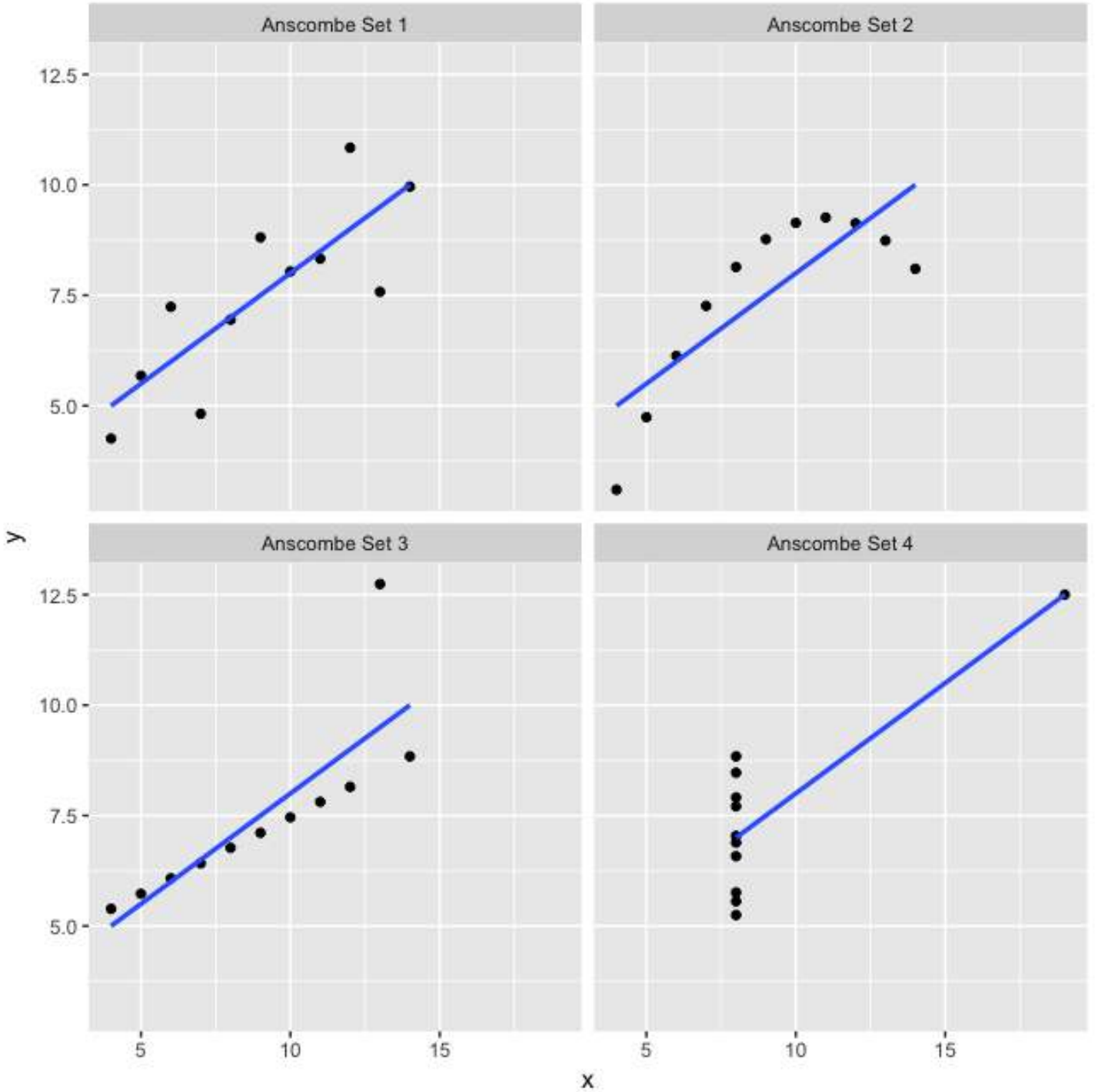


Quinn & Keough 2002

Quarteto de Anscombe

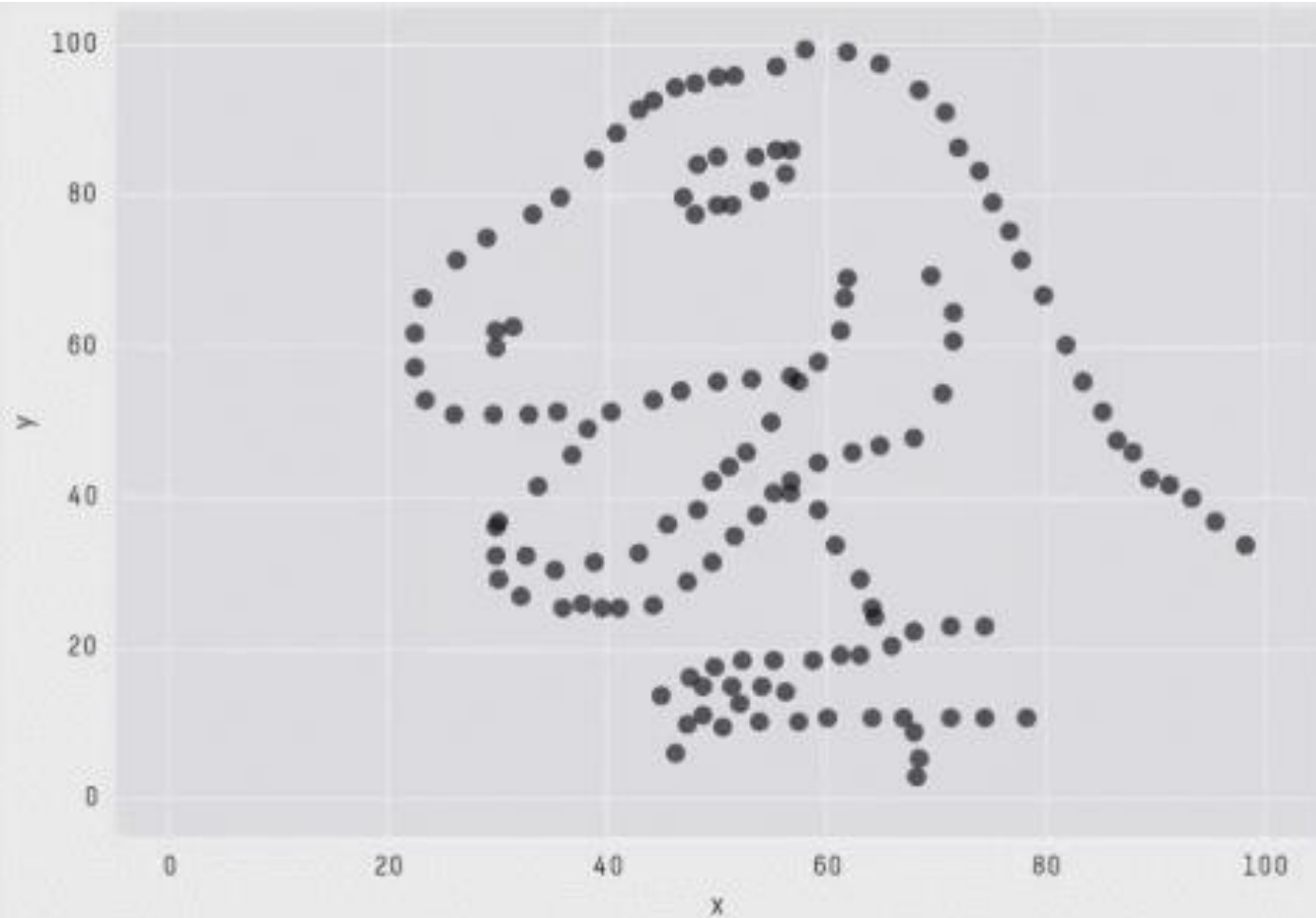
Sempre inspecione seus dados antes de fazer qualquer análise!!

https://rstudio-pubs-static.s3.amazonaws.com/52381_36ec82827e4b476fb968d9143aec7c4f.html



Todos esses plots têm o mesmo coeficiente de correlação

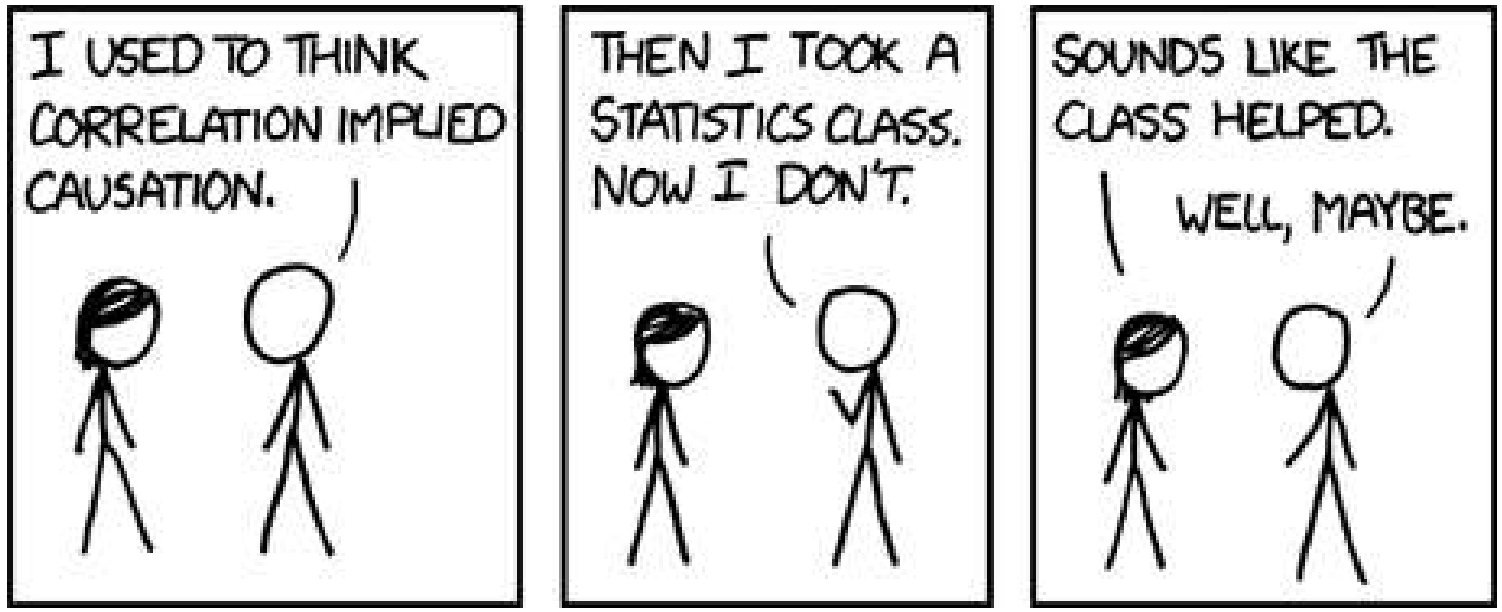
r = 0.816



X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526

https://www.autodeskresearch.com/publications/samestats

CORRELATION DOES NOT IMPLY CAUSATION

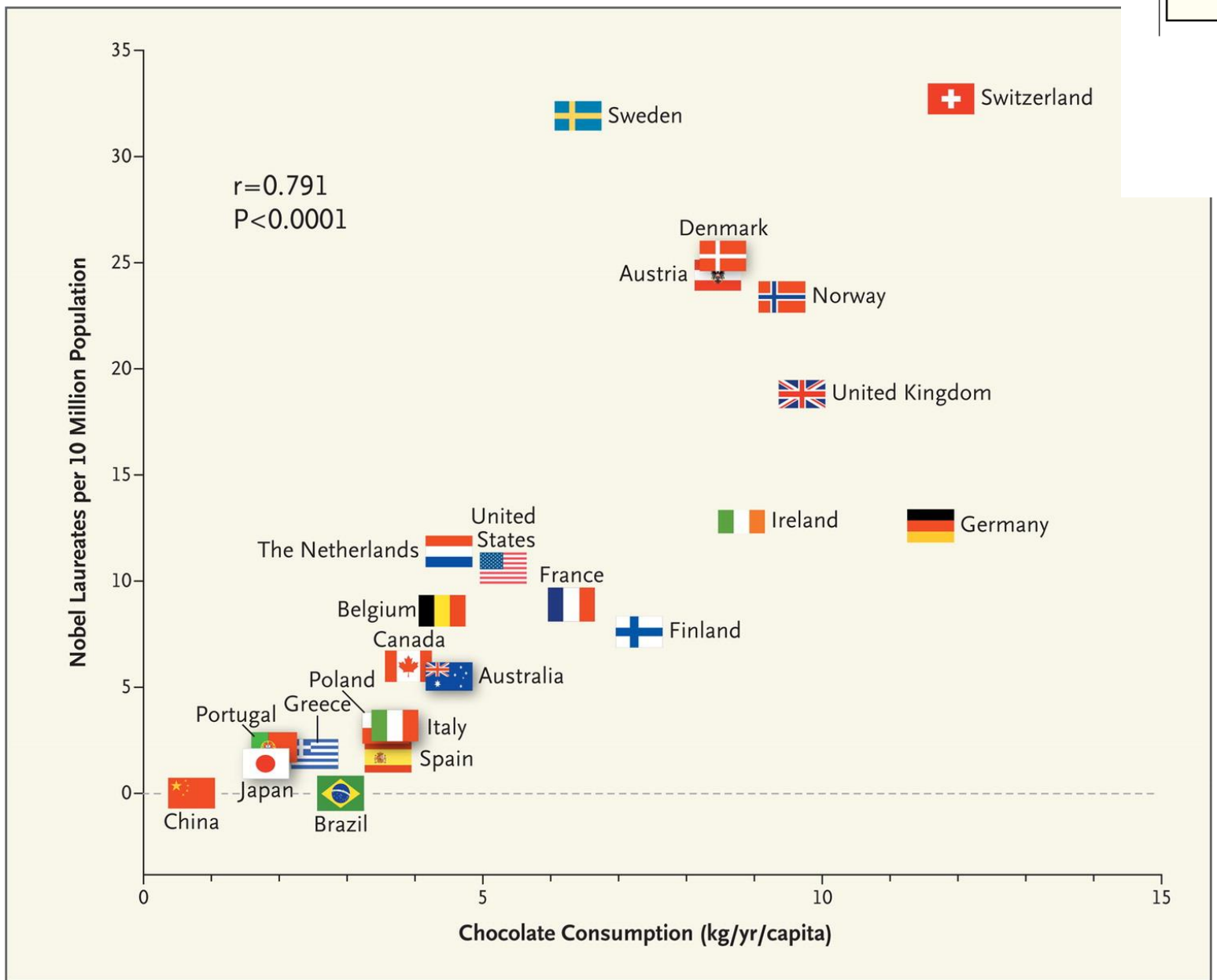


The NEW ENGLAND JOURNAL of MEDICINE

OCCASIONAL NOTES

Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.



spurious correlations

correlation is not causation

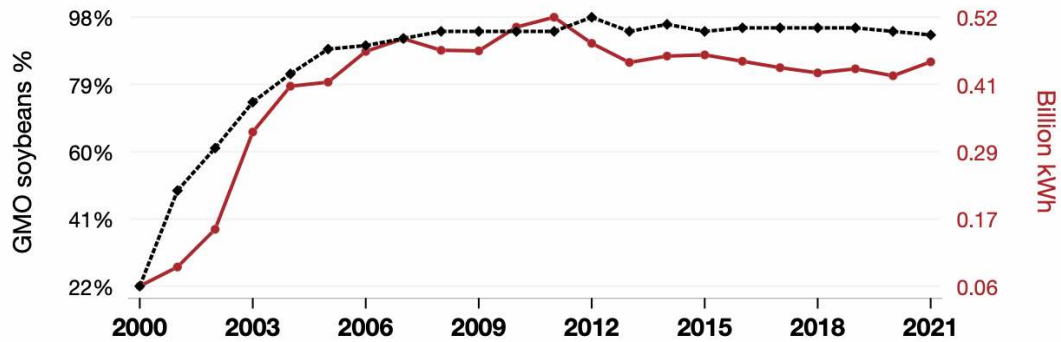
random · discover · *next page →*

don't miss spurious scholar,
where each of these is an academic paper

GMO use in soybeans in North Dakota

correlates with

Geothermal power generated in Russia



Percent of soybeans planted in North Dakota that are genetically modified to be herbicide-tolerant (HT), but not insect-resistant (Bt) · Source: USDA

Total geothermal power generated in Russia in billion kWh · Source: Energy Information Administration

2000-2021, r=0.951, r^2=0.905, p<0.01 · tylervigen.com/spurious/correlation/2699

View details about correlation #2,699

Show scatterplot

What else correlates?
GMO use in soybeans in North Dakota · all food
Geothermal power generated in Russia · all energy

https://www.tylervigen.com/spurious-correlations

Teste de hipótese do r de Pearson

- Utiliza um teste *t* de uma amostra:

$$t = \frac{r}{s_r}$$

Erro padrão do *r*

$$s_r = \sqrt{\frac{(1 - r^2)}{(n - 2)}}$$

- Testa a hipótese nula de que $t = 0$

Regressão linear simples

Objetivo de uma regressão linear

- Analisar a *relação* entre duas variáveis contínuas
  - Qual a forma desta relação?, qual o grau de relação entre as variáveis?
- Variável preditora (X) e uma variável resposta (Y) ambas contínuas
- *Predizer* um valor da variável resposta com base na variável preditora
- Modelo linear
  - Um dado valor da variável resposta é descrito por uma combinação linear de um conjunto de parâmetros (inclinação e intercepto), no qual nenhum deles aparece como um exponente ou é multiplicado por outro parâmetro

Equação da regressão linear passo-a-passo

Variável resposta = modelo + erro

Equação da regressão linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

Equação da regressão linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

Equação da regressão linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

Variável resposta

Variável preditora

Equação da regressão linear passo-a-passo

$$y_i = \alpha + \beta x_i + \varepsilon$$

intercepto

Inclinação (slope)

Exemplo

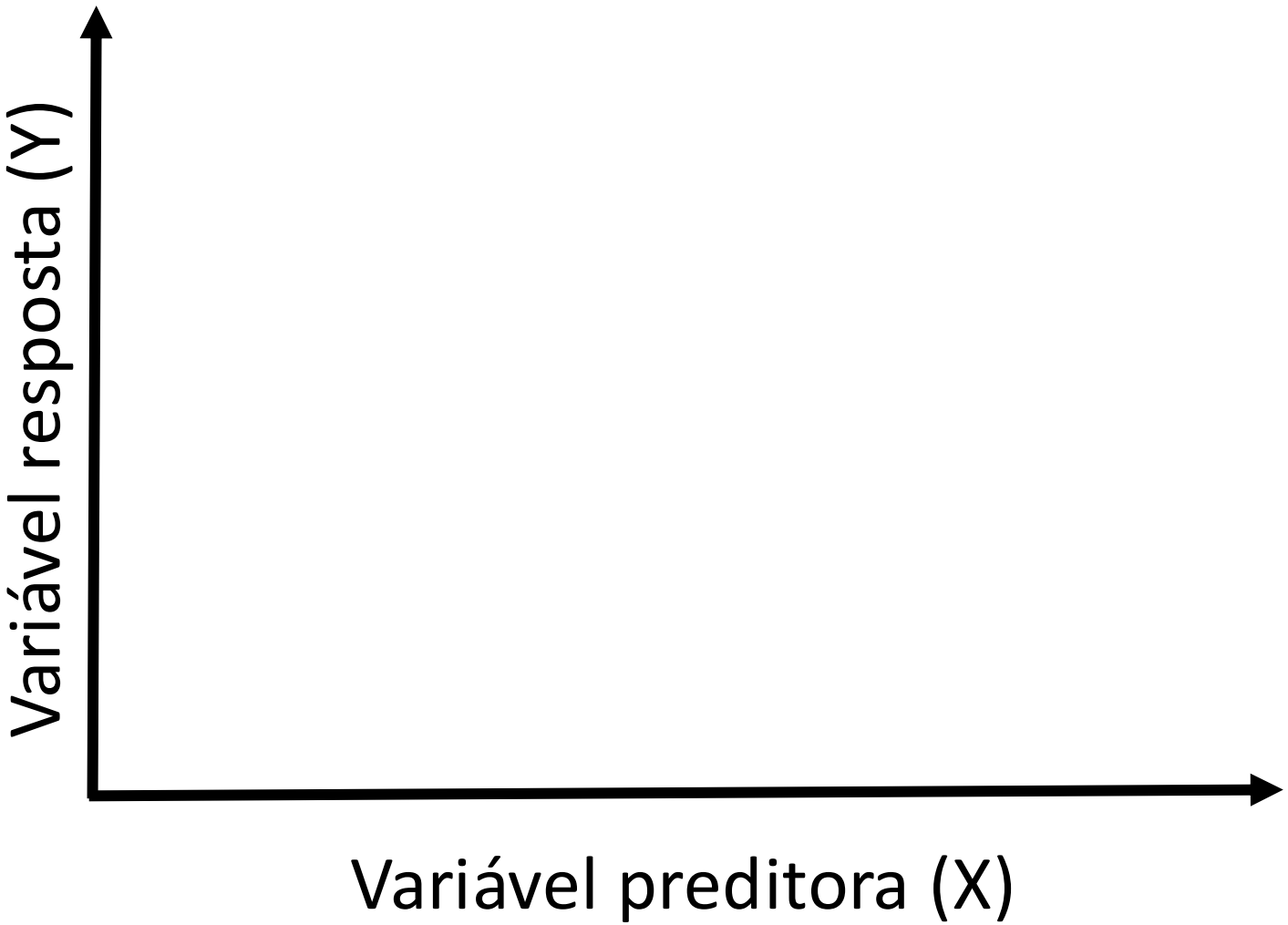
Altura dos filhos = $\alpha + \beta$*altura dos pais + resíduos

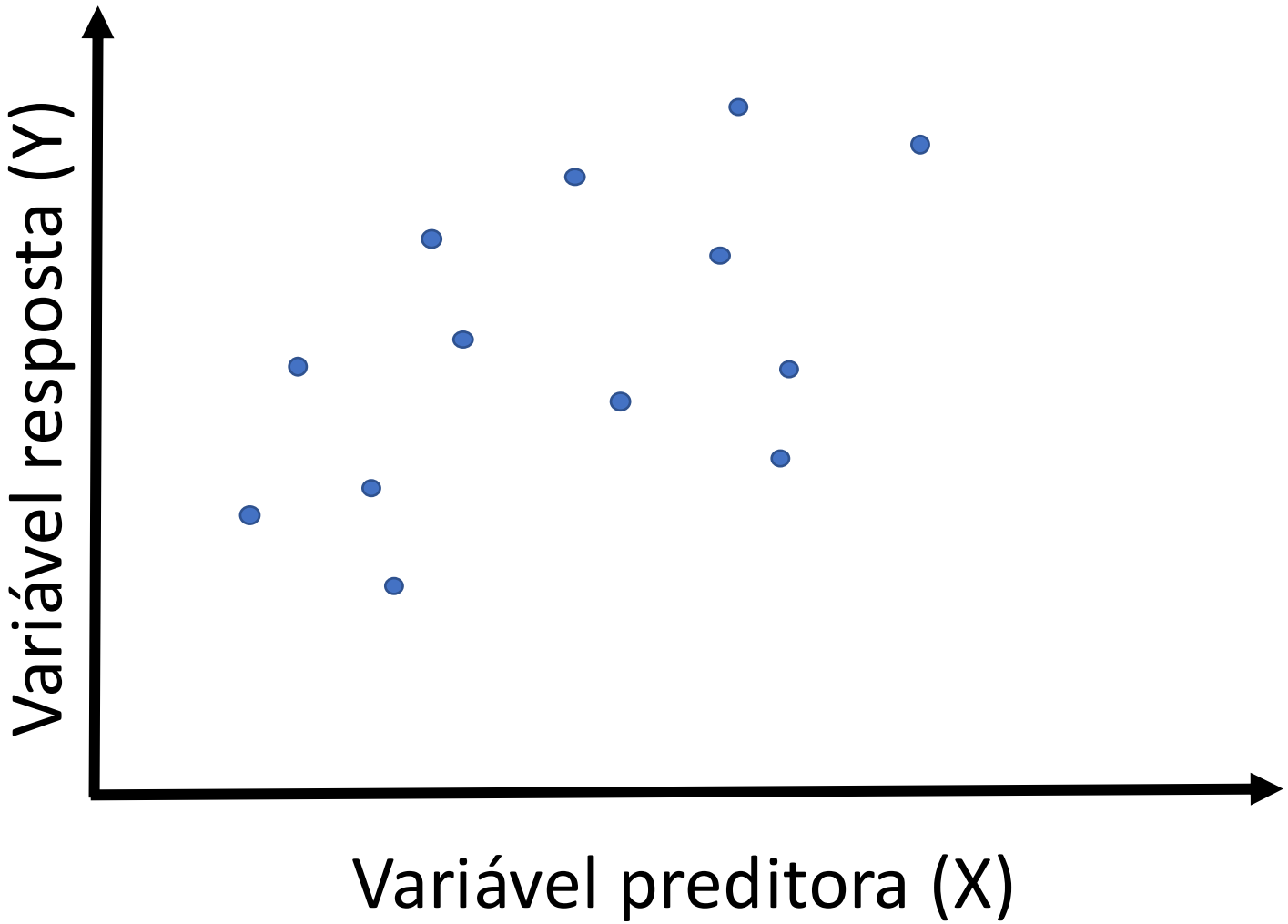
Parâmetros do modelo

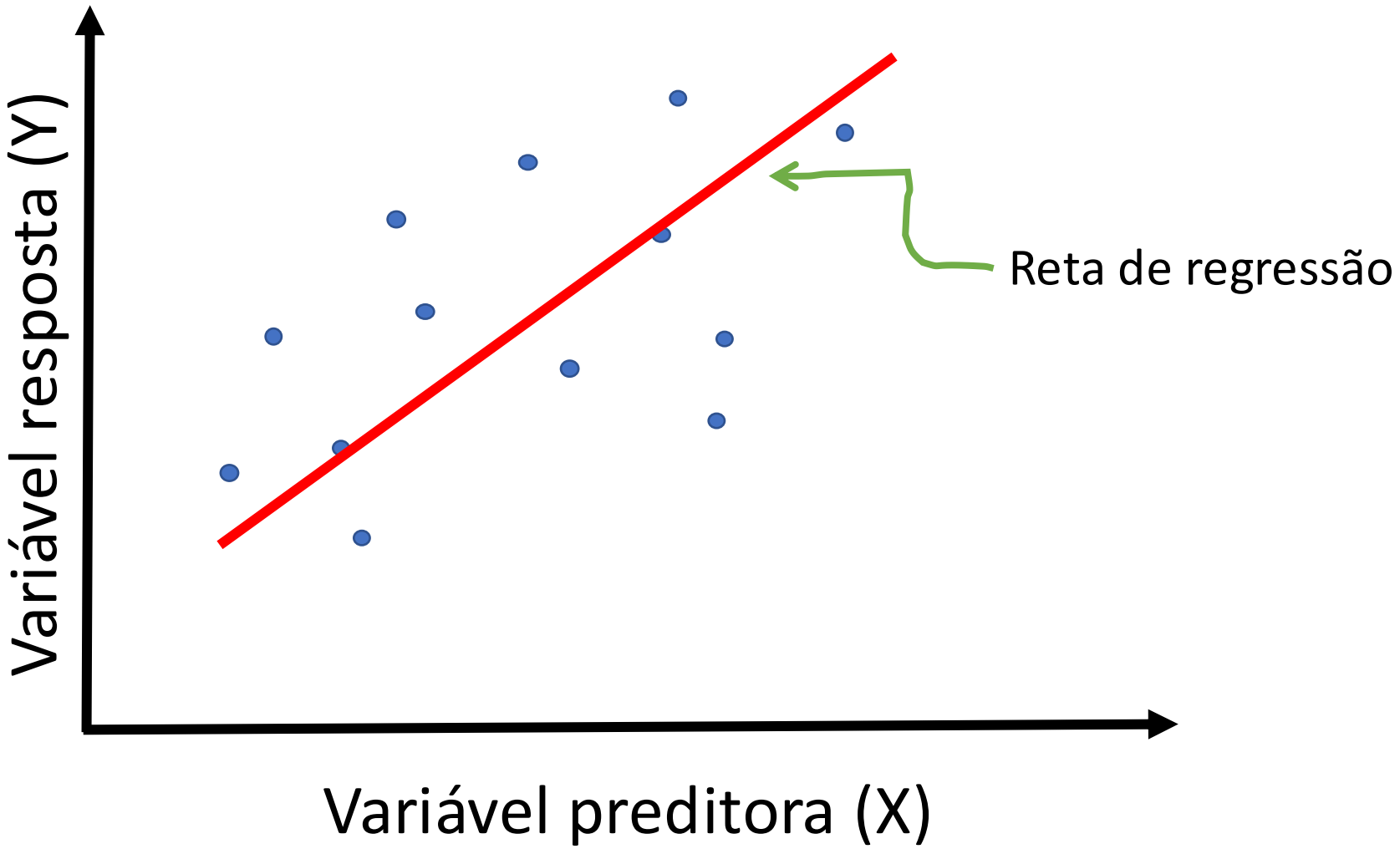
- Intercepto = valor médio da variável resposta quando a variável preditora é zero

- Inclinação (slope) = inclinação da reta predita pelo modelo. Mede o quanto a variável resposta muda em função do aumento da variável preditora.
  - Pode-se calcular o IC95% para o slope

Gráfico de dispersão


Variável resposta (Y)
Variável preditora (X)


Variável resposta (Y)
Variável preditora (X)


Variável resposta (Y)
Reta de regressão
Variável preditora (X)

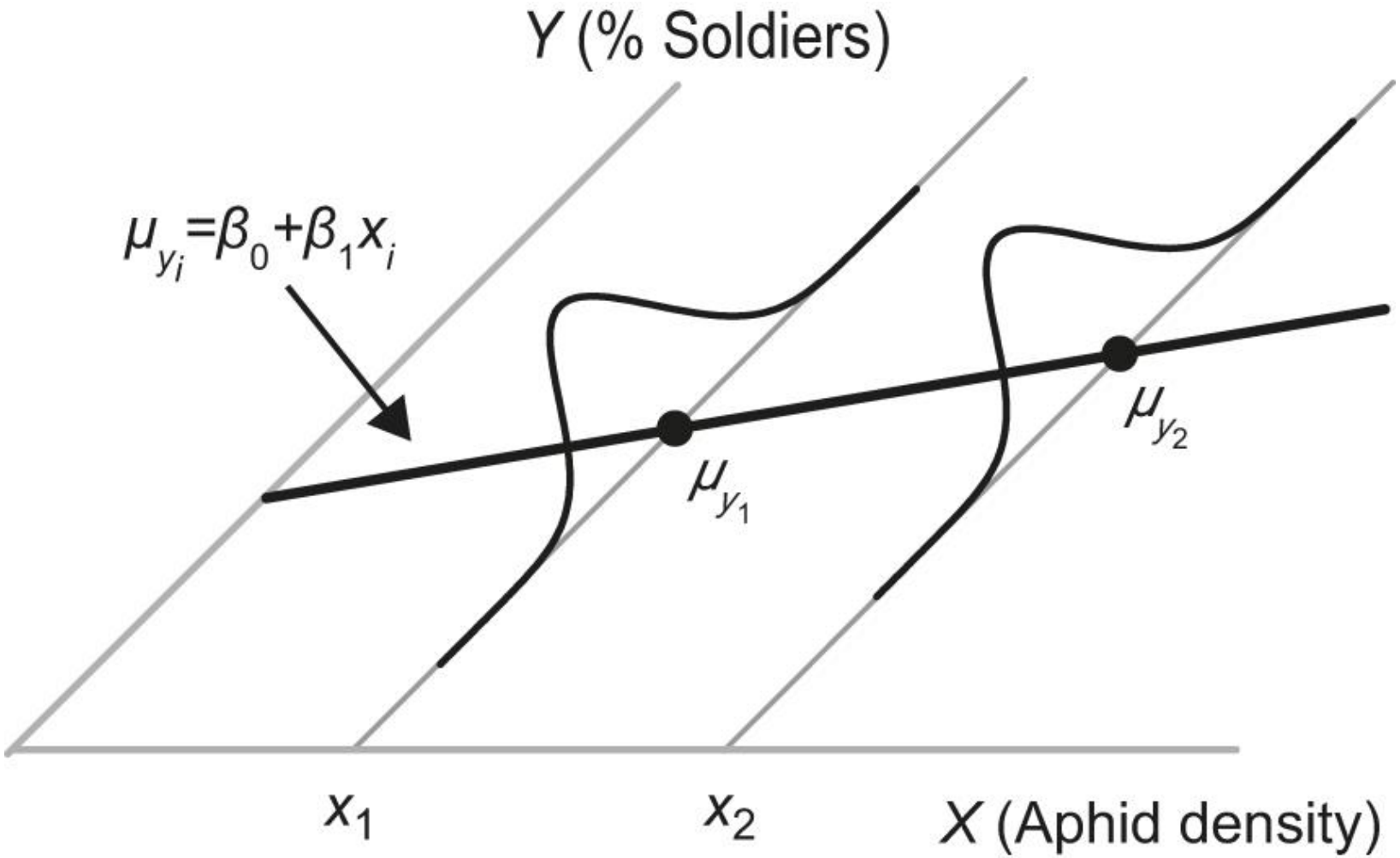


Fig. 4.1 Quinn & Keough

População de valores de Y pra cada valor de x_i tem uma distribuição normal, com a mesma variância

Será importante no cálculo dos intervalos de confiança destes parâmetros e no teste de hipótese

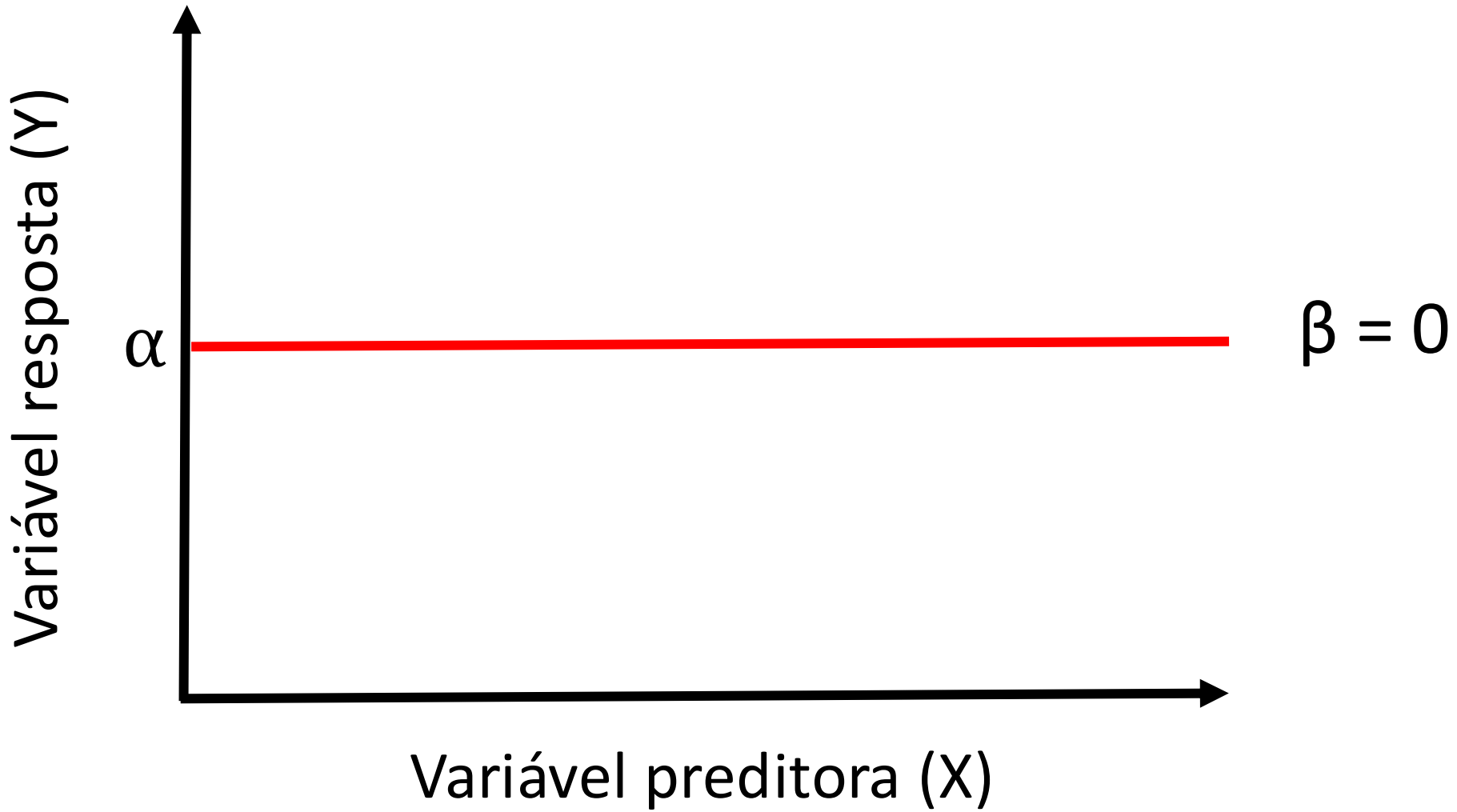
Teste de hipótese

- A hipótese nula que é testada na regressão linear é a de que o slope é diferente de zero, ou:

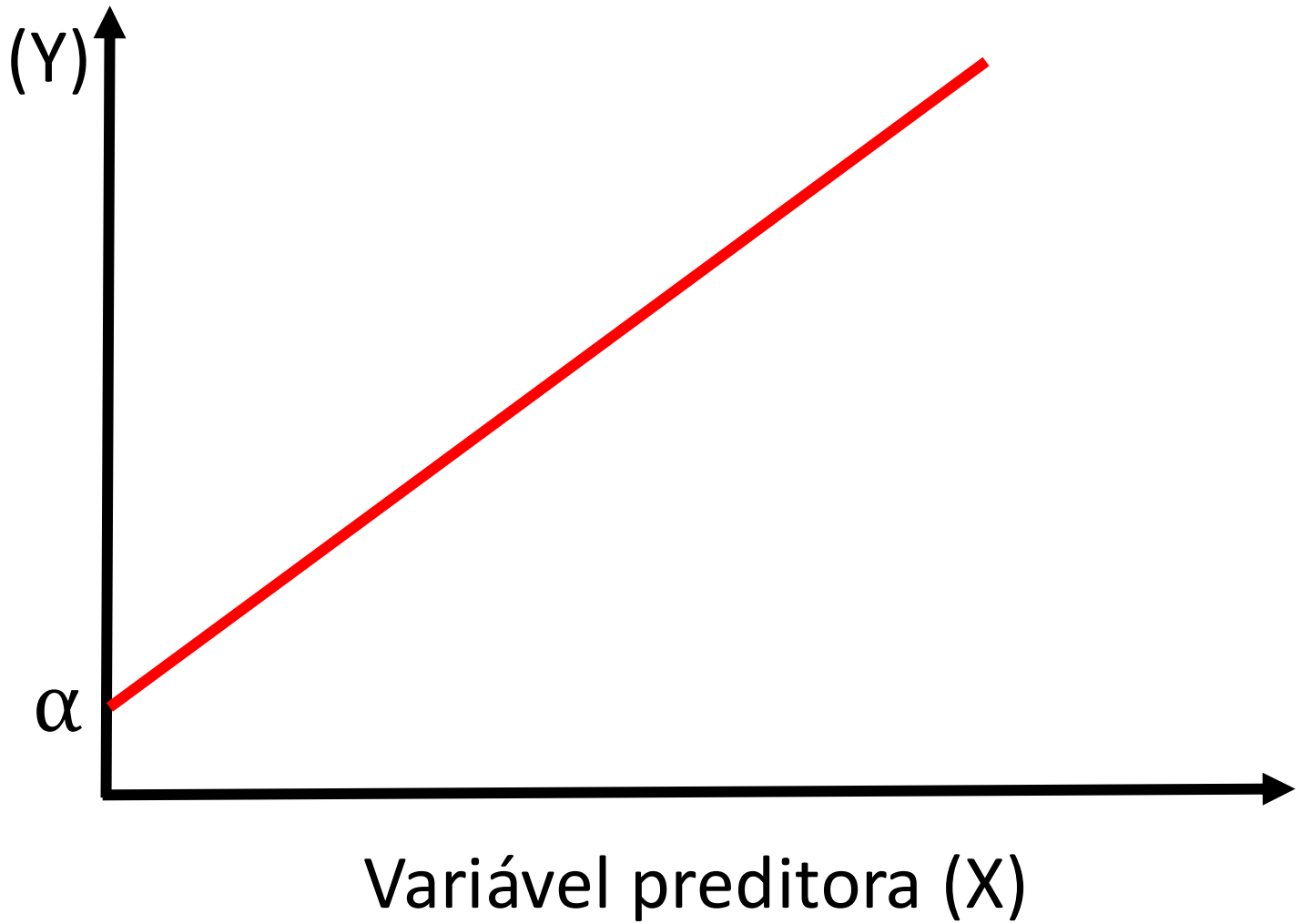
$$\beta \neq 0$$

- Mas o que diabos isso significa?
  - Se não há relação entre Y e X, a reta da regressão não deve ter inclinação, ou seja, à medida que o X muda, o valor de Y permanece o mesmo.

Hipótese nula

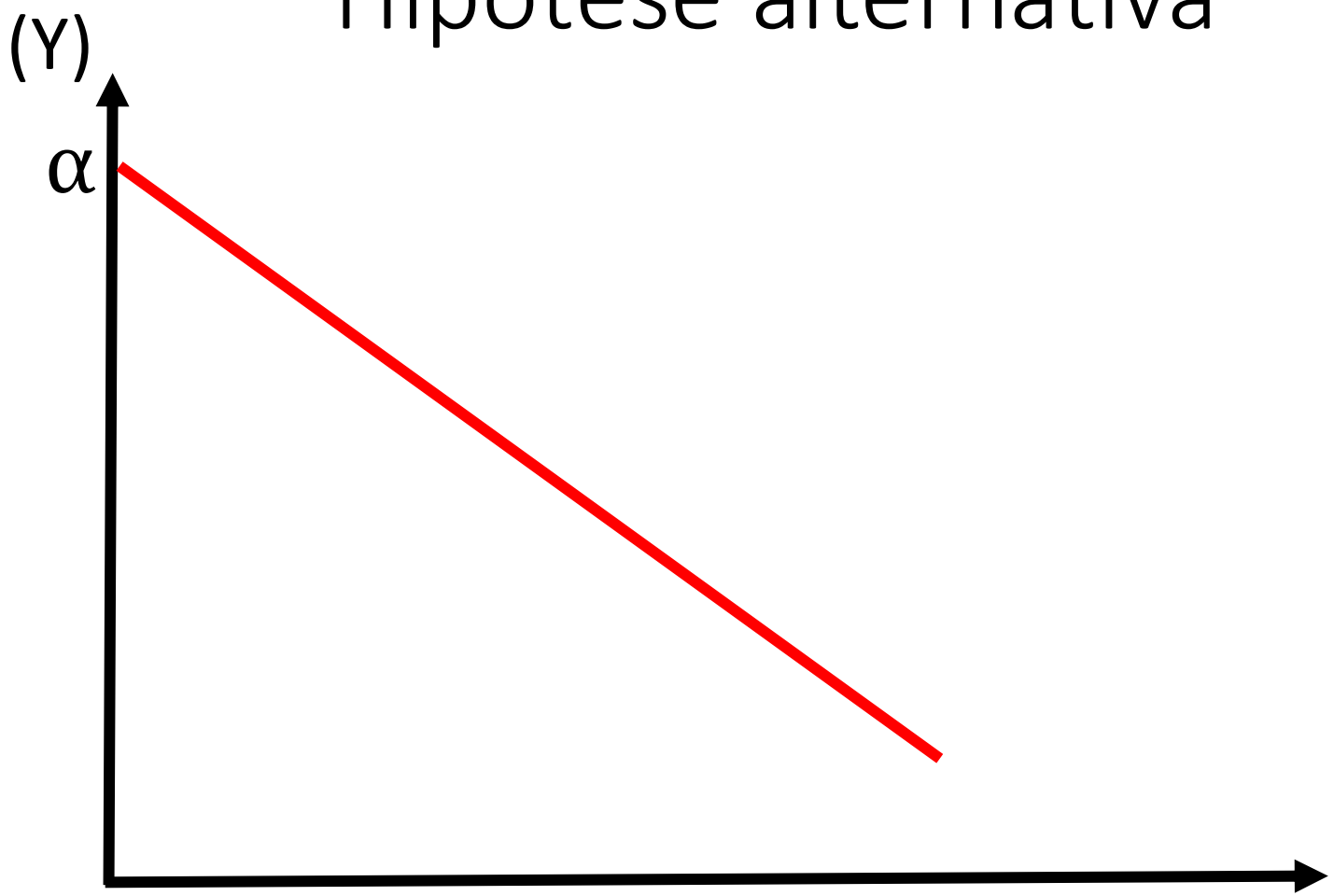


Hipótese alternativa



$\beta > 0$

Hipótese alternativa



$\beta < 0$

Como funciona o método dos mínimos quadrados

- Método para estimar os parâmetros do modelo que minimiza o desvio entre os dados observados e o predito
- Calcula a distância entre uma observação e o predito pelo modelo
- Minimiza a distância entre o ponto e a reta do modelo
- Essa distância é o resíduo

Melhor estimativa de y e x são as suas respectivas médias

É por onde a reta de regressão vai passar

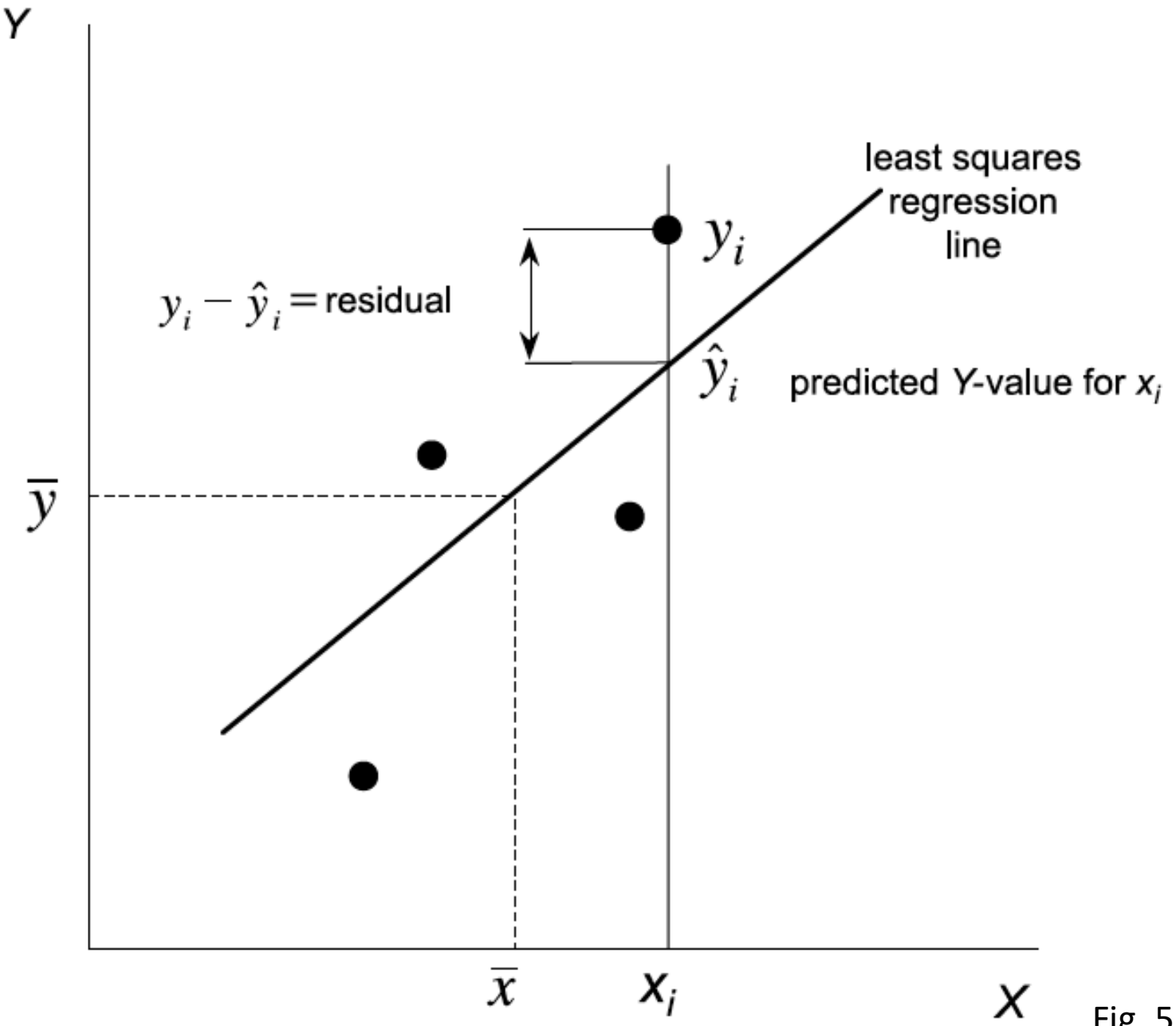
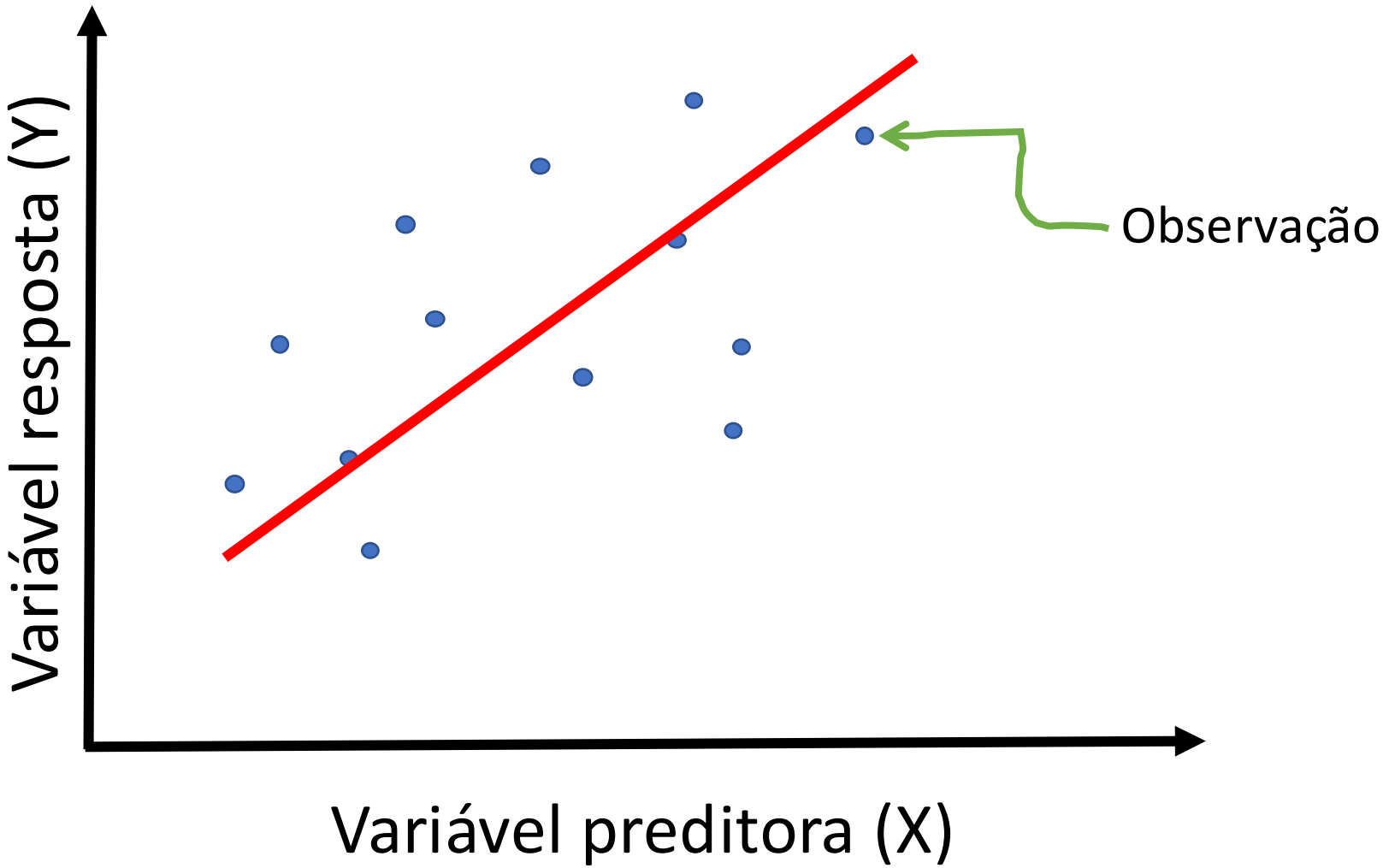
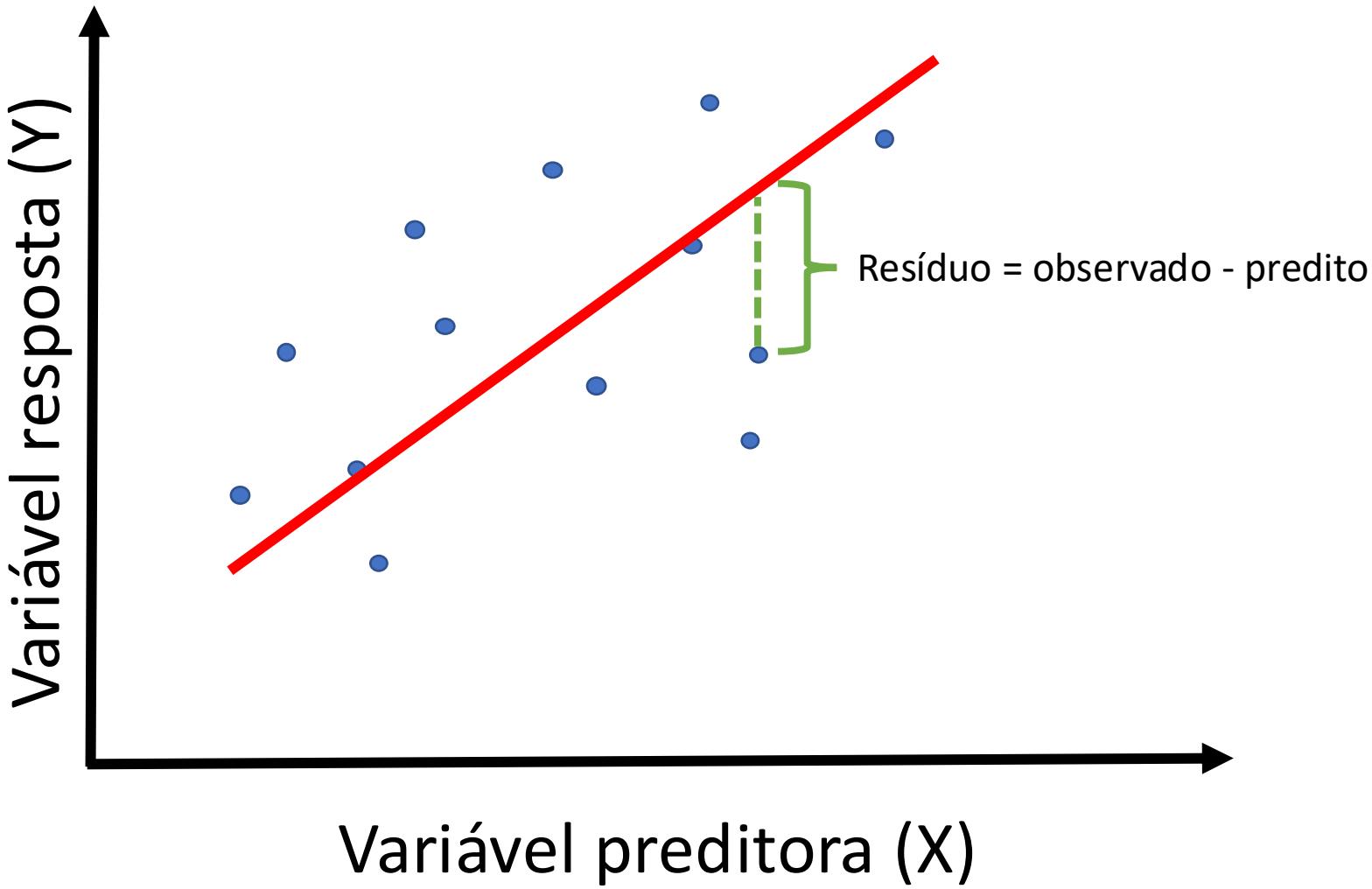
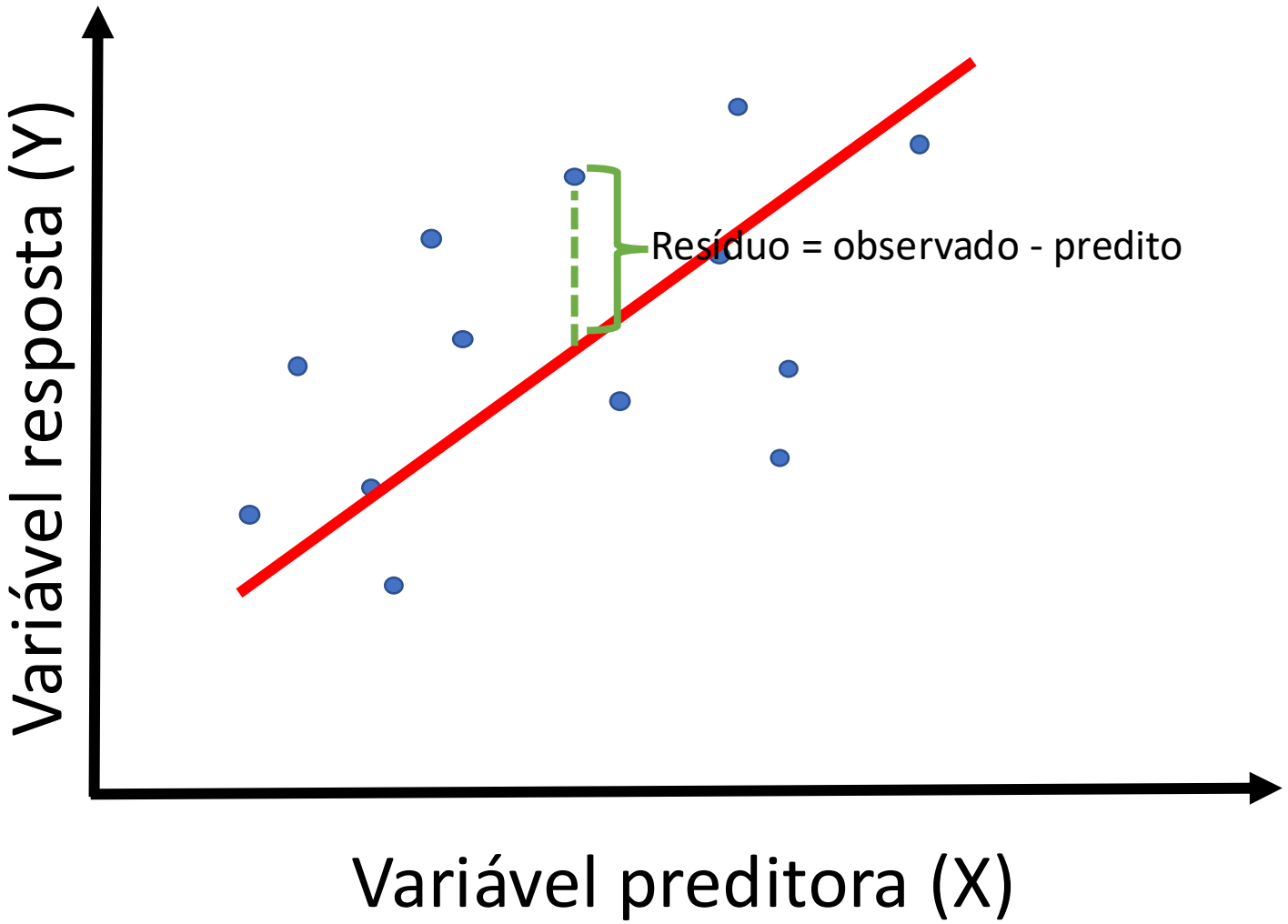


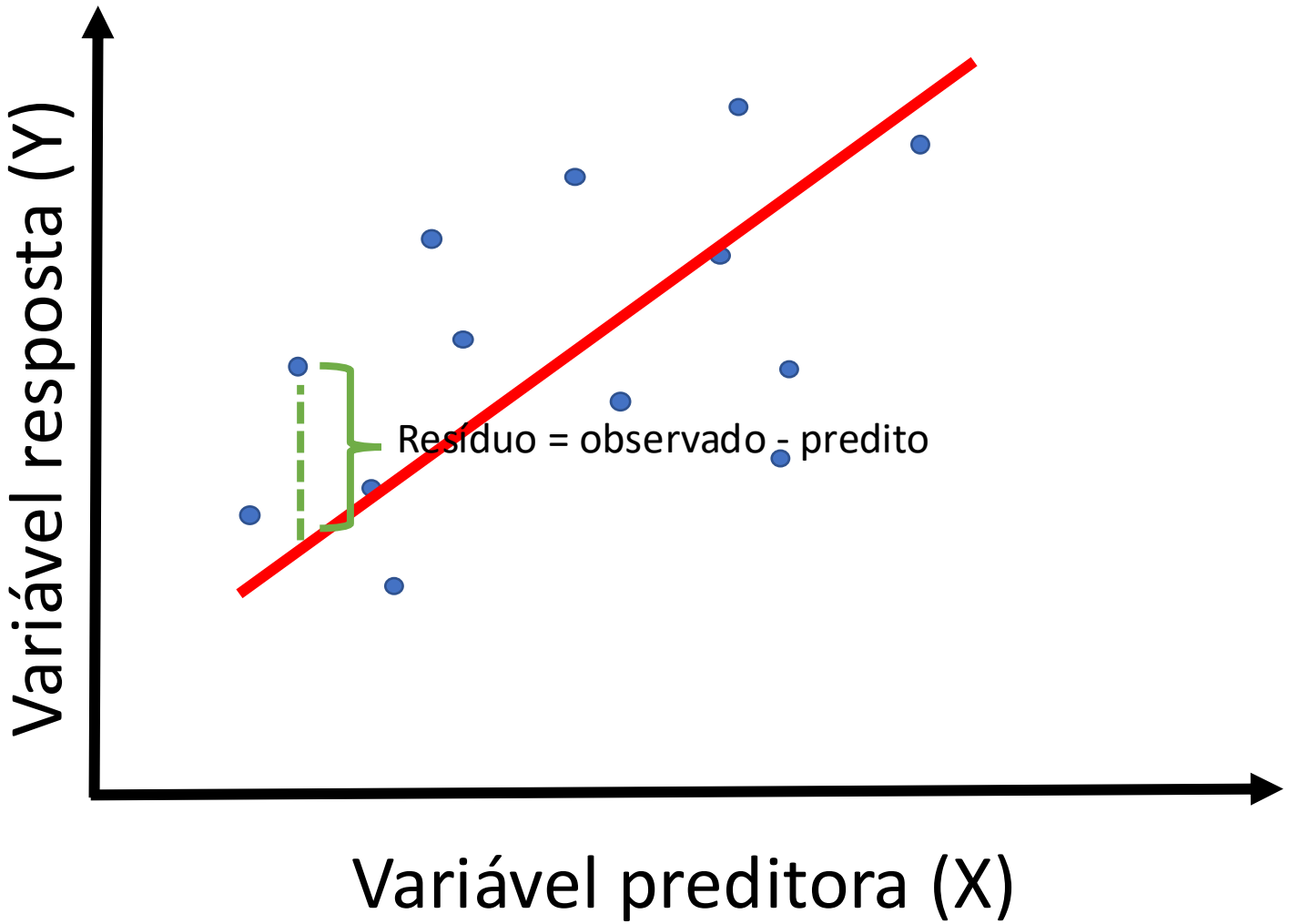
Fig. 5.6 Quinn & Keough

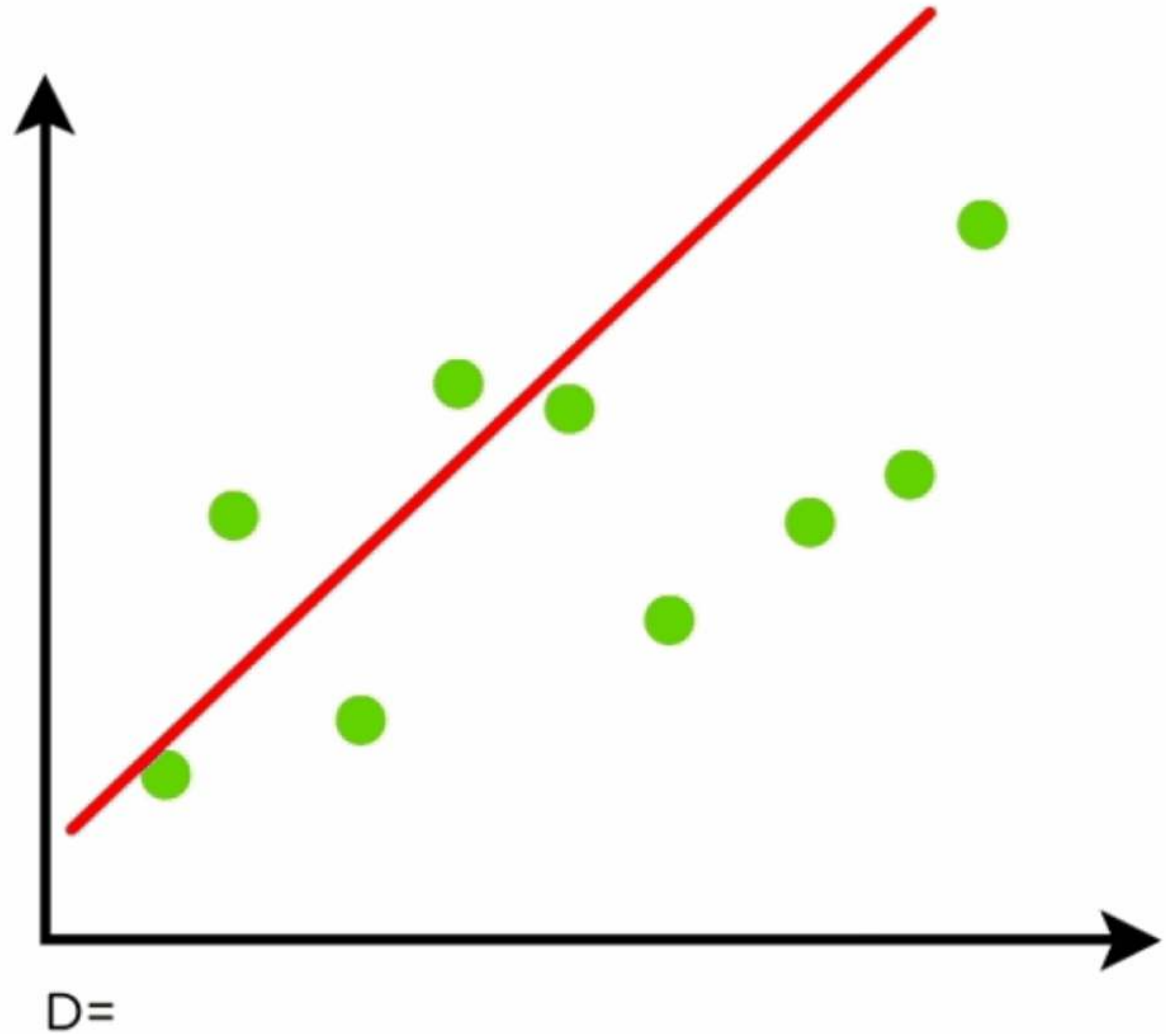

Variável resposta (Y)
Observação
Variável preditora (X)



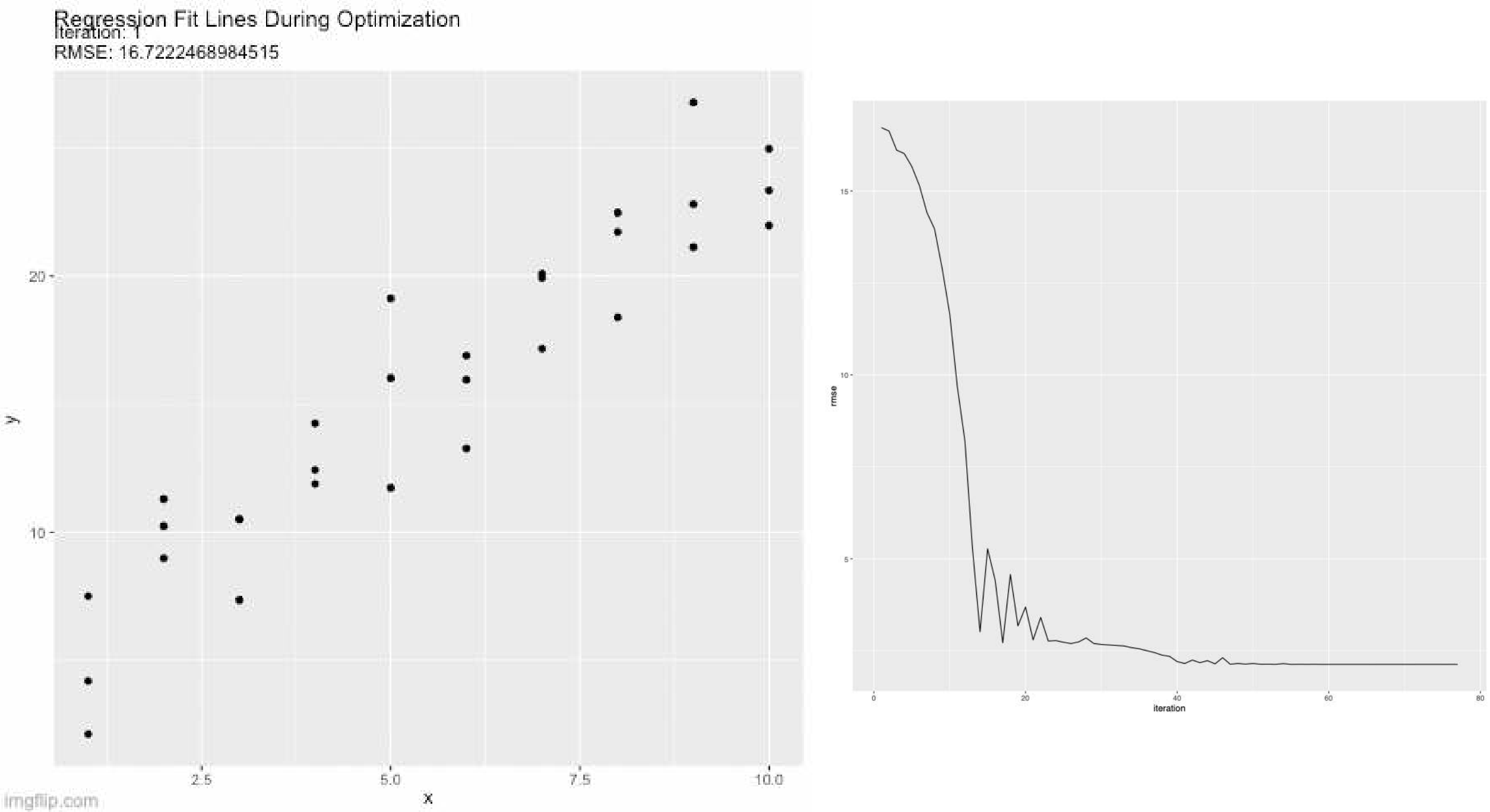
Assume que o erro de medida está na variável resposta (Y)


Variável resposta (Y)
Resíduo = observado - predito
Variável preditora (X)


Variável resposta (Y)
Resíduo = observado - predito
Variável preditora (X)



https://medium.com/@aulakhkaran5a/linear-regression-b712b88bb940


Regression Fit Lines During Optimization
Iteration: 1
RMSE: 16.7222468984515
20
10
y
2.5
5.0
7.5
10.0
x
imgflip.com
15
10
5
rmse
0
20
40
60
80
iteration

Soma de quadrados dos resíduos

Quadrado do Resíduo (cada dado)

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

Critério de otimização (estimador) para ajuste da reta de regressão (Mínimos Quadrados)

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

https://setosa.io/ev/ordinary-least-squares-regression/

Outras formas de calcular os resíduos
Soma de quadrados tipo 2 e 3

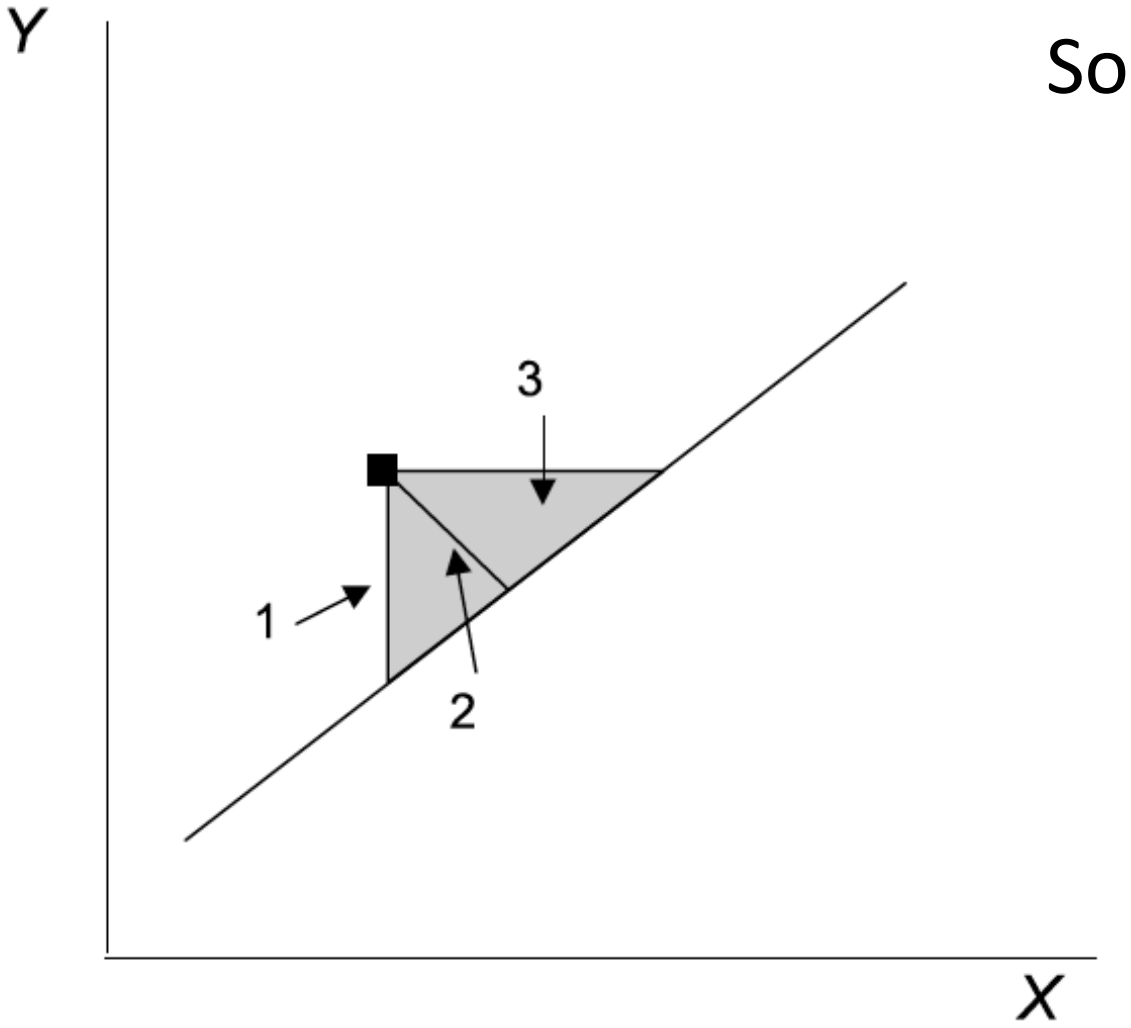


Figure 5.12 Distances or areas minimized by OLS (1), MA (2) and RMA (shaded area 3) linear regressions of *Y* on *X*.

https://stats.stackexchange.com/questions/20452/how-to-interpret-type-i-type-ii-and-type-iii-anova-and-manova

p. 100 Quinn & keough

Table 5.2 | Parameters of the linear regression model

Parameter	OLS estimate
β_1	$b_1 = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$

Diferença entre um valor de x e a sua média

*

Diferença entre um valor de y e a sua média

Diferença entre um valor de x e a sua média, ao quadrado

Table 5.2 | Parameters of the linear regression model

Parameter	OLS estimate
β_1	$b_1 = \dfrac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
β_0	$b_0 = \bar{y} - b_1\bar{x}$

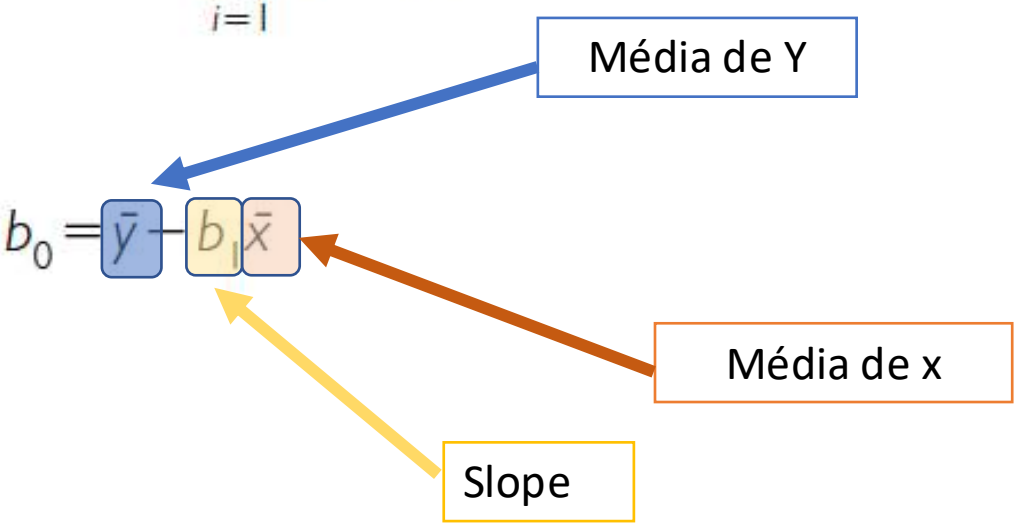


Table 5.2 | Parameters of the linear regression model

Parameter	OLS estimate
β_1	$b_1 = \dfrac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
β_0	$b_0 = \bar{y} - b_1\bar{x}$
ε_i	$e_i = y_i - \hat{y}_i$

Cada valor individual de Y

Valor predito de Y pelo modelo (reta)

Table 5.2 Parameters of the linear regression model and their OLS estimates with standard errors

Parameter	OLS estimate	Standard error
β_1	$b_1 = \dfrac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$	$s_{b_1} = \sqrt{\dfrac{MS_{Residual}}{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$
β_0	$b_0 = \bar{y} - b_1\bar{x}$	$s_{b_0} = \sqrt{MS_{Residual}\left[\dfrac{1}{n} + \dfrac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right]}$
ε_i	$e_i = y_i - \hat{y}_i$	$\sqrt{MS_{Residual}}$ (approx.)

Quinn & Keough 2002

Table 5.2 | Parameters of the linear regression model and their OLS estimates with standard errors

Parameter	OLS estimate	Standard error
β_1	$b_1 = 0.64629$	$s_{b_1} = 0.04114$
β_0	$b_0 = 68.09\text{-}0.6463*68.31 = 23.94153$	$s_{b_0} = 2.81088$
ε_i	Vetor para cada observação	2.239

Dados para o exemplo de altura dos filhos e pais de Galton

Quinn & Keough 2002

Teste de hipótese

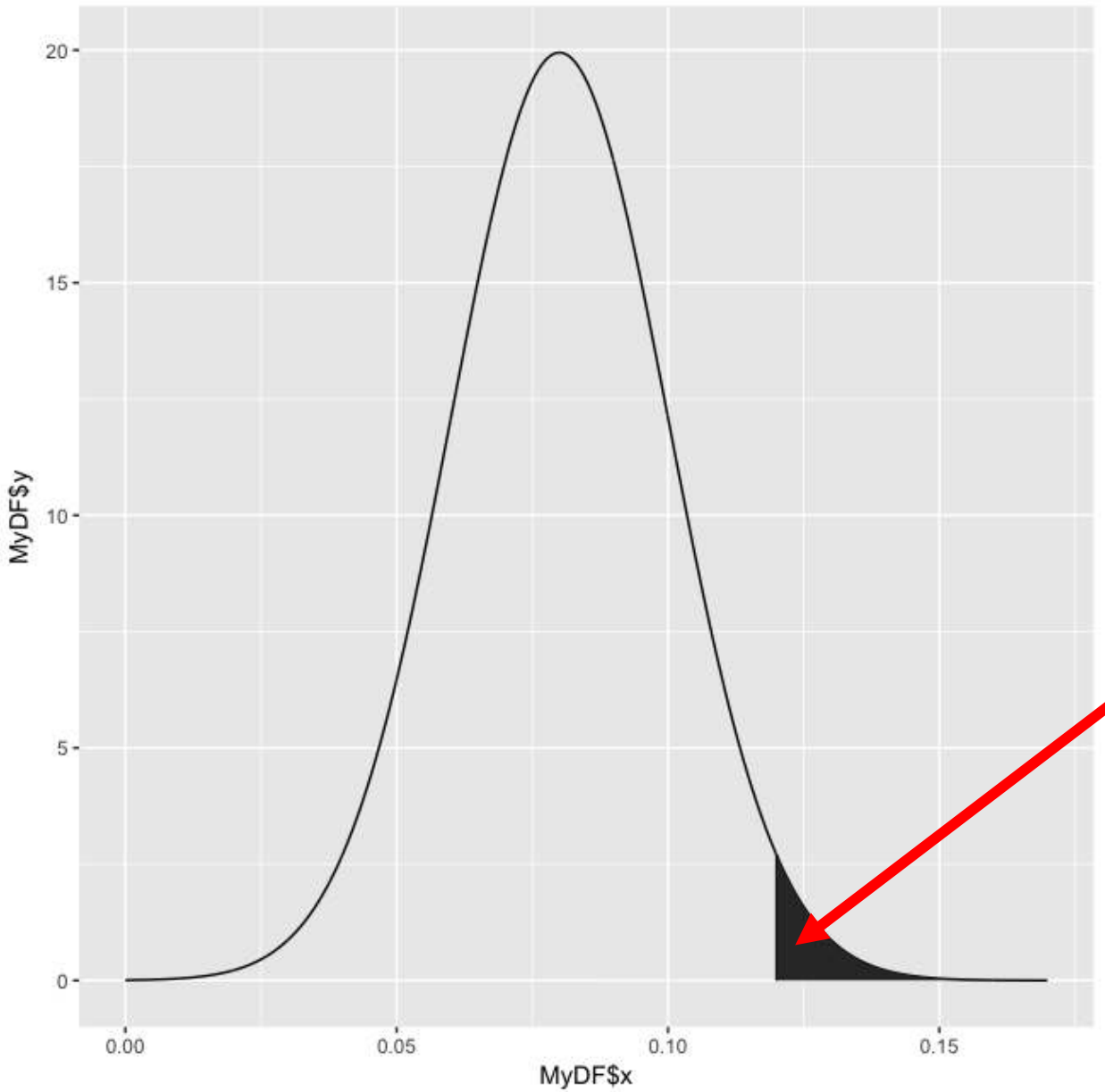
- Estatística do teste:

$$t = \sqrt{\frac{\beta - 0}{Erro\ padrão\ \beta}}$$

Teste *t* para uma amostra

Usada para testar a hipótese nula de que β = 0

Teste de hipótese



Valor para o conjunto de dados original do Galton

$t = 15.711$

Distribuição *t*

Tabela de Análise de Variância

- Particionando a variância total em componentes explicados e residuais -

Tabela de ANOVA generalizada para modelos lineares

Tabela 4.2 Quinn & Keough 2023

Source of variation	Sum-of-squares	Degrees of freedom	Mean square
Explained by linear combination of predictor(s)	$SS_{Explained}$	$df_{Explained}$	$MS_{Explained}$
Unexplained or residual or error	$SS_{Residual}$	$df_{Residual}$	$MS_{Residual}$
Total	SS_{Total}	df_{Total}	

Particionando a variância numa regressão

- Como todo modelo linear, numa regressão também é possível particionar a variância total (soma de quadrados) num componente explicado pelo(s) preditor(es) e num erro
- Chamamos isso de tabela de ANOVA
- Aqui, ao invés de testar a hipótese nula utilizando um teste *t* de um parâmetro, usamos o teste F com os mesmos graus de liberdade
  - O teste F neste caso compara um modelo utilizado a um em que só o intercepto é incluído
  - $F = t^2$
- Numa regressão, isso é feito particionando a soma de quadrados:
  - $SQ_{total} = SQ_{regressão} + SQ_{resíduos}$
  - $GL_{total} = GL_{regressão} + GL_{resíduos}$

Particionando a variância numa regressão

- No entanto, a SQ é dependente do número de observações, e aumenta com o aumento das observações
- Logo, precisamos de uma medida de variabilidade *independente* do número de observações
  - Este é o Quadrado Médio (Mean squares)
- Porém, o Quadrado Médio (QM) não tem a mesma propriedade aditiva, como a SQ
  - $QM_{regressão} + QM_{resíduos} \neq QM_{total}$
- O $QM_{resíduos}$ estima a variância do erro ($\boldsymbol{\sigma}^2_{\varepsilon}$)

Table 5.3 Analysis of variance (ANOVA) table for simple linear regression of Y on X

Source of variation	SS	df	MS	Expected mean square
Regression	$\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2$	1	$\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{1}$	$\sigma_\varepsilon^2+\beta_1^2\sum_{i=1}^{n}(x_i-\bar{x})^2$
Residual	$\sum_{i=1}^{n}(y_i-\hat{y}_i)^2$	$n-2$	$\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}$	σ_ε^2
Total	$\sum_{i=1}^{n}(y_i-\bar{y})^2$	$n-1$		

Table 5.3 Analysis of variance (ANOVA) table for simple linear regression of Y on X

Source of variation	SS	df	MS	Expected mean square
Regression	1236.9	927	1236.93	$\sigma_\varepsilon^2 + \beta_1^2 \sum_{i=1}^{n} (x_i - \bar{x})^2$
Residual	4640.3	926	5.01	5.005
Total	5877.2	926		

Dados para o exemplo de altura dos filhos e pais de Galton

Coeficiente de determinação

$$r^2 = \frac{SS_{Regression}}{SS_{Total}} = 1 - \frac{SS_{Residual}}{SS_{Total}}$$

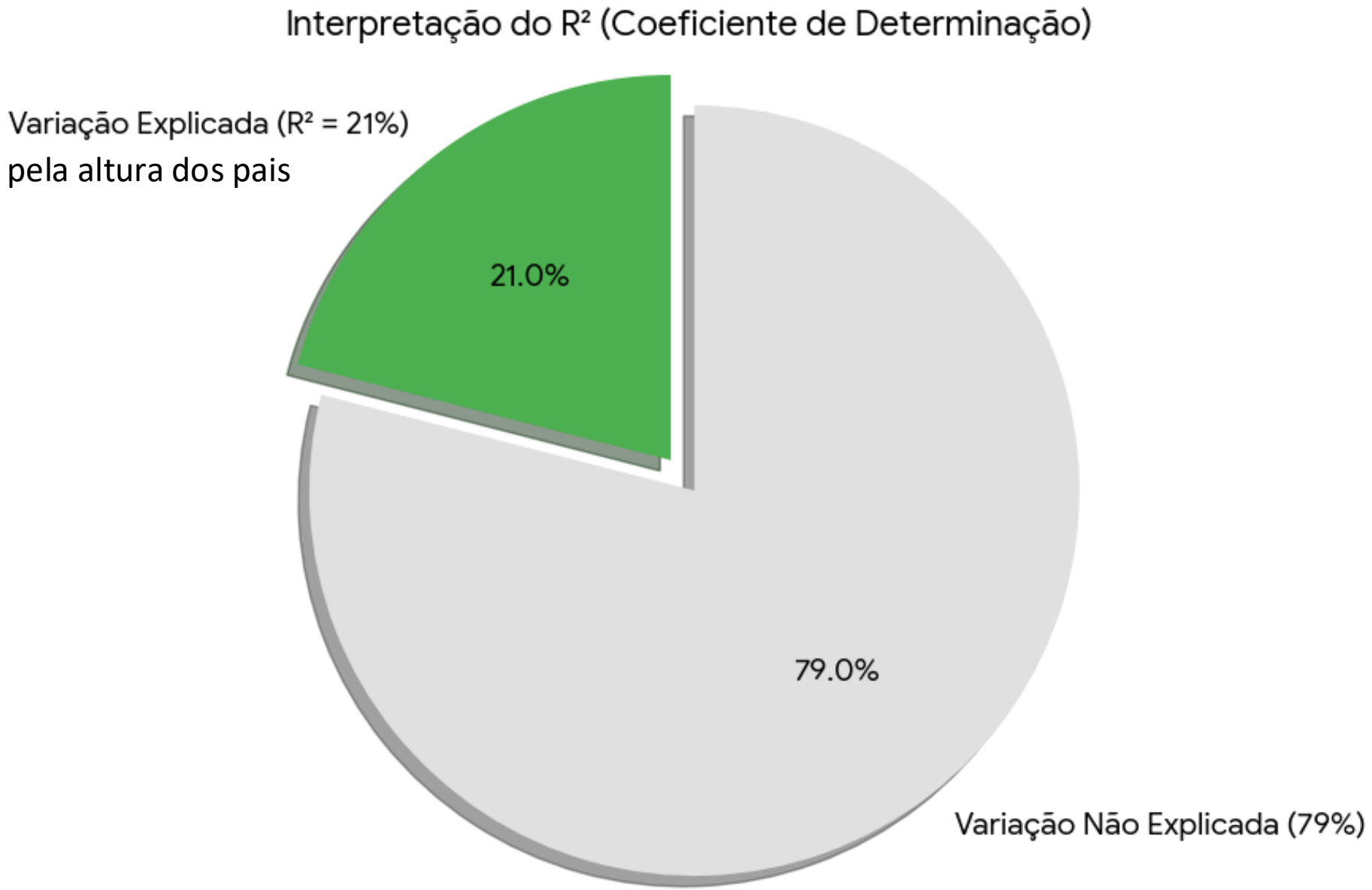
$0 < r^2 > 1$

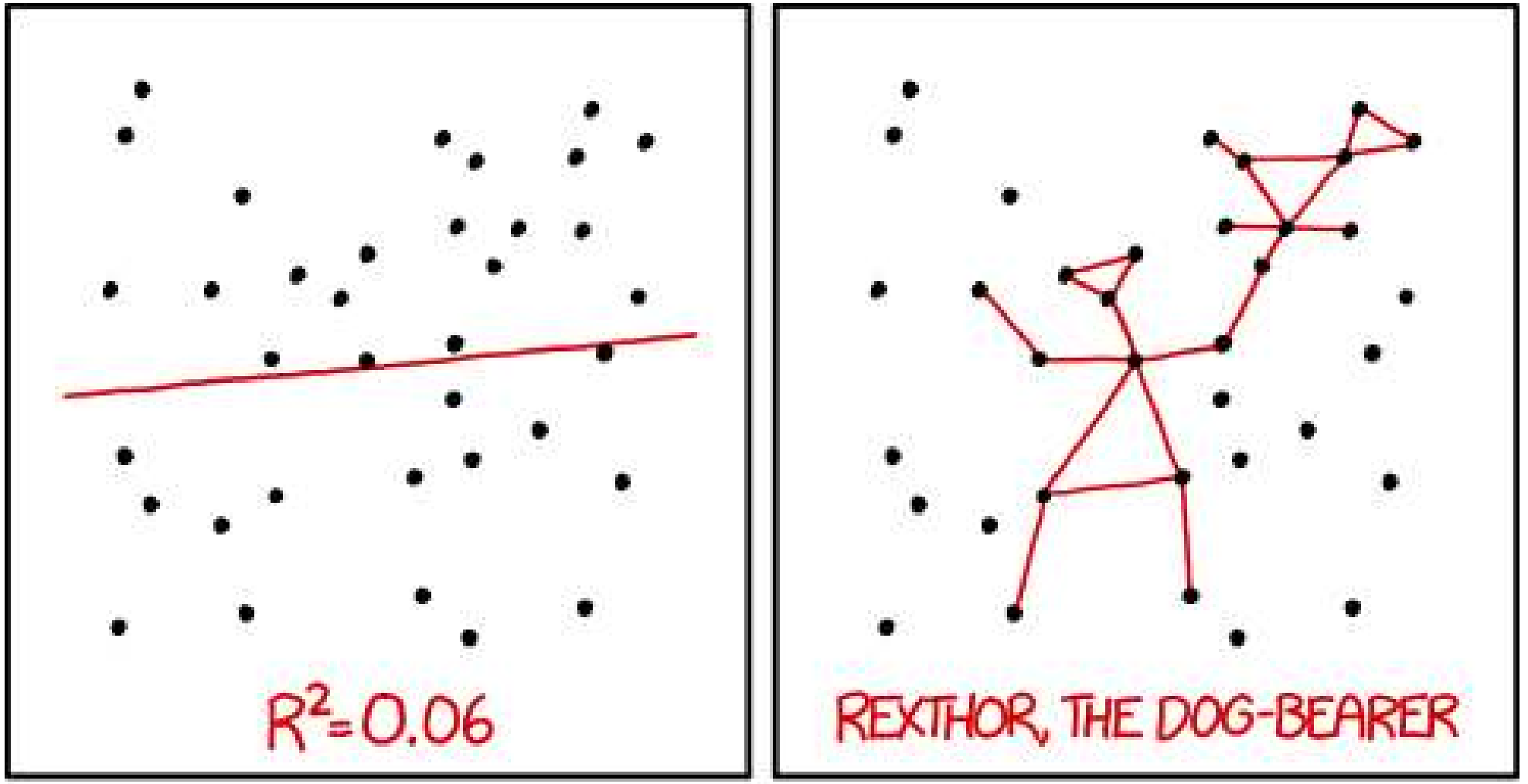
1236.9/5877.2
= 0.2105

1-4640.3/5877.2
= 0.2105

r^2 = *r* Pearson ao quadrado

A variação na altura dos pais explicou cerca de 21% da variação da altura dos filhos. Os outros 79% são devidos a outros fatores (nutrição, aleatoriedade, genética não mendeliana simples, etc.) que não medimos


Interpretação do R² (Coeficiente de Determinação)
Variação Explicada (R² = 21%)
pela altura dos pais
21.0%
79.0%
Variação Não Explicada (79%)



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

https://www.explainxkcd.com/wiki/index.php/1725:_Linear_Regression

Pressupostos dos modelos lineares

- Variâncias homogêneas dos resíduos
- Relação entre X e Y é linear
- Variável preditora (X) é medida sem erros (Soma de Quadrados tipo I)

- Resíduos independentes e identicamente distribuídos (i.i.d)
  - Seguem uma distribuição normal
  - $E[e] = 0$ e $Var[e] = \boldsymbol{\sigma}^2$

- Independência das unidades amostrais
  - Pseudoréplicas, lembram da última aula?

p 275-7 Gotelli & Ellison
Pag. 360-1 Zar
Pág 92-3 Quinn & Keough

Diagnósticos dos resíduos

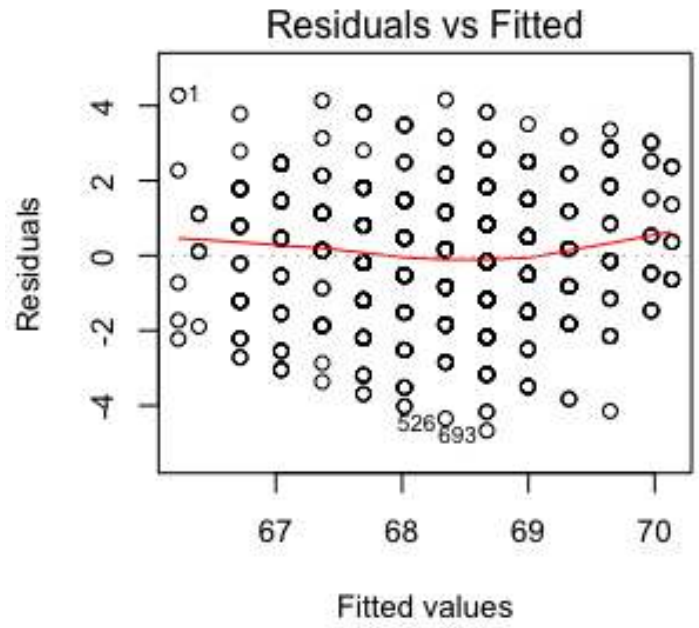
P. 94-6 Quinn & Keough

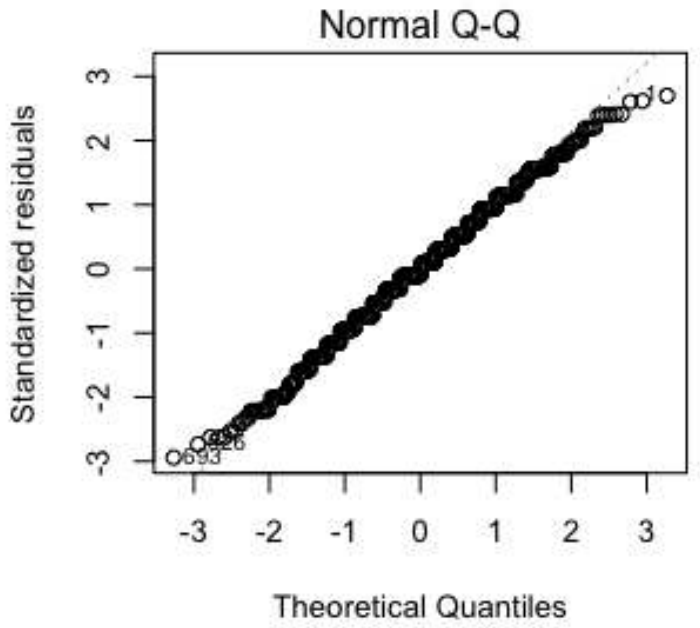
Homogeneidade de variância

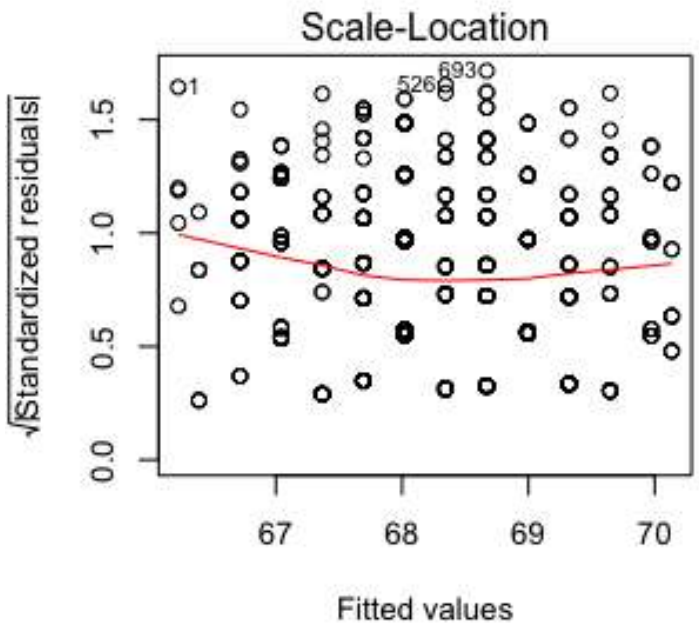
Normalidade

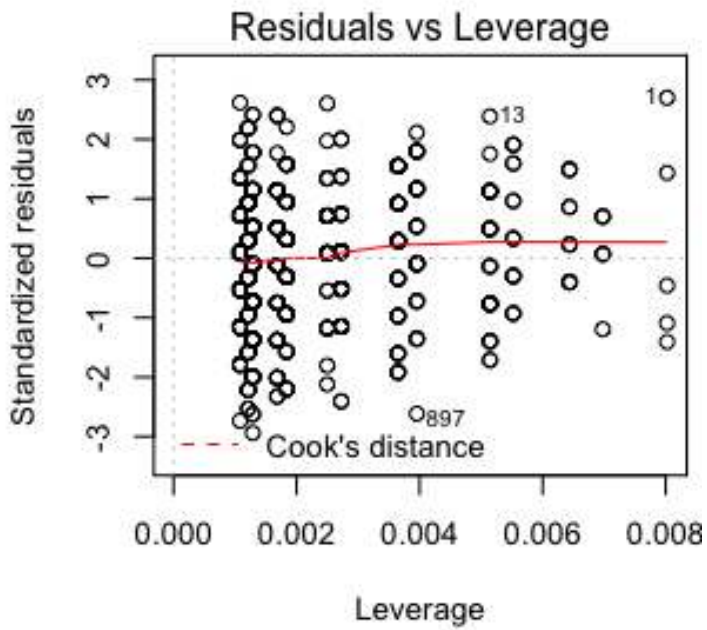
Homogeneidade de variância (com resíduos padronizados)

Valores extremos (ouliers)



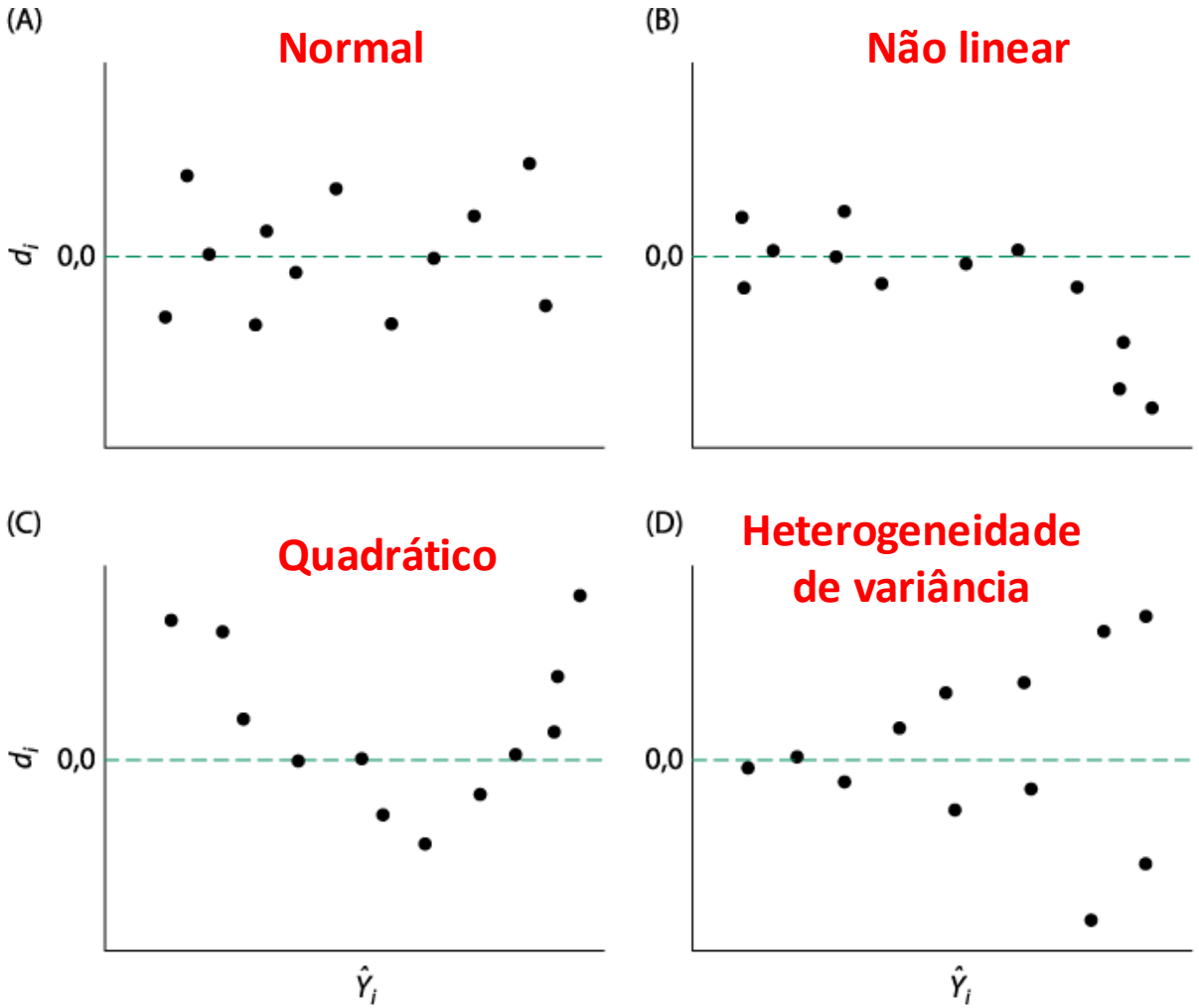






No R:
par(mfrow=c(2,2))
plot(modelo)

Dados para o exemplo de altura dos filhos e pais de Galton



Gotelli & Ellison

Figura 9.5 Padrões hipotéticos para gráficos de diagnóstico de resíduos (d_i), *versus* os valores ajustados ($\hat{Y}_i$) em regressão linear. (A) Distribuição esperada dos resíduos para um modelo linear com distribuição normal dos erros. Se os dados são bem ajustados pelo modelo linear, este é o padrão que deve ser encontrado nos resíduos. (B) Resíduos para um ajuste não linear. Neste caso, o modelo sistematicamente superestima os valores reais do *Y* conforme o *X* aumenta. Uma transformação matemática (p. ex., logaritmo, raiz quadrada ou inverso) pode produzir uma relação mais linear. (C) Resíduos para uma relação quadrática ou polinomial. Neste caso, os resíduos positivos grandes ocorrem para valores da variável *X*, que são muito pequenos, e para valores muito grandes. Uma transformação polinomial da variável *X* (X^2 ou alguma potência maior de *X*) pode produzir um ajuste linear. (D) Resíduos com heterocedasticidade (variância aumentando). Neste caso, os resíduos não são consistentemente negativos nem positivos, indicando que o ajuste do modelo é linear. Contudo, o tamanho médio dos resíduos aumenta com o *X* (heterocedasticidade), sugerindo que erros de medidas podem ser proporcionais ao tamanho da variável *X*. Uma transformação logarítmica ou da raiz quadrada pode corrigir esse problema. As transformações não são uma panaceia para as análises de regressão e nem sempre resultam em relações lineares.

Outras formas de regressão

- Não-linear simples (relação entre Y e X dado por um parâmetro elevado a alguma potência)
  - Quadrática
  - Polinomial
  - Exponencial
- Linear múltipla ($Y \sim X_1 + X_2 + \ldots X_n$)
- Logística
- (semi)Parcial => particiona a variação em termos dos preditores isolados e em conjunto
- Ponderada (weigthed regression)
- Robusta
- De quantis (ajusta uma reta pra cada parte dos dados)
- Step-wise regression => utiliza um critério de seleção para incluir/remover preditores
- Árvore de regressão (regression tree)
- Análise de rotas (Path analysis) => diagrama de relação entre as variáveis
- Partição hierarquica

Regressão múltipla

- Relaciona vários preditores à variável resposta
- Coeficientes de regressão parcial (padronizados) => influência de cada preditor na resposta
- Problema com multicolinearidade
  - Como diagnosticar? Variation Inflation Factor (VIF)

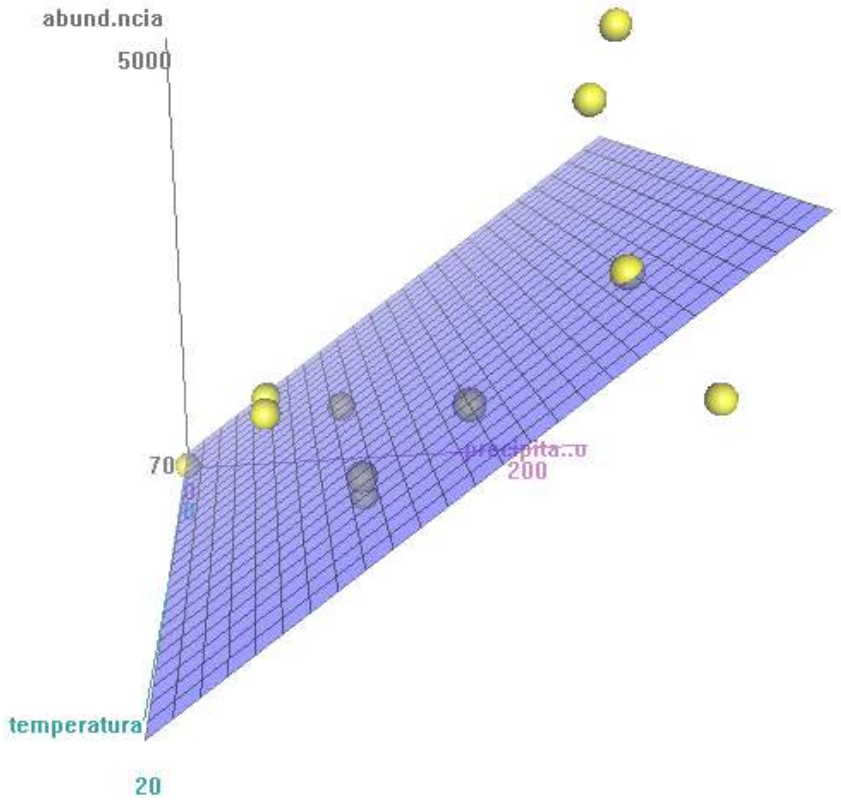
R^2 ajustado

$$1 - \frac{SS_{Residual}/[n-(p+1)]}{SS_{Total}/(n-1)}$$

- Modelo fica mais complicado à medida que se adiciona preditores
- Normalmente se usa o R^2 ajustado, que leva em conta o número de variáveis e unidades amostrais

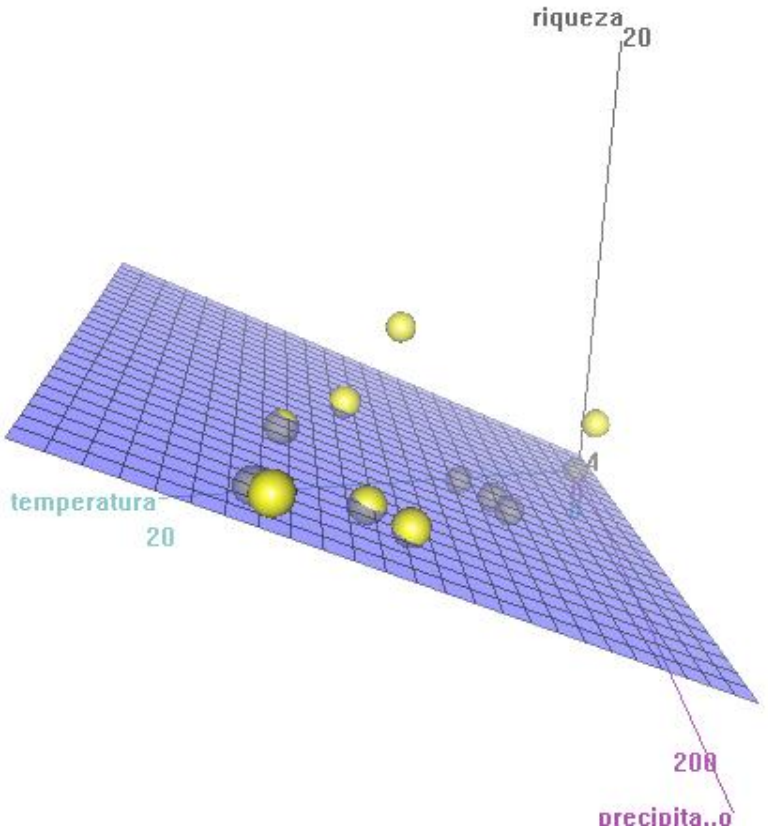
Resultados

Abundância



(r^2=0.516; $F_{(2,9)}$=4.811;
P=0.07

Riqueza



(r^2=0.823; $F_{(2,9)}$=21.05;
P=0.048

Temperatura do ar afetou a riqueza

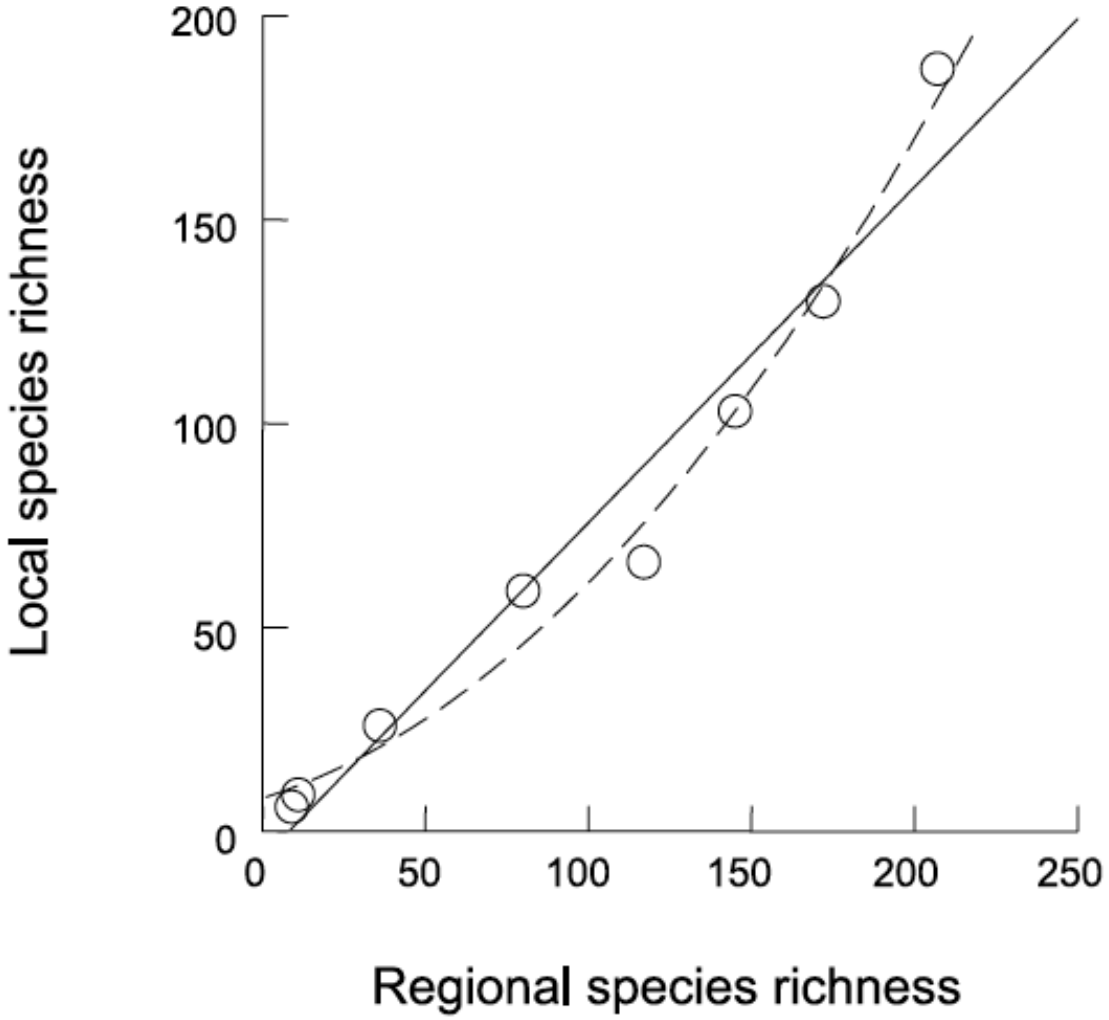


Figure 6.5 Scatterplot of local species richness against regional species richness for 10% of regions sampled in North America for a range of taxa (Caley & Schluter 1997) showing linear (solid line) and second-order polynomial (quadratic; dashed line) regression functions.

Regressão polinomial

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1}^2 + \varepsilon_i$$

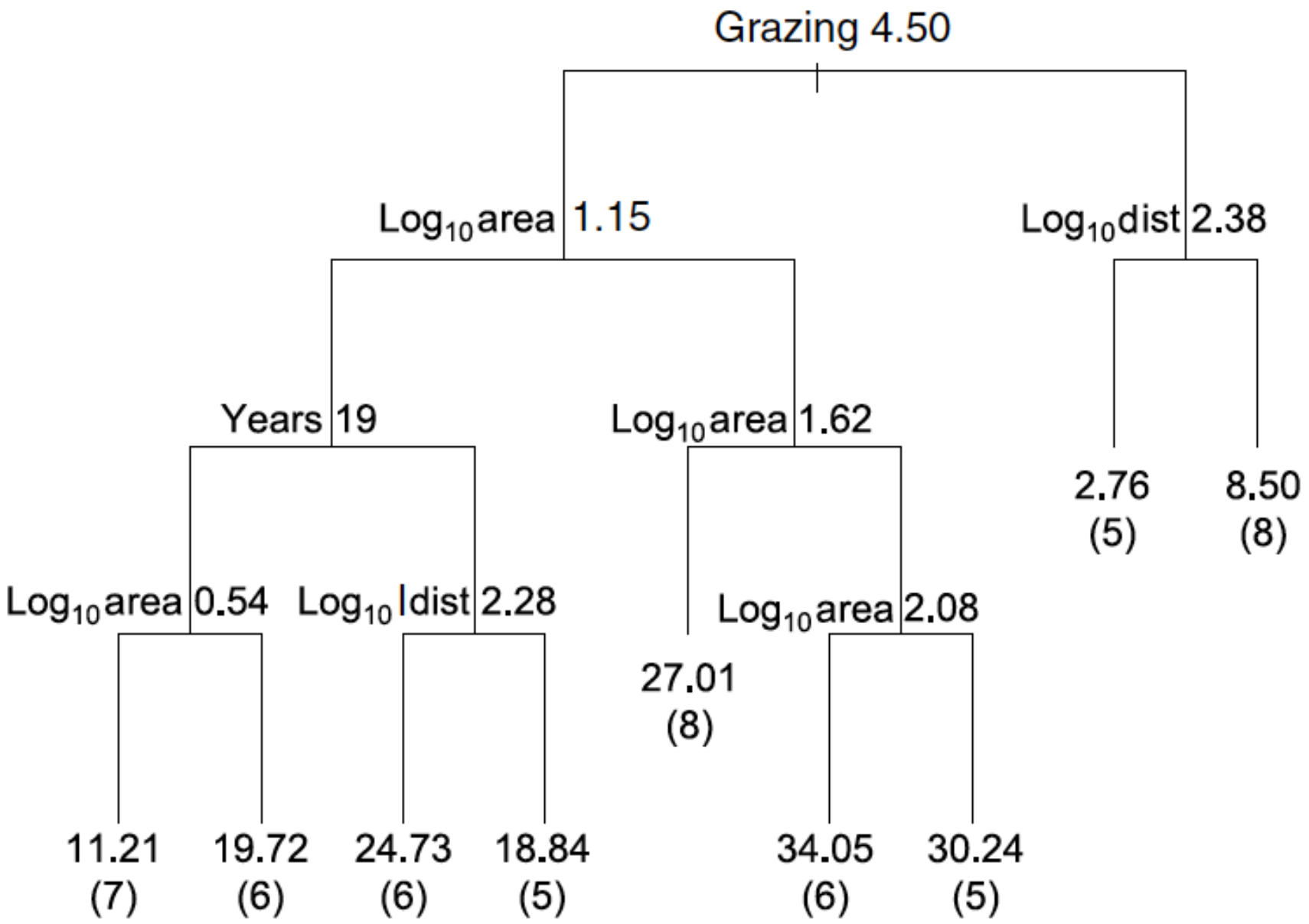


Figure 6.7 Regression tree based on the same data as in Figure 6.6, except that branching was continued to a lower level.

Árvore de regressão
- Uma boa alternativa para regressão múltipla

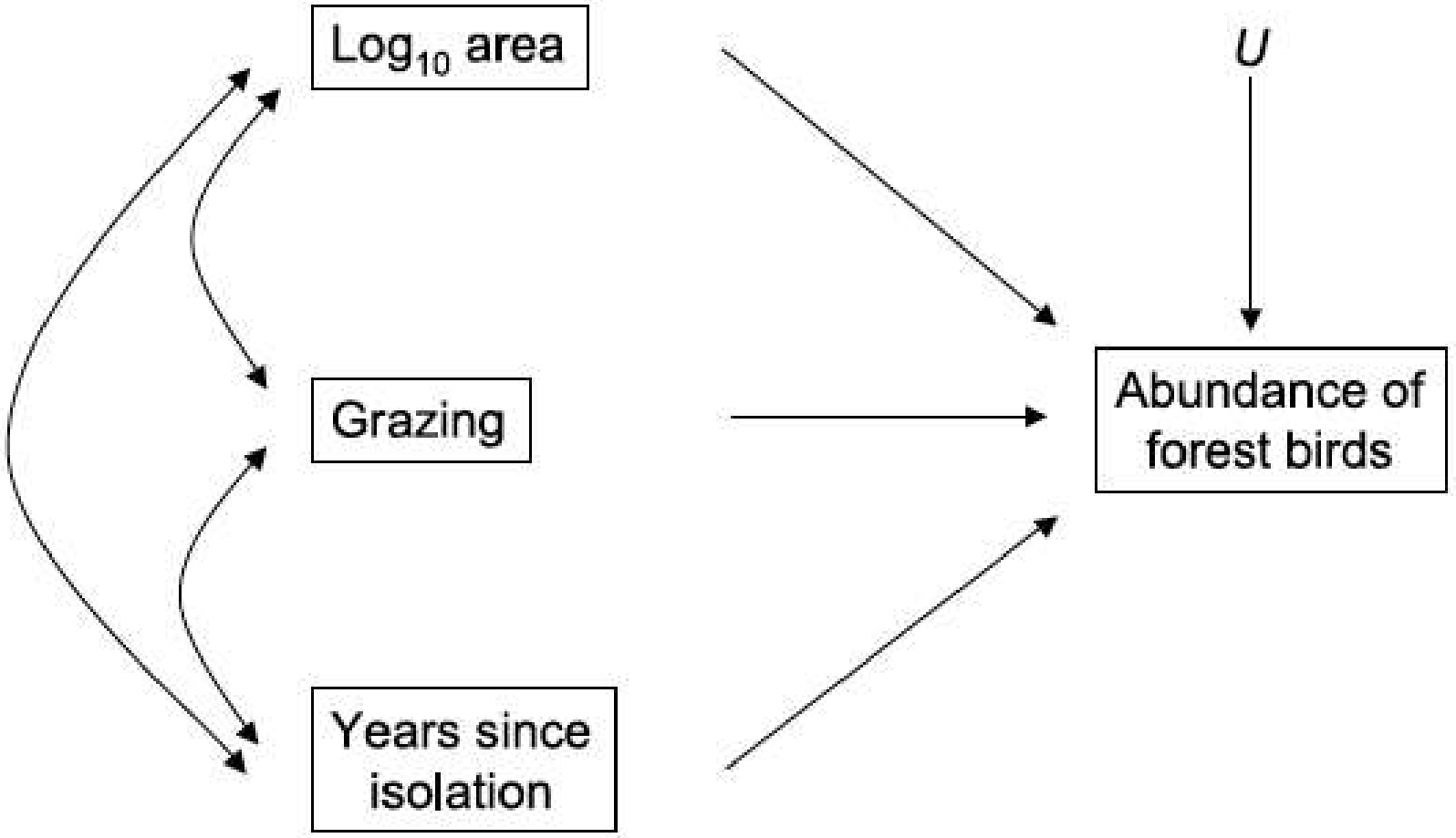


Figure 6.9 Path diagram for simple multiple regression model relating three predictor variables (log_{10} patch area, grazing, years since isolation) to one response variable (abundance of forest birds) using the data from Loyn (1987).

Análise de rota
(Path analysis)

Recapitulando

- Correlação linear mede grau de associação entre duas variáveis contínuas
  - r de Pearson varia de -1 a 1, sendo 0 significando ausência de relação
- Regressão linear é um modelo linear em que o preditor é uma variável contínua
- É possível participar a variância num componente explicado pelo modelo e num resíduo
  - R^2 mede o quanto o(s) preditor(es) explicam a variável resposta
- Atenção para os pressupostos!