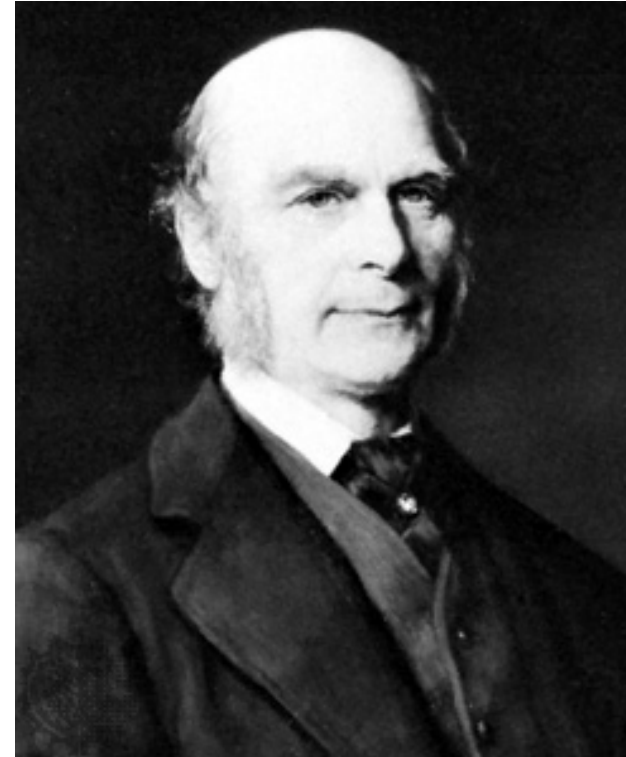


Regressão e correlação linear



Karl Pearson (1857-1936)

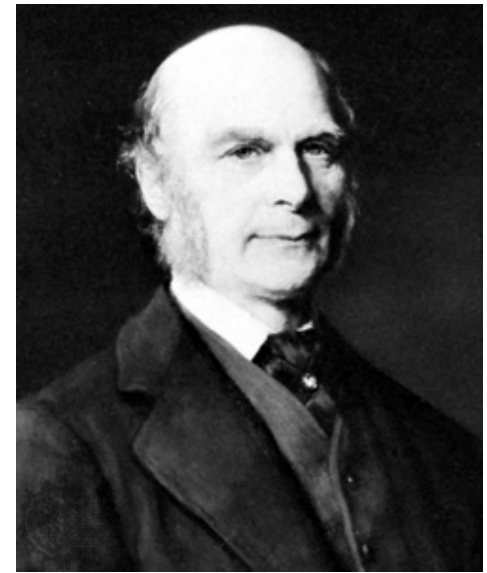


Sir Francis Galton (1822-1911)

Sumário da aula

- Breve histórico
- Introdução aos Modelos Lineares
- Equação da regressão linear (Ordinary Least Squares, OLS)
 - Parâmetros estimados
 - Intercepto ("a", α , β_0)
 - Inclinação (slope: b , β , β_1)
 - Erro e como a não independência afeta a matriz VCV
- 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)

Um pouco de história



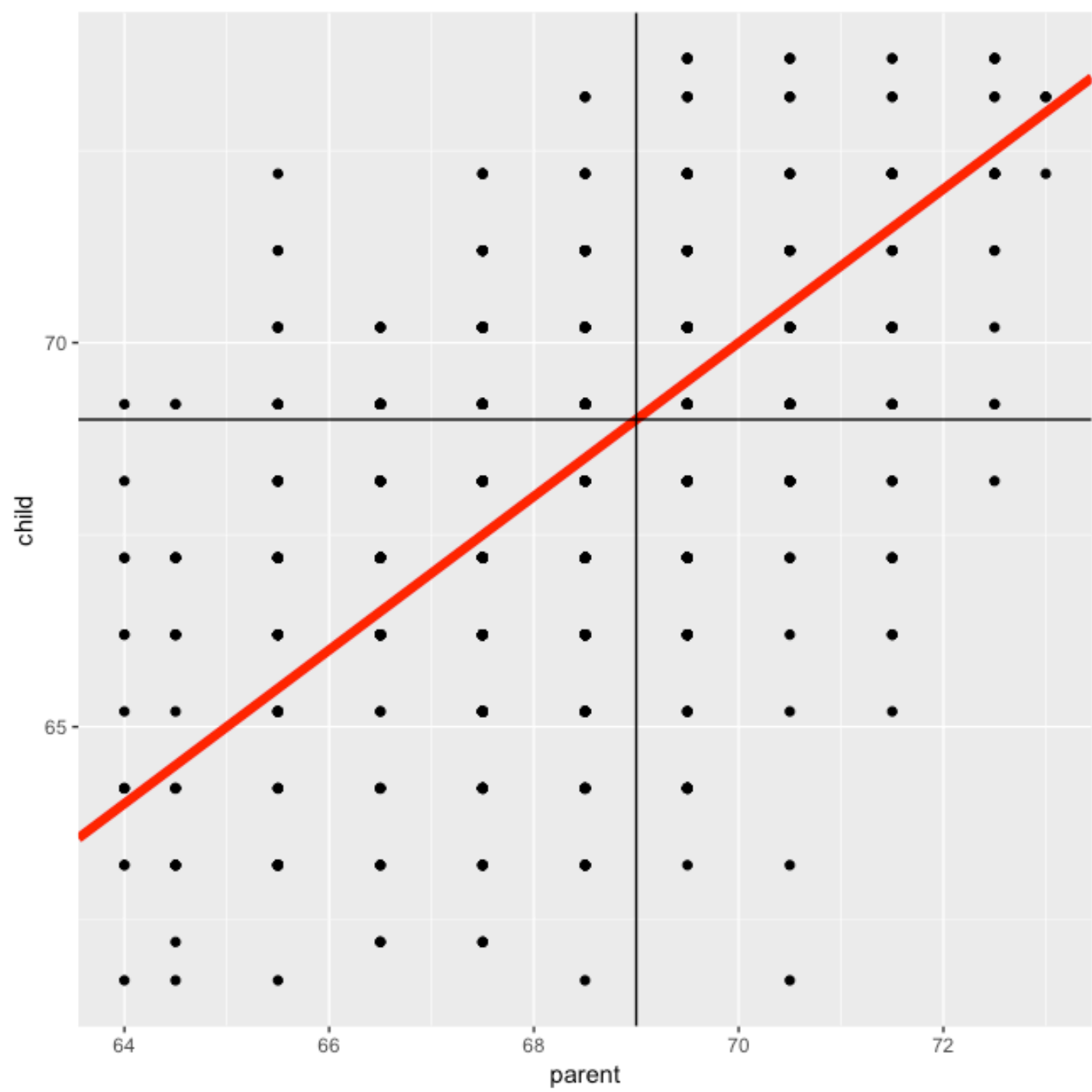
- Francis Galton (1822-1911): Antropólogo inglês, primo de Charles Darwin, defendia idéias de *eugenia*
- Cunhou o termo "regressão" no contexto de estudos de herdabilidade de caracteres quantitativos em 1886
- Mediu **altura** de filhos e seus pais para investigar o padrão de herdabilidade desta característica

“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.



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)

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.

Define coeficiente de regressão

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-Wallis)
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.

Correlação de Pearson

Coeficiente de Correlação

$$r_{Y_1Y_2} = \frac{\sum_{i=1}^n [(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)]}{\sqrt{\sum_{i=1}^n (y_{i1} - \bar{y}_1)^2 \sum_{i=1}^n (y_{i2} - \bar{y}_2)^2}}$$

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

$$-1 < r > 1$$

Coeficiente de Correlação

$$r_{Y_1Y_2} = \frac{\sum_{i=1}^n [(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)]}{\sqrt{\sum_{i=1}^n (y_{i1} - \bar{y}_1)^2 \sum_{i=1}^n (y_{i2} - \bar{y}_2)^2}}$$

Soma dos produtos cruzados
(Covariância)

Desvio padrão das duas variáveis

Coeficiente de Correlação

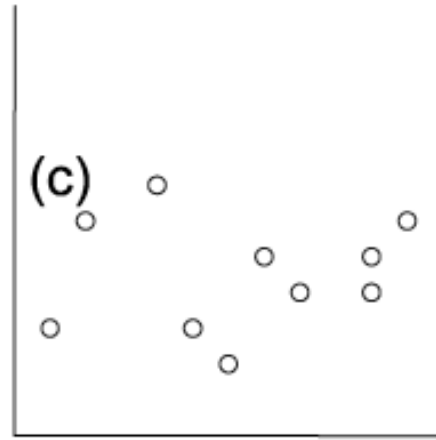
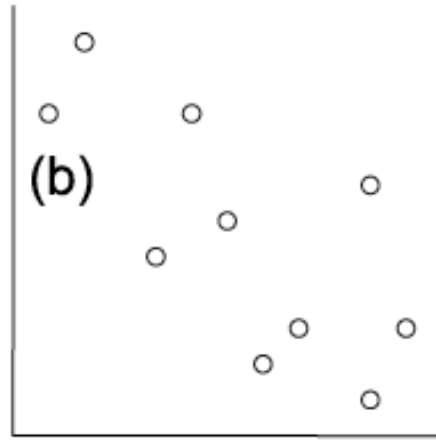
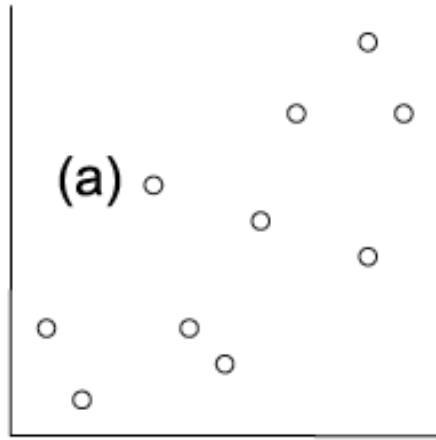
$$r_{Y_1Y_2} = \frac{\sum_{i=1}^n [(y_{i1} - \bar{y}_1)(y_{i2} - \bar{y}_2)]}{\sqrt{\sum_{i=1}^n (y_{i1} - \bar{y}_1)^2 \sum_{i=1}^n (y_{i2} - \bar{y}_2)^2}}$$

Pode ser positivo ou negativo

Sempre positivo

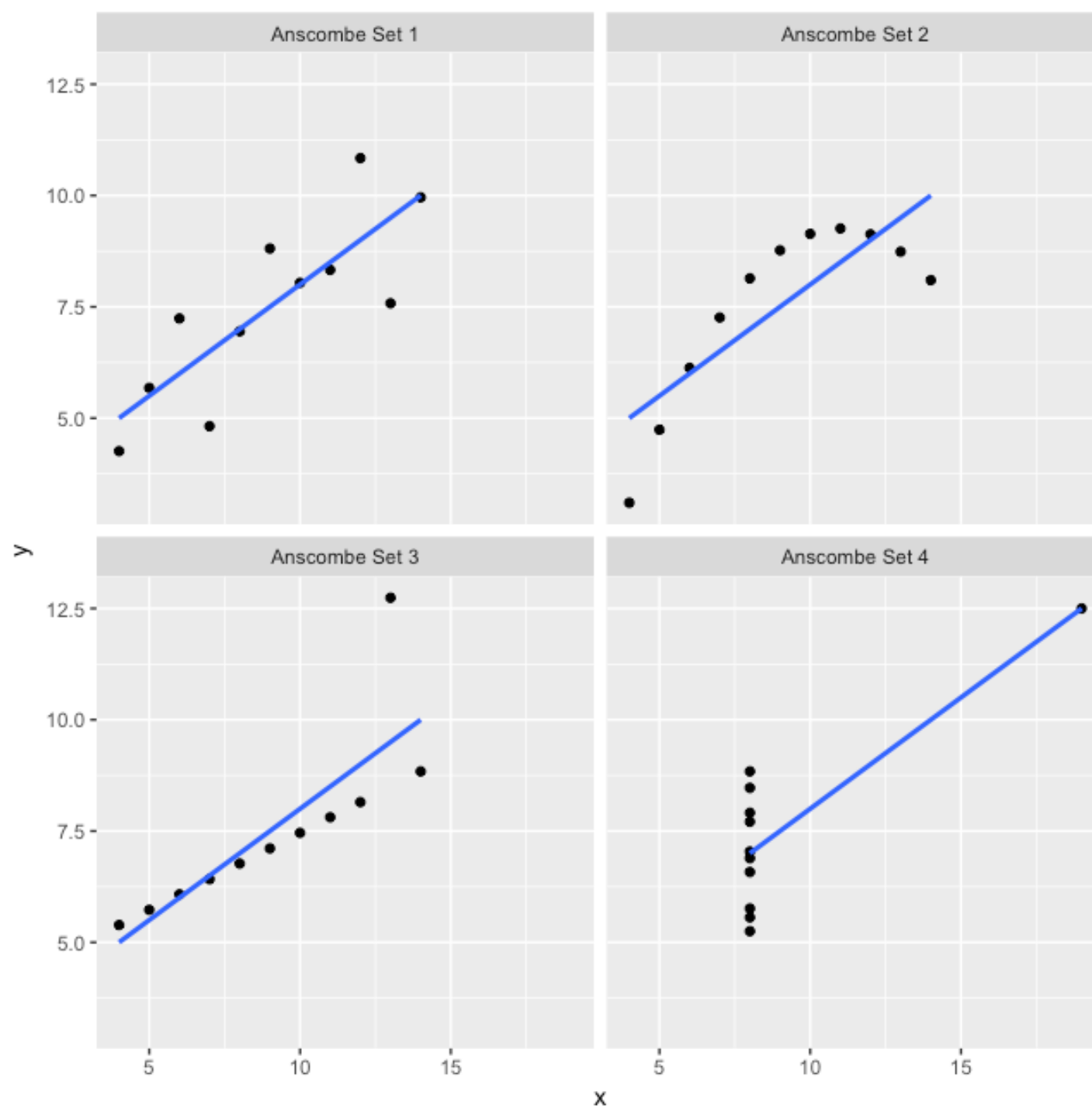
Se o numerador for positivo o r será negativo e vice-versa

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 ($r = 0.08$).



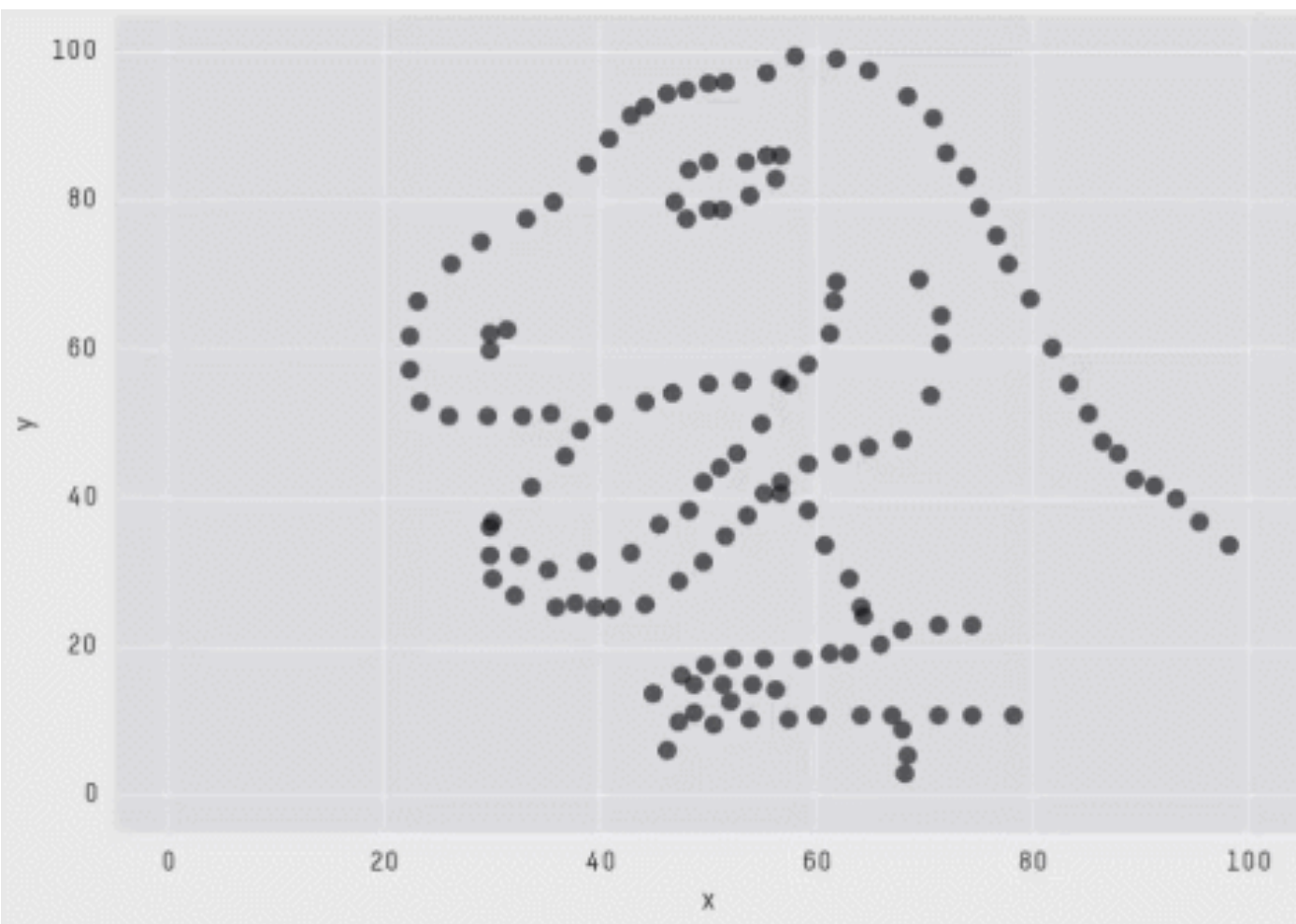
Quarteto de Anscombe

!!Sempre inspecione seus dados antes de fazer qualquer análise!!

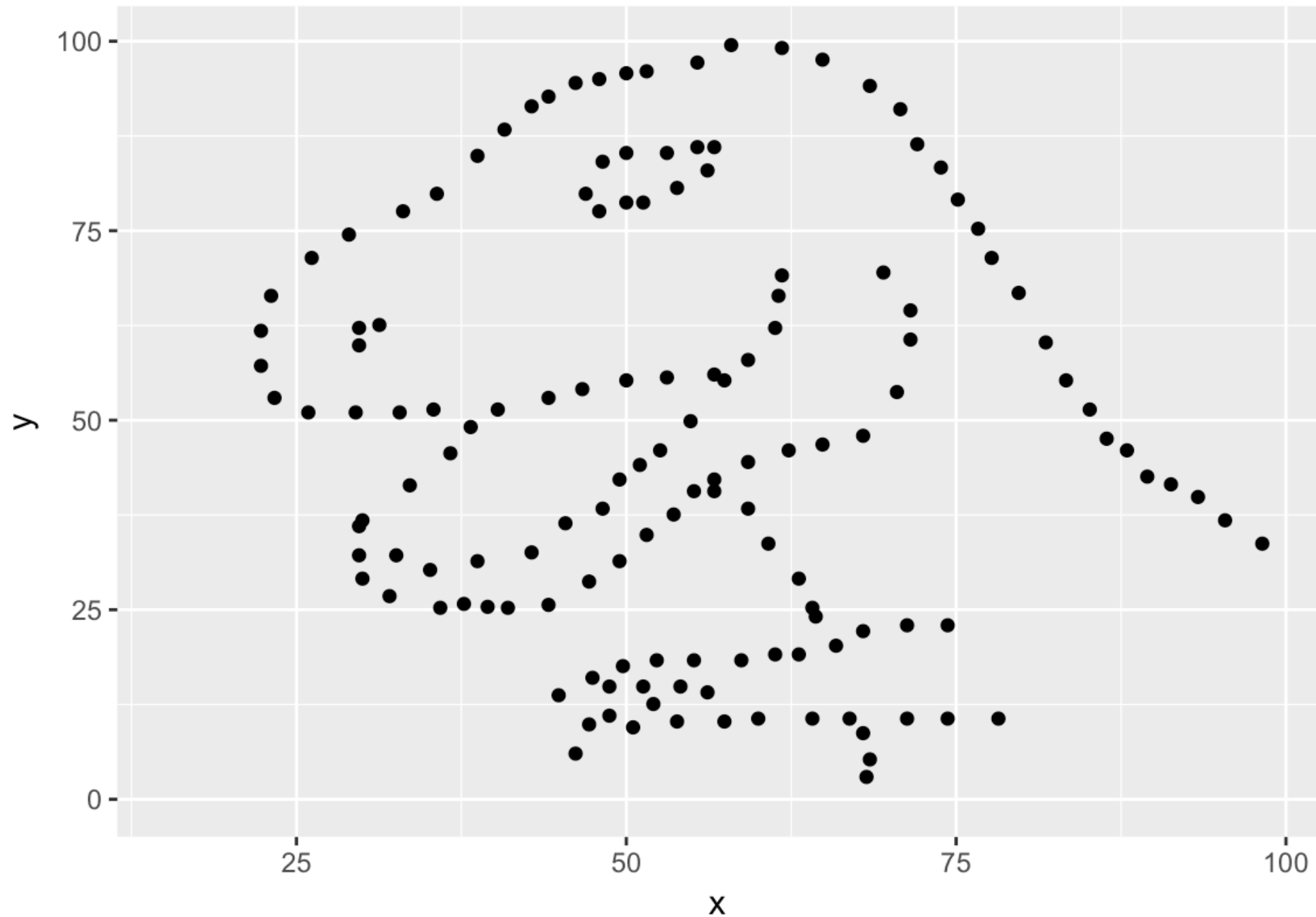


Todos esses plots têm o mesmo coeficiente de correlação

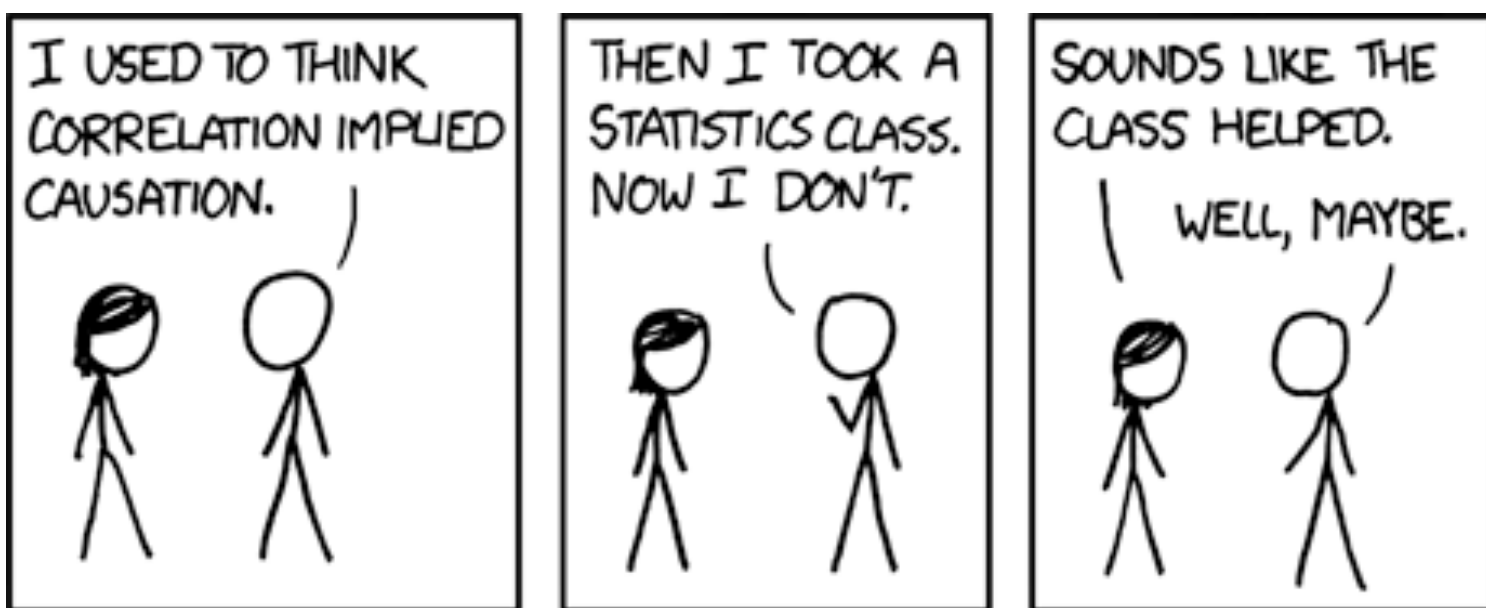
$$r = 0.816$$



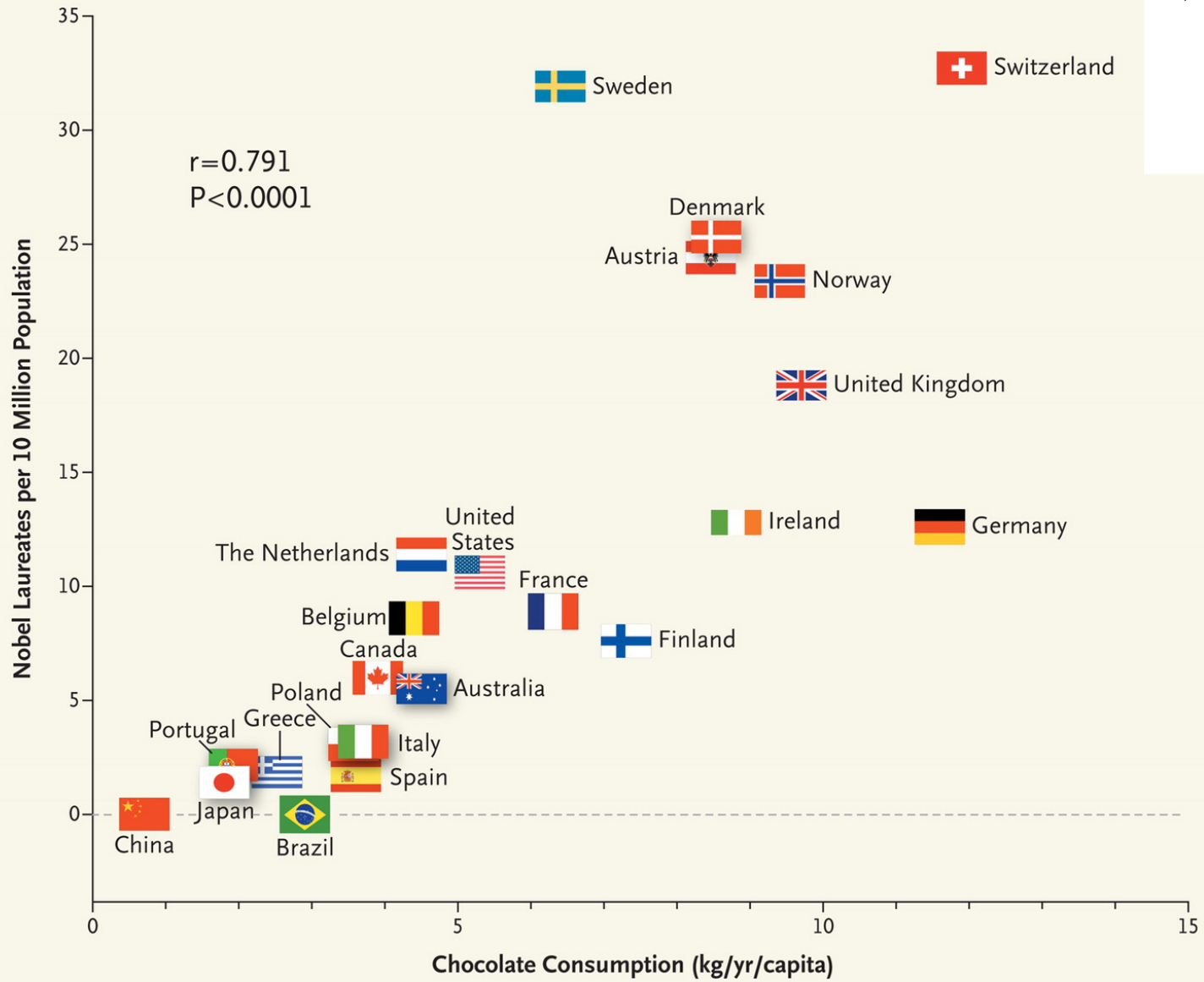
X Mean: 54.2659224
Y Mean: 47.8313999
X SD : 16.7649829
Y SD : 26.9342120
Corr. : -0.0642526



CORRELATION **DOES NOT** IMPLY CAUSATION



OCCASIONAL NOTES



Chocolate Consumption, Cognitive Function, and Nobel Laureates

Franz H. Messerli, M.D.

[Compare US states](#)**[Compare weekly time series](#)**[Compare monthly time series](#)Shift series weeksCountry: **Documentation**[Comic Book](#)[FAQ](#)[Tutorial](#)[Whitepaper](#)[Correlate Algorithm](#)**Correlate Labs**[Search by Drawing](#)

Find searches that correlate with real-world data

Google Correlate finds search patterns which correspond with real-world trends.

Compare time series

Many search terms vary in popularity over time. To find terms that vary in a similar way to your own time series, enter your data using the link above. Or take a look at these examples to see which search terms:

- [...are more popular in winter](#)
- [...were most likely to be issued in 2005](#)
- [...match the pattern of actual flu activity](#)
(this is how we built Google Flu Trends!)

You can also enter a query into the search box above to find search terms that have a similar pattern of activity, or try one of these:

- [mittens](#)
- [losing weight](#)
- [ribosome](#)

Compare US states

Search terms are often popular in some states and less popular in others. To find terms whose pattern of activity across the United States reflects your own US states dataset, enter your data using the link above. Or, you can find terms correlated with:

- [...the state's latitude](#)
- [...being in New England](#)
- [...annual rainfall in the state](#)

You can also use the search box above to see which searches correlate state-by-state to any query, or try one of these:

- [mittens](#)
- [hunting season](#)
- [southern cooking](#)



Teste de hipótese do r de Pearson

- Utiliza um teste t :

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

Erro padrão do r

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

- Testa a hipótese nula de que $\rho = 0$

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 expoente 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)

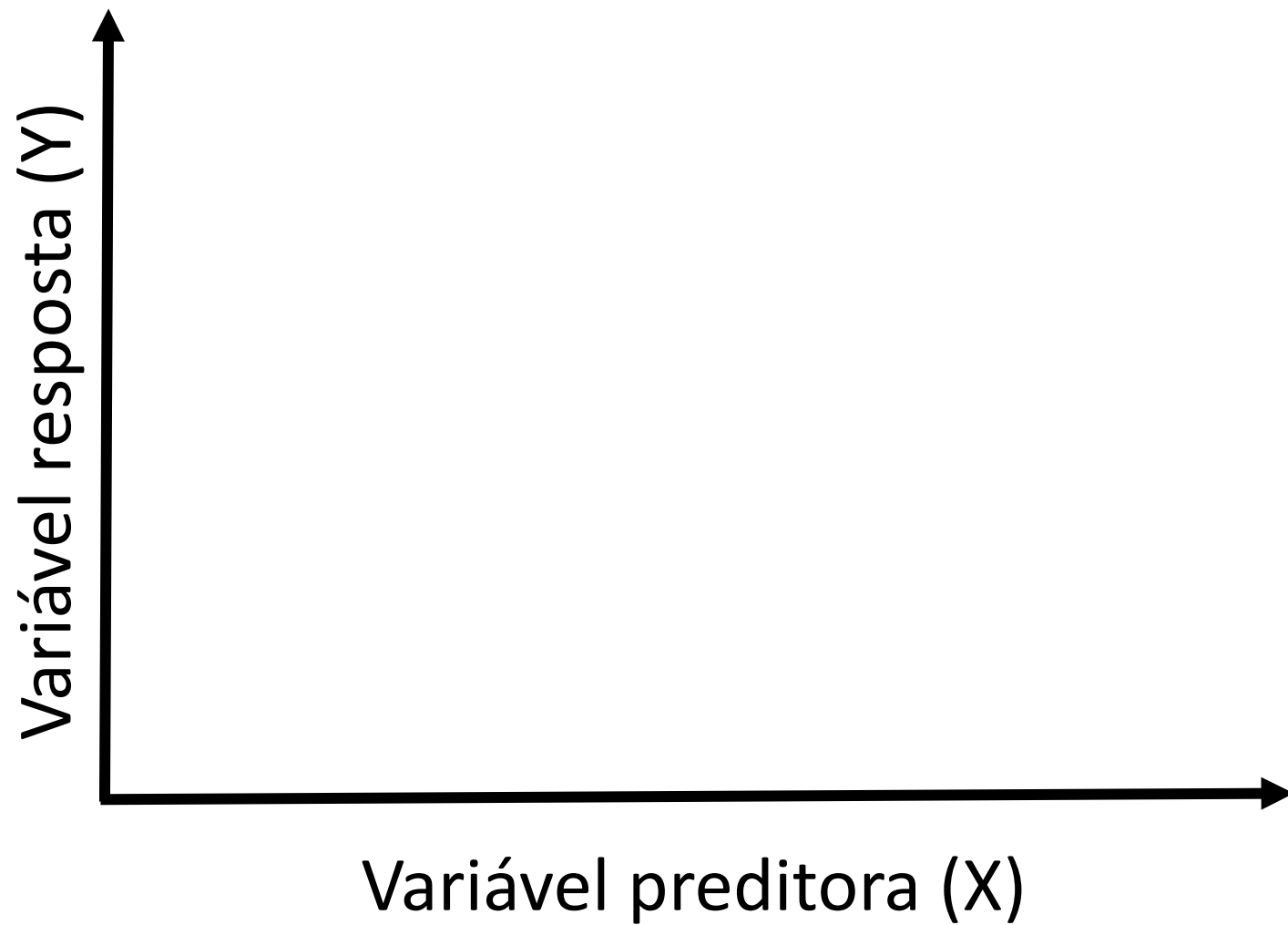
Exemplo

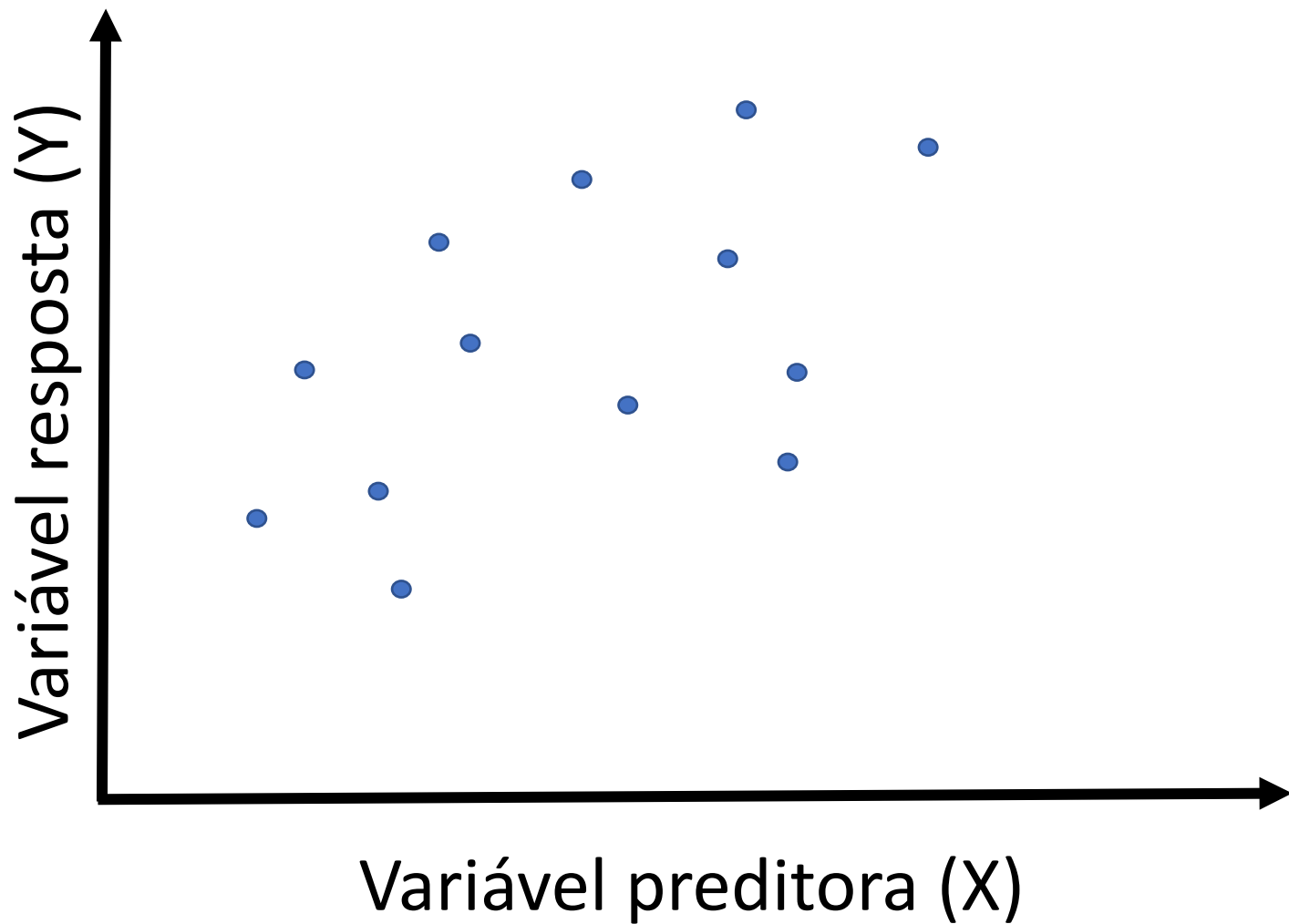
Altura dos filhos = $\alpha + \beta^*$ altura dos pais + resíduos

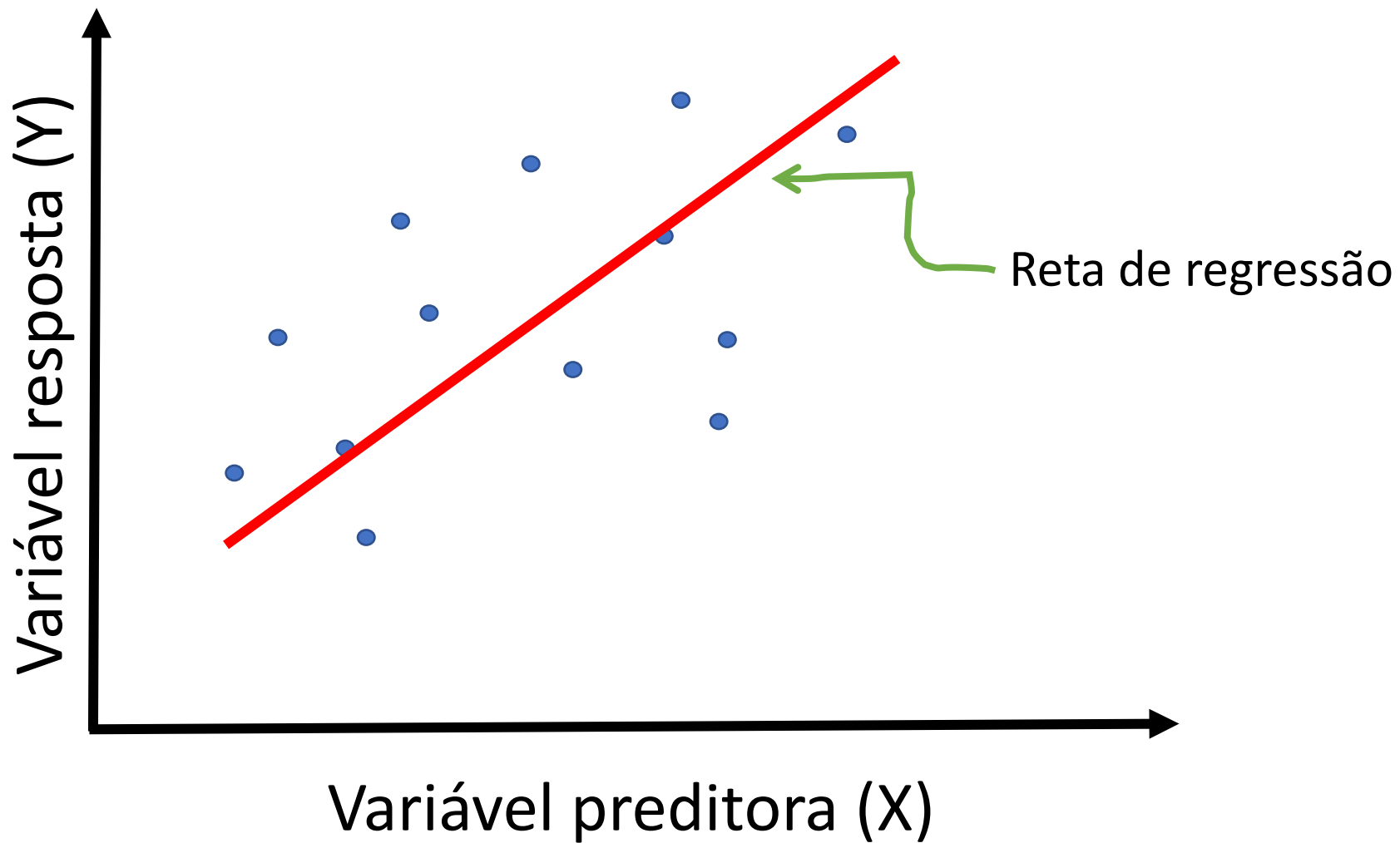
Parâmetros do modelo

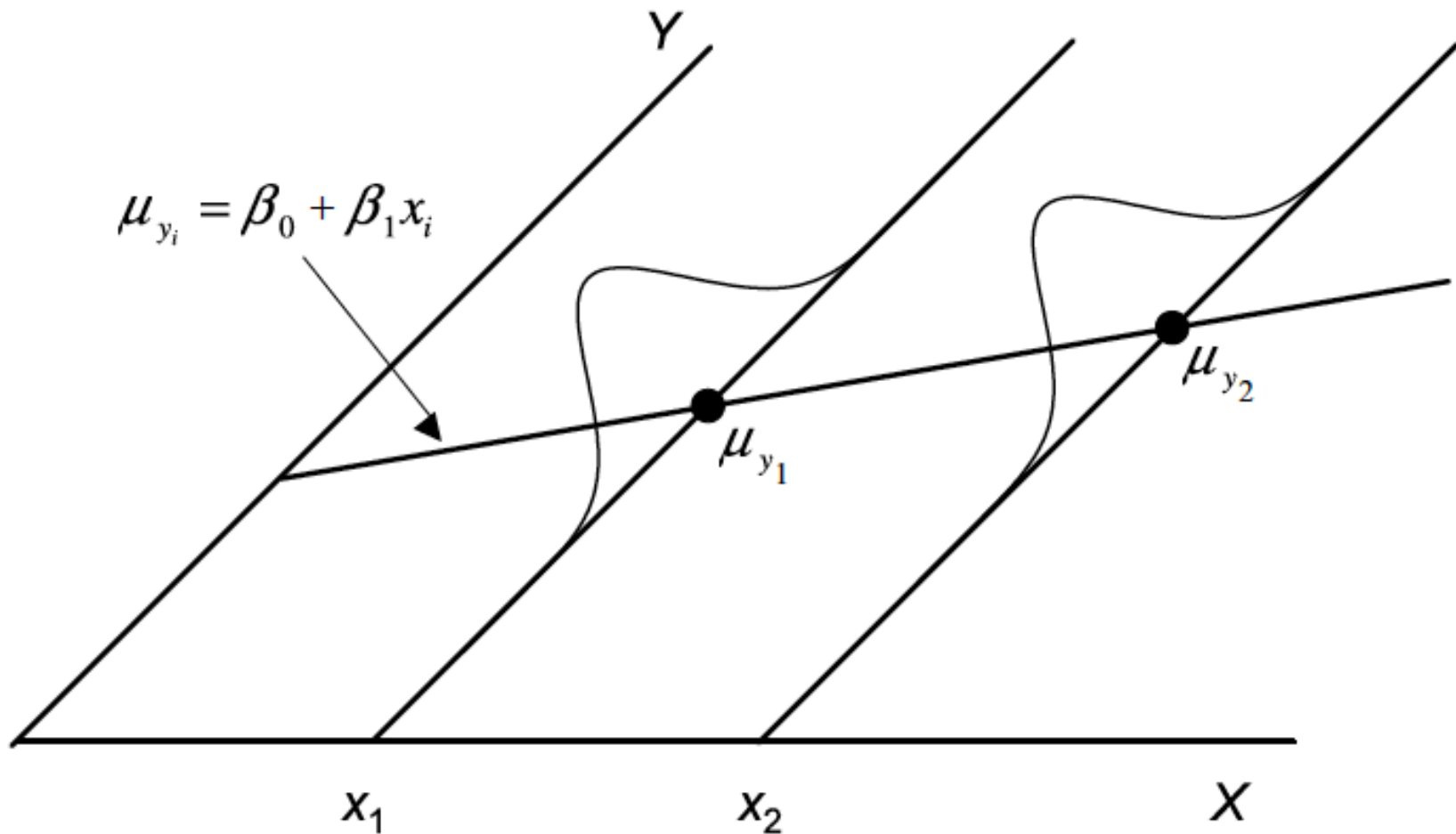
- Intercepto = valor médio da variável resposta quando a variável preditora é zero
- Inclinação (slope) = inclinação da reta predita pelo modelo. Mede o quanto a variável resposta muda em função do aumento da variável preditora.
 - Pode-se calcular o IC95% para o slope

Gráfico de dispersão









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

Fig. 5.5 Quinn & Keough

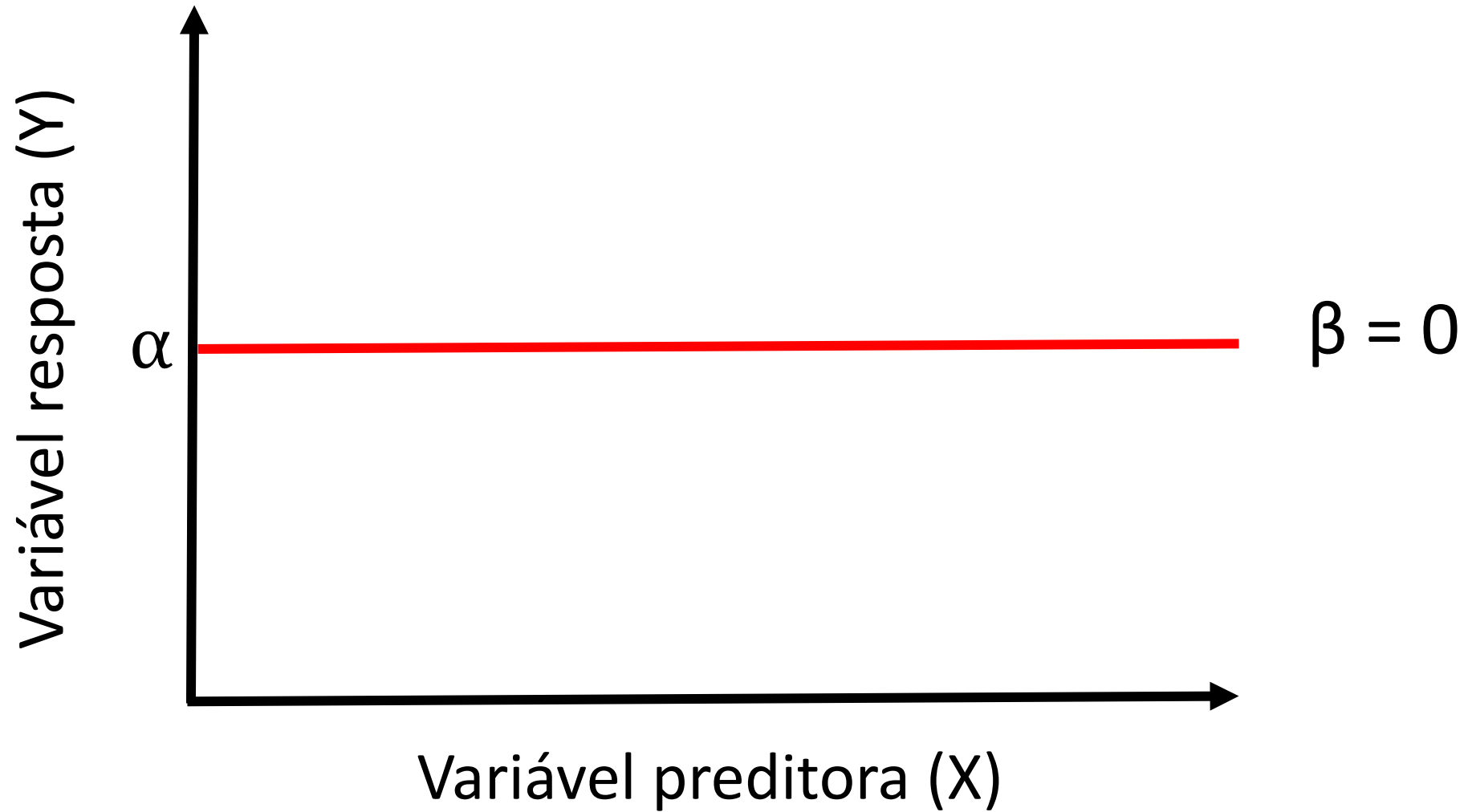
Teste de hipótese

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

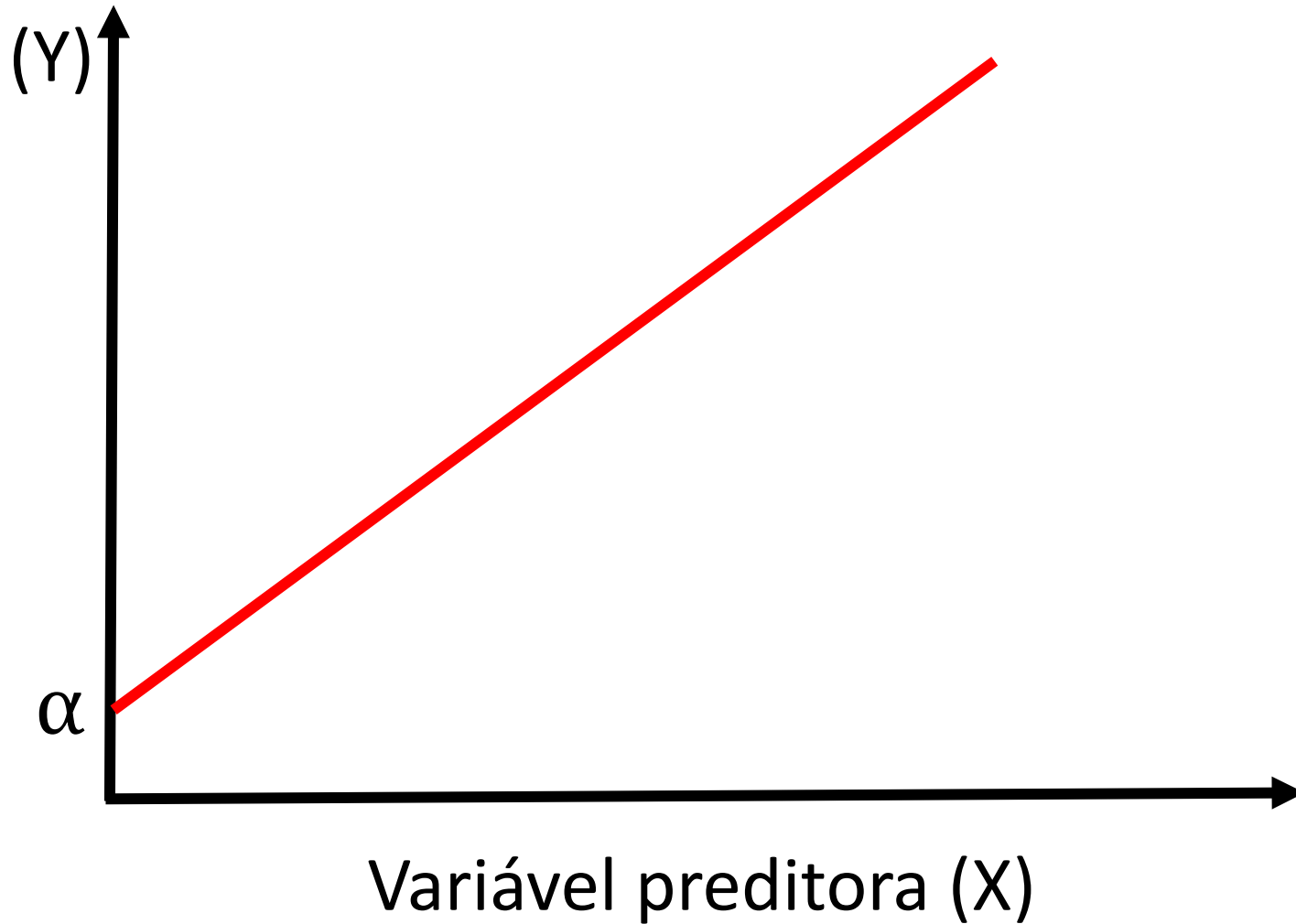
$$\beta \neq 0$$

- 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

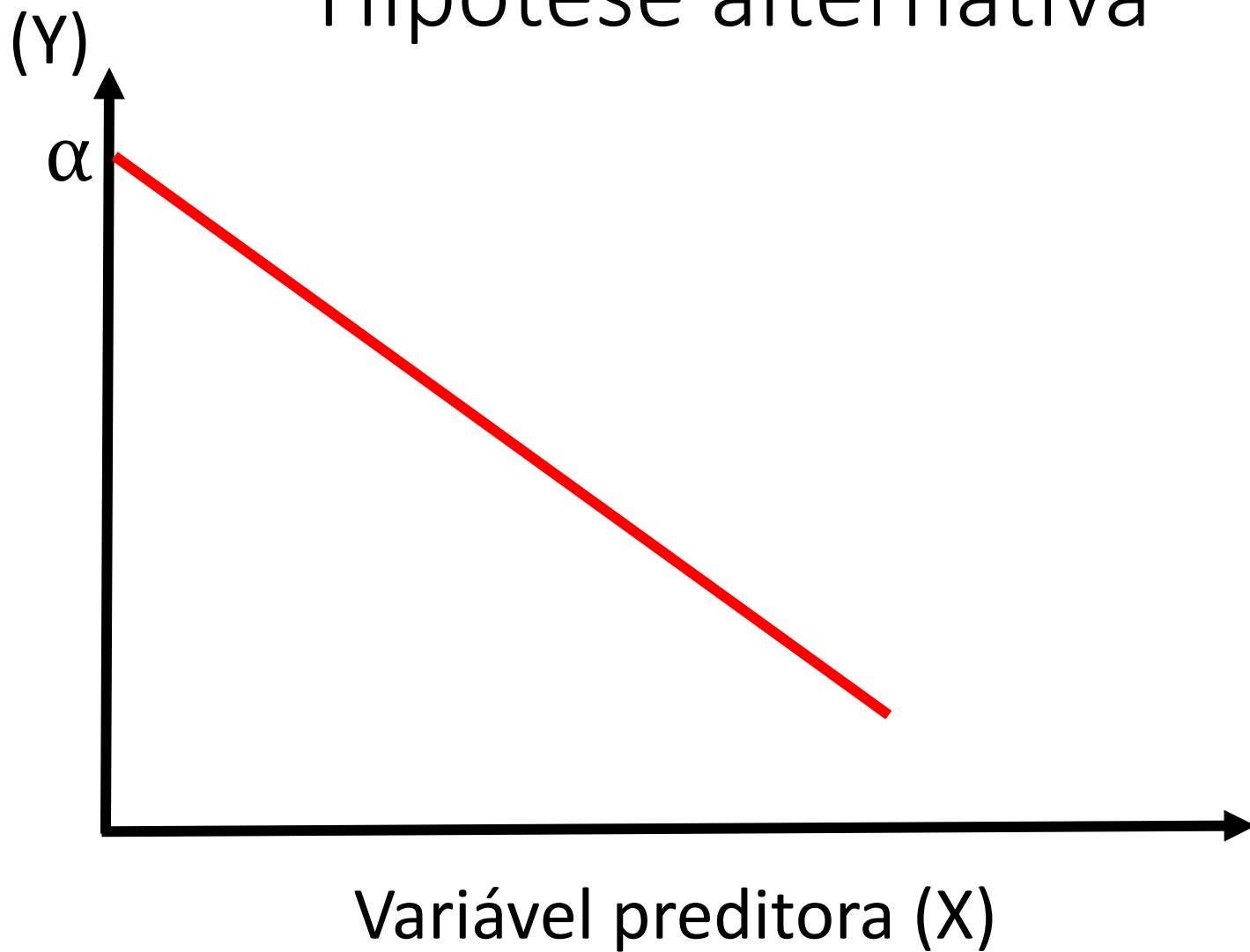


Hipótese alternativa



$$\beta > 0$$

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

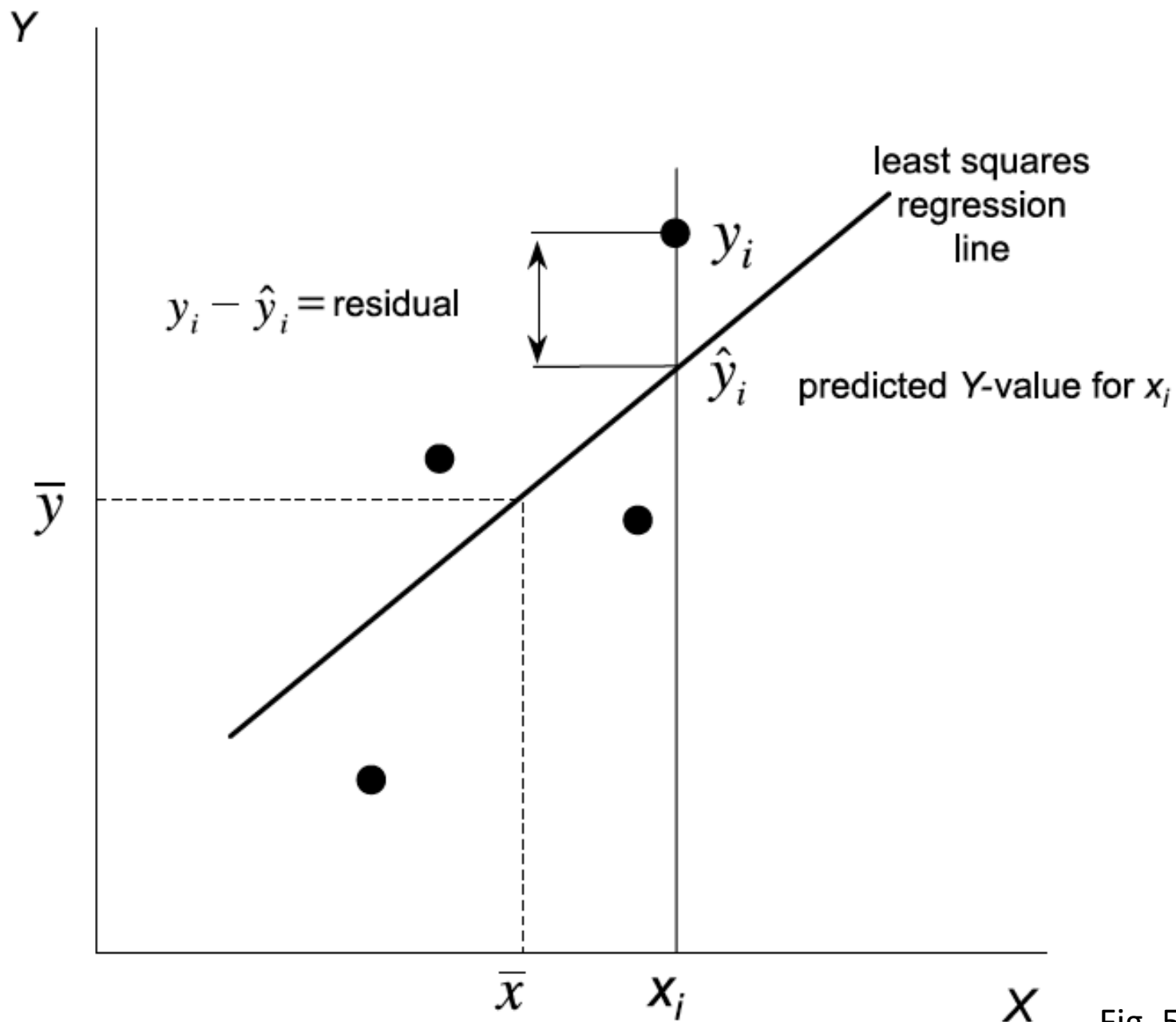
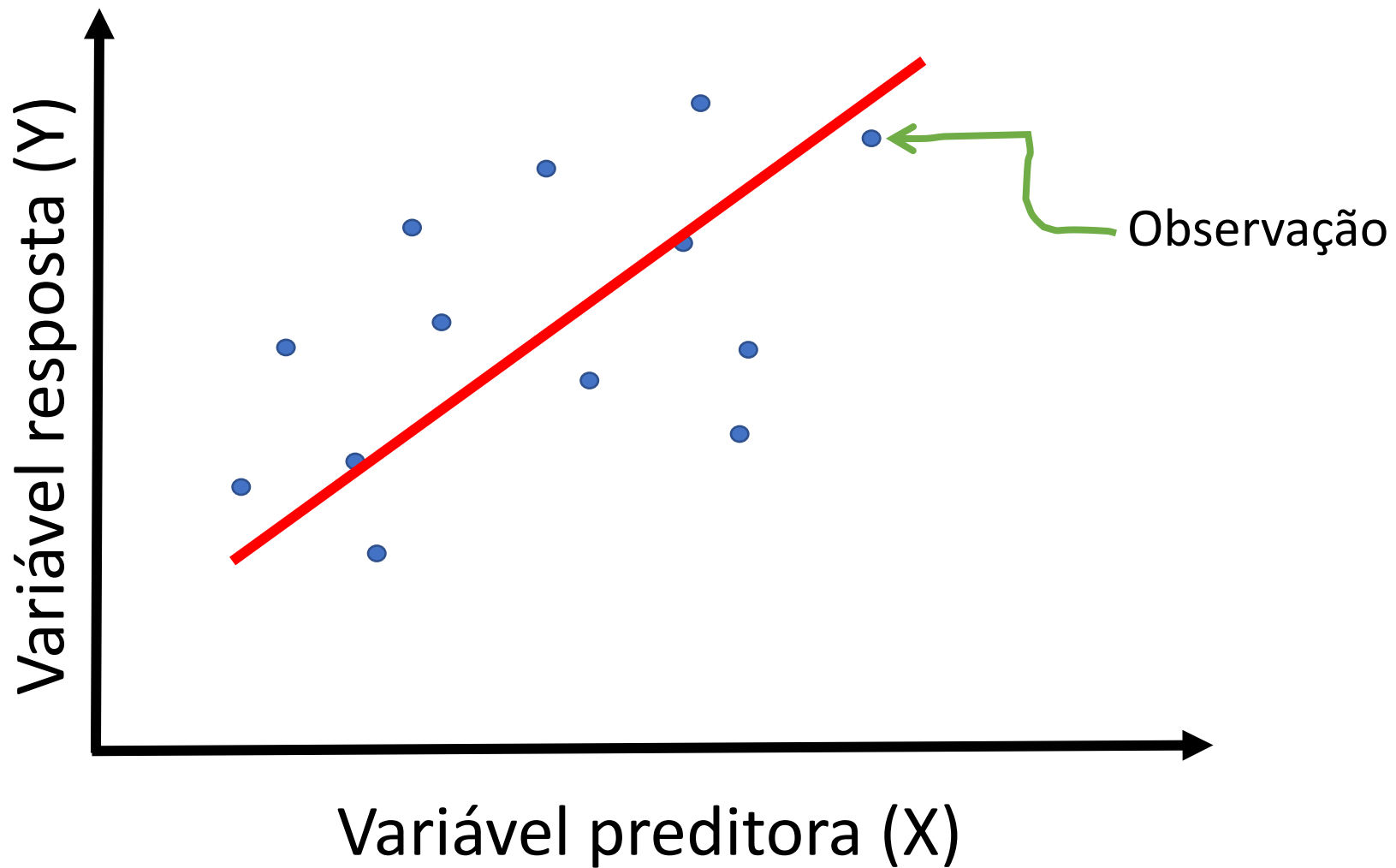
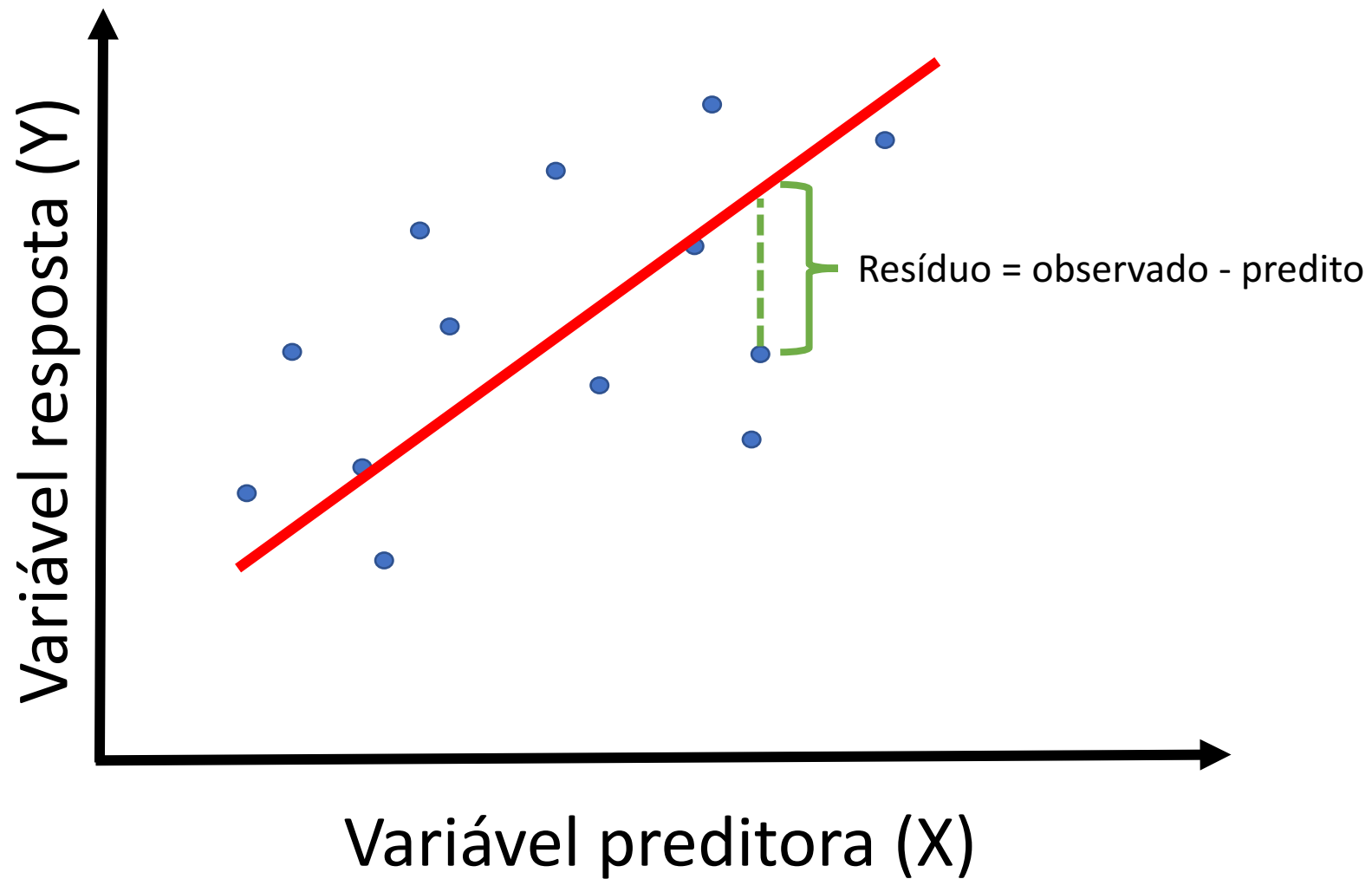
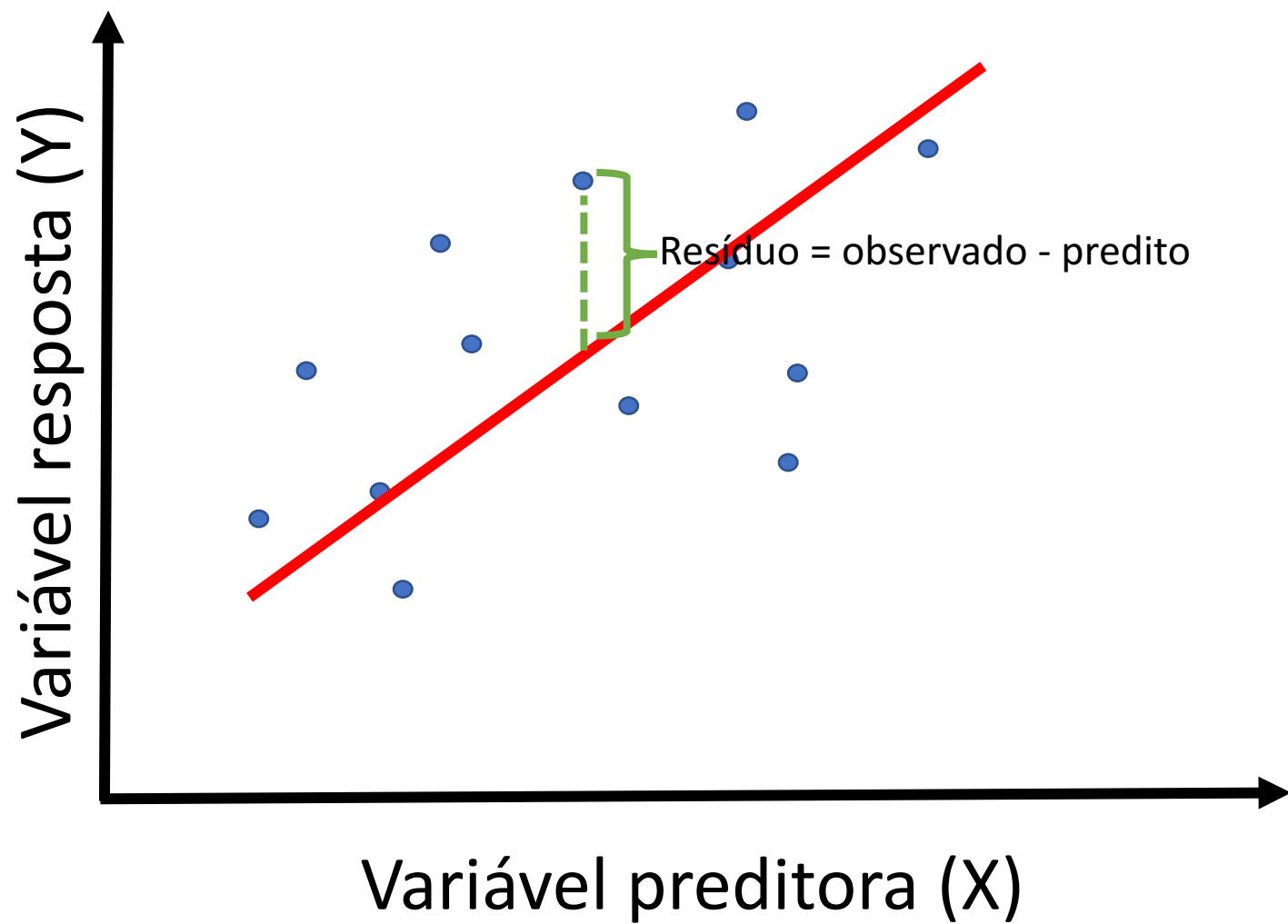


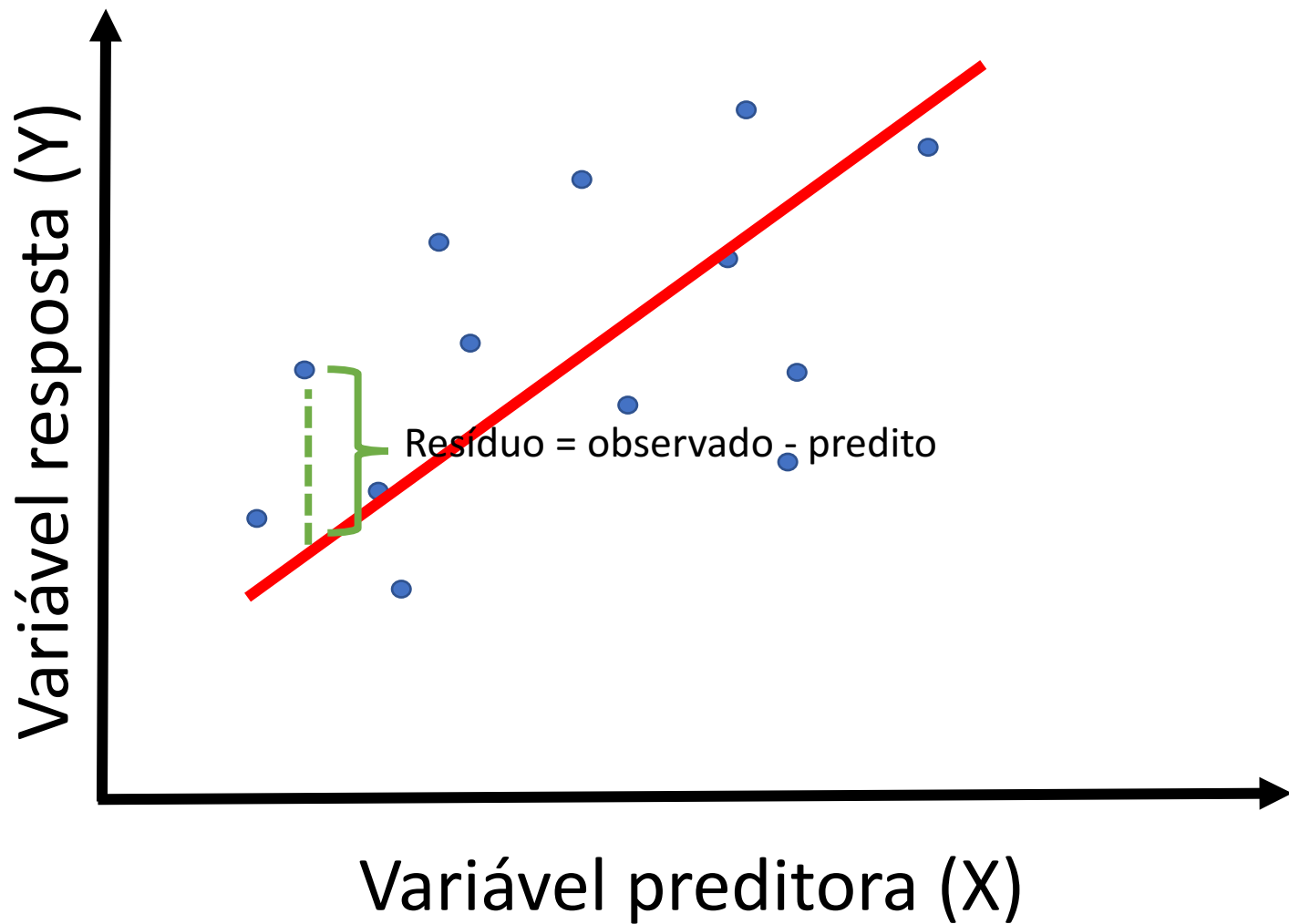
Fig. 5.6 Quinn & Keough





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





Soma de quadrados dos resíduos

$$SQ_{res} = \sum (\text{observação} - \text{predito})^2$$

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

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

β_0

$$b_0 = \bar{y} - b_1 \bar{x}$$

Média de Y

Média de x

Slope

Table 5.2 | Parameters of the linear regression model

Parameter

OLS estimate

β_1

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

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

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

$$s_{b_1} = \sqrt{\frac{MS_{\text{Residual}}}{\sum_{i=1}^n (x_i - \bar{x})^2}}$$

 β_0

$$b_0 = \bar{y} - b_1 \bar{x}$$

$$s_{b_0} = \sqrt{MS_{\text{Residual}} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

 ε_i

$$e_i = y_i - \hat{y}_i$$

$$\sqrt{MS_{\text{Residual}}} \text{ (approx.)}$$

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	$b_0 = 68.09 - 0.6463 * 68.31 = 23.94153$	$s_{b_0} = 2.81088$
ε_i	Vetor para cada observação	2.239

Teste de hipótese

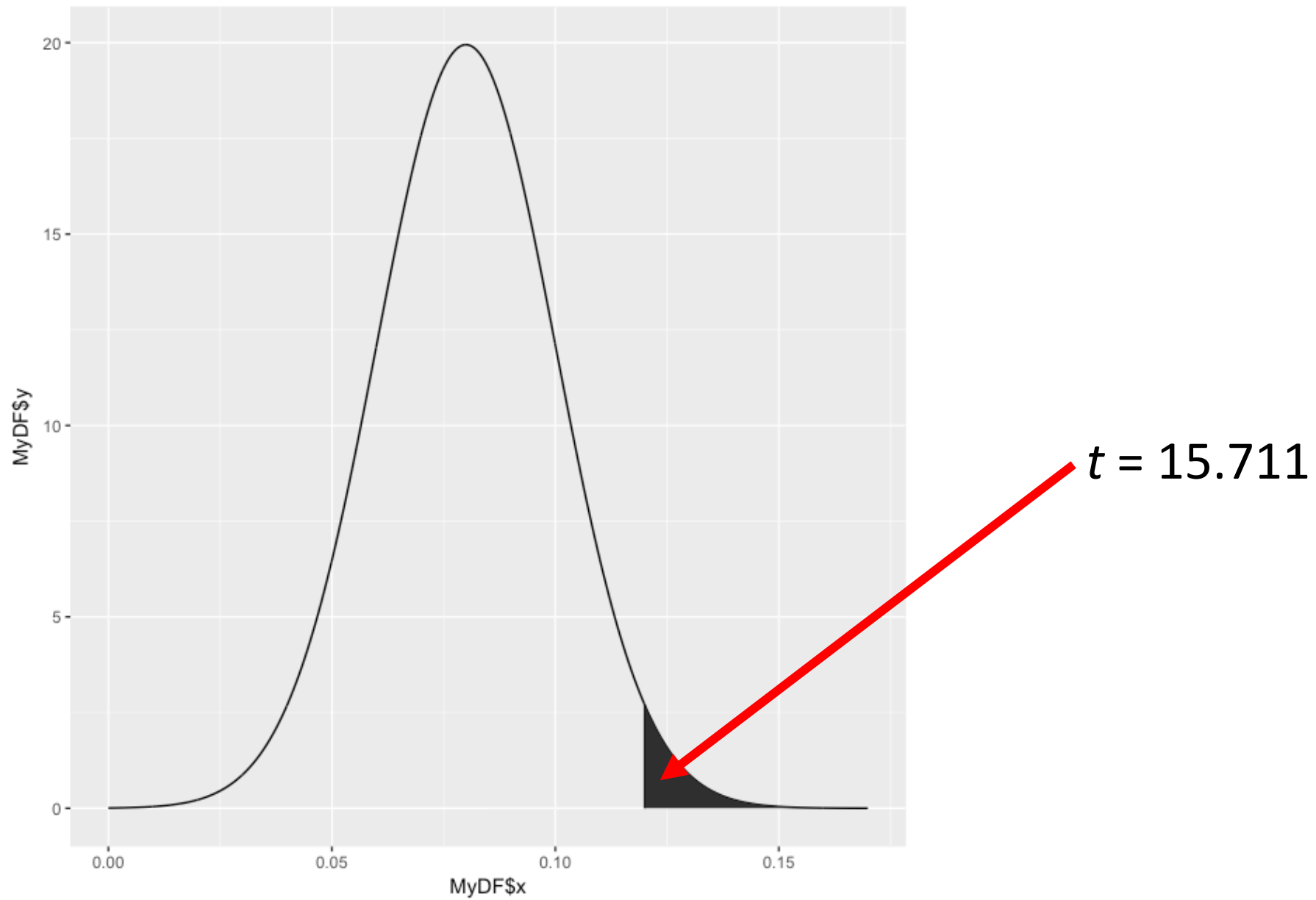
- Estatística do teste:

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

Teste t de um parâmetro

Usada para testar a hipótese nula de que $\beta = 0$

Teste de hipótese



Análise de variância

- Particionando a variância total em componentes explicados e residuais -

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 preditor 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 ao um em que só um 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
 - Apresento-lhes o Quadrado Médio (Mean squares)
- Porém, o Quadrado Médio (QM) não tem essa propriedade aditiva, como a SQ
 - $QM_{\text{regressão}} + QM_{\text{resíduos}} \neq QM_{\text{total}}$
- O $QM_{\text{resíduos}}$ estima a variância do erro (σ^2_{ε})

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

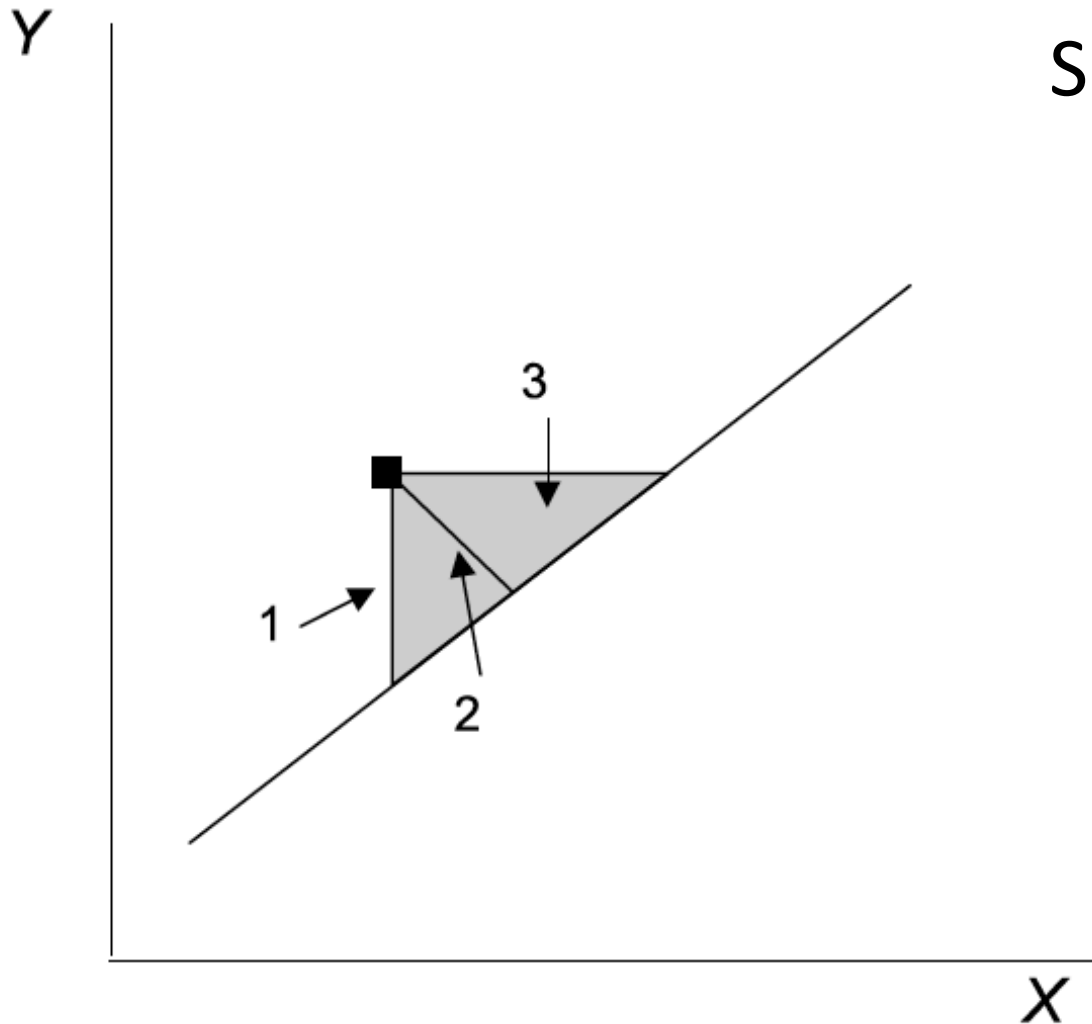
Source of variation	SS	df	MS	Expected mean square
Regression	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	1	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{1}$	$\sigma_{\epsilon}^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$
Residual	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - 2$	$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2}$	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	$\sigma_{\epsilon}^2 + \beta_1^2 \sum_{i=1}^n (x_i - \bar{x})^2$
Residual	4640.3	926	5.01	5.005
Total	5877.2	926		

Outras formas de calcular os resíduos

Soma de quadrados tipo 2 e 3



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

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

Coeficiente de determinação

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

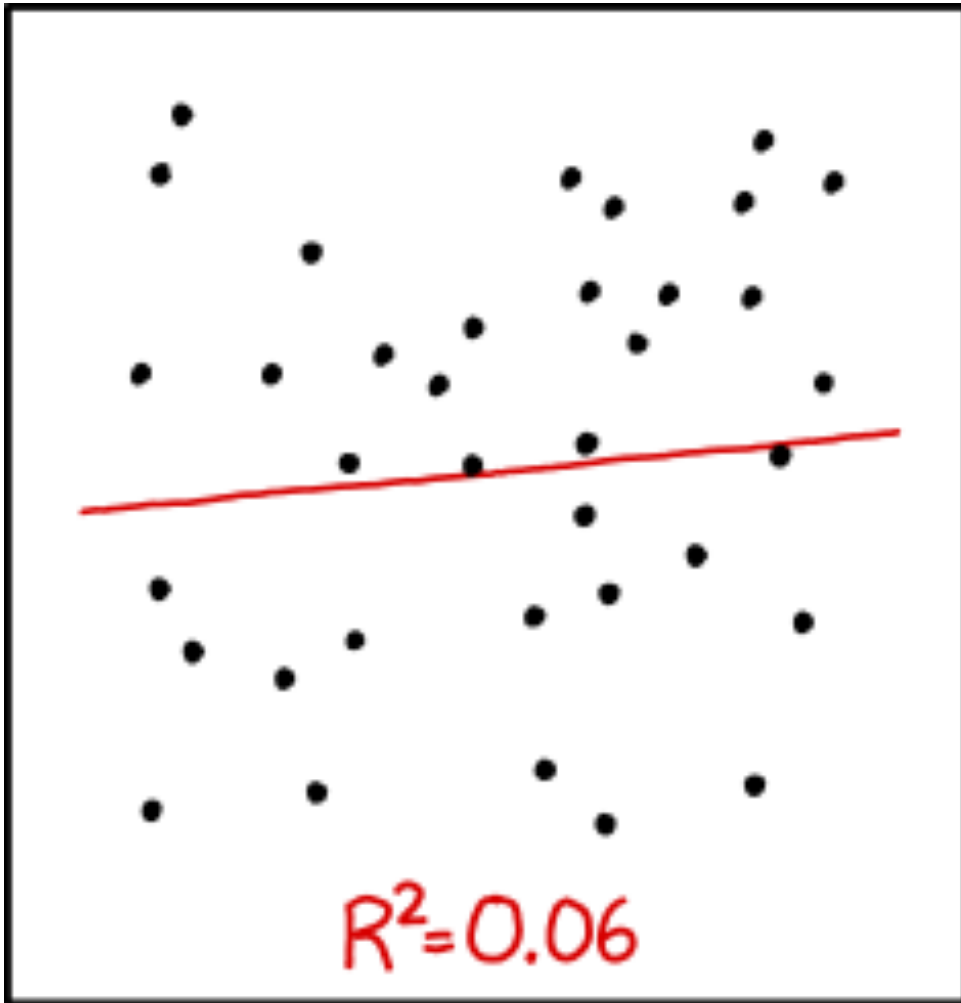
$$0 < r^2 < 1$$

$$\frac{1236.9}{5877.2} = 0.2105$$

$$1 - \frac{4640.3}{5877.2} = 0.2105$$

$r^2 = r$ Pearson
ao quadrado

A variação na altura dos pais explicou cerca de 21%
da variação da altura dos filhos



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.

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 $\text{Var}[e] = \sigma^2$
- Independência das unidades amostrais
 - Pseudoréplicas, lembrem da última aula?

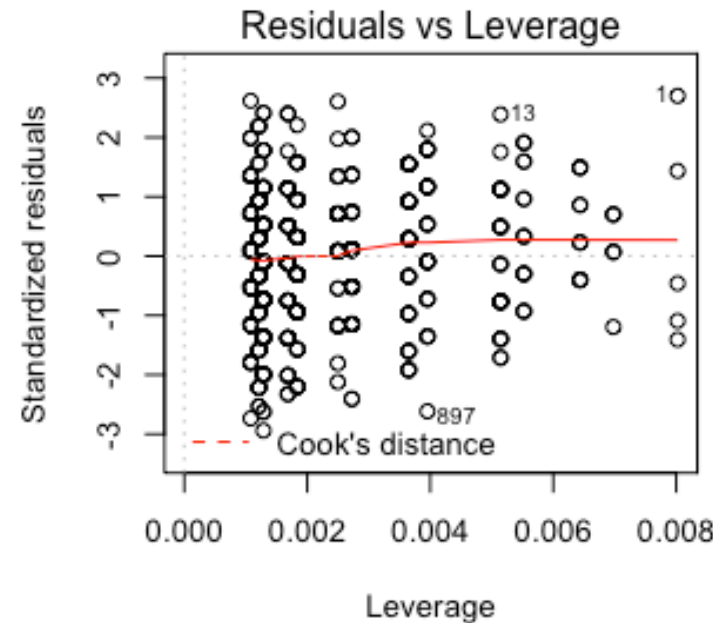
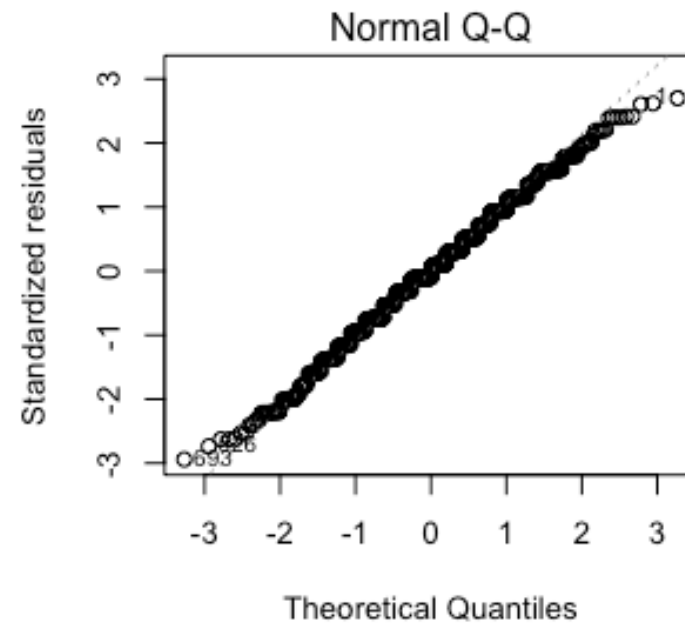
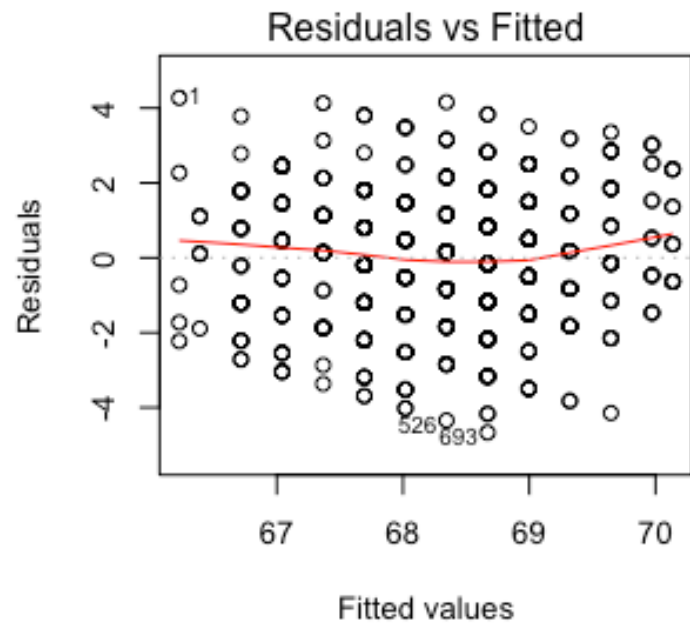
Diagnósticos dos resíduos

Homogeneidade de
variância

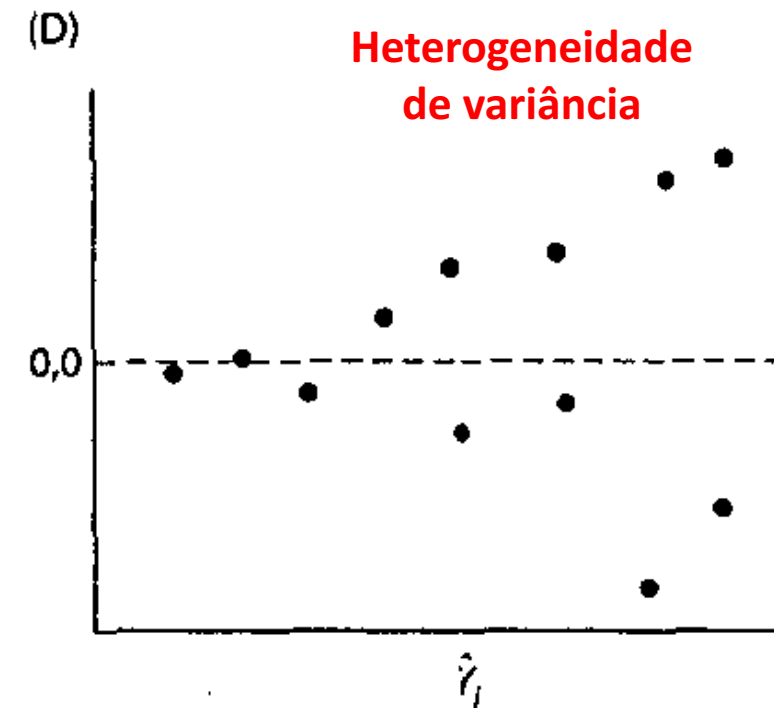
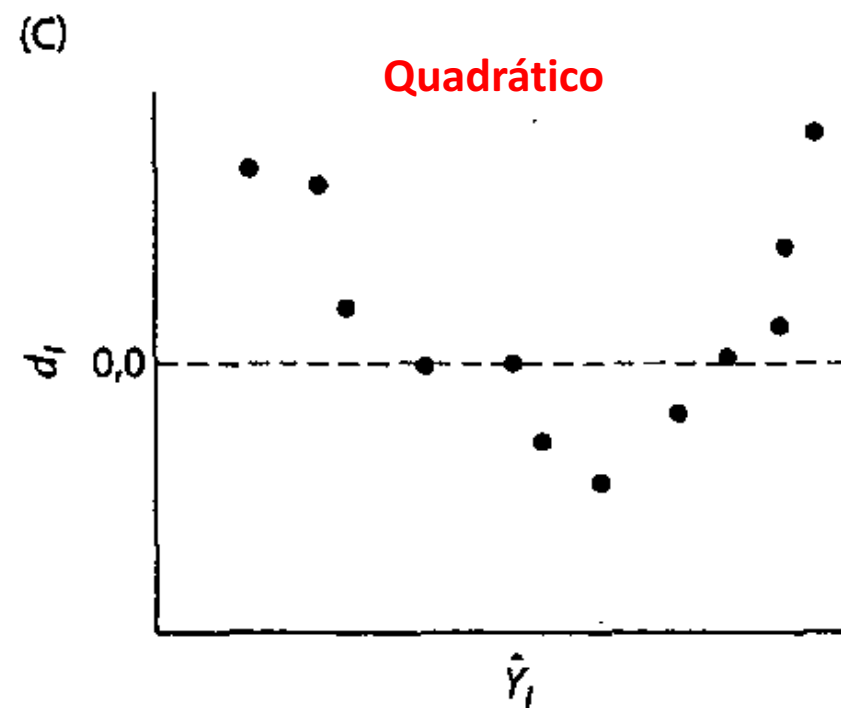
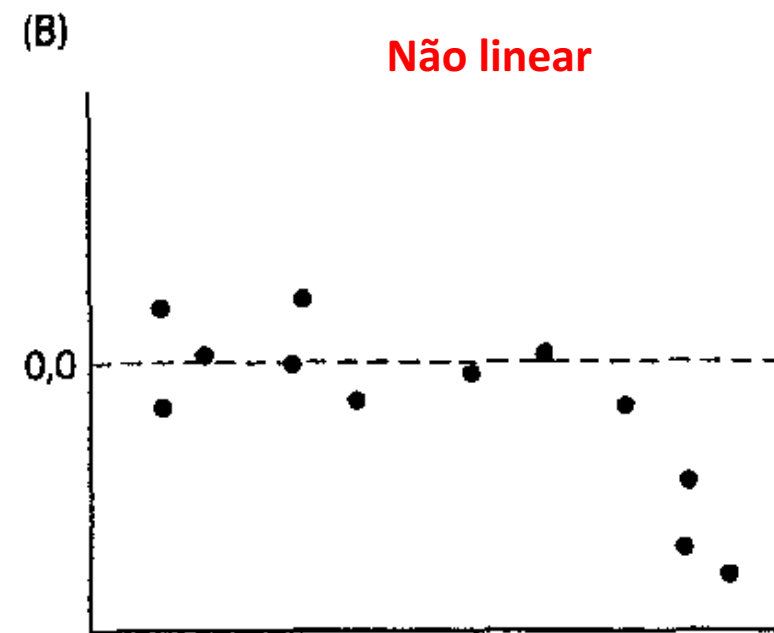
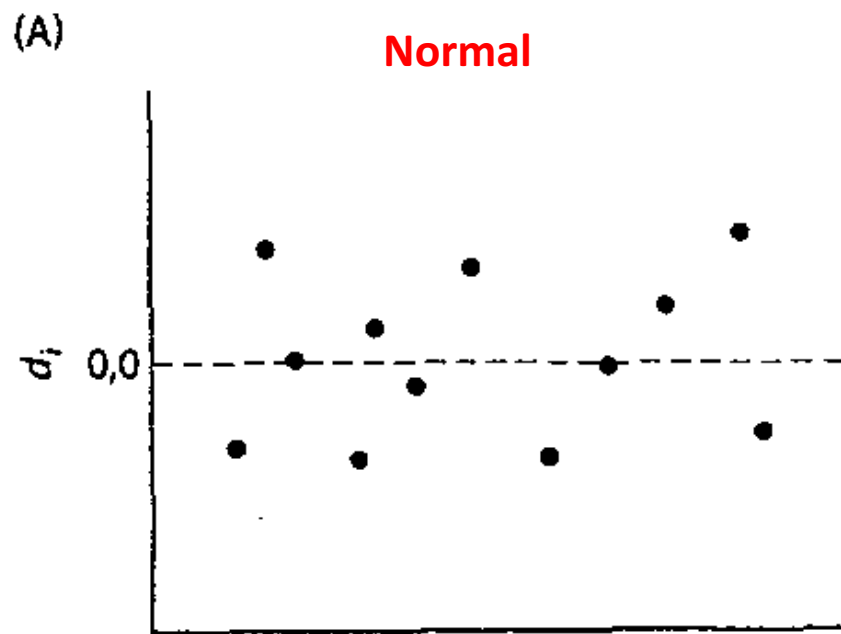
Normalidade

Homogeneidade de
variância (com resíduos
padronizados)

Valores extremos
(outliers)



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



Bora botar a mão na massa

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 ($Y \sim X_1 + X_2 + \dots + X_n$)
- Logística
- (semi)Parcial => particiona a variação em termos dos preditores isolados e em conjunto
- Ponderada (weighed 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 hierárquica

Regressão múltipla

- Relaciona vários preditores à variável resposta
- Coeficientes de regressão parcial (padronizados) => influência de cada preditor na resposta

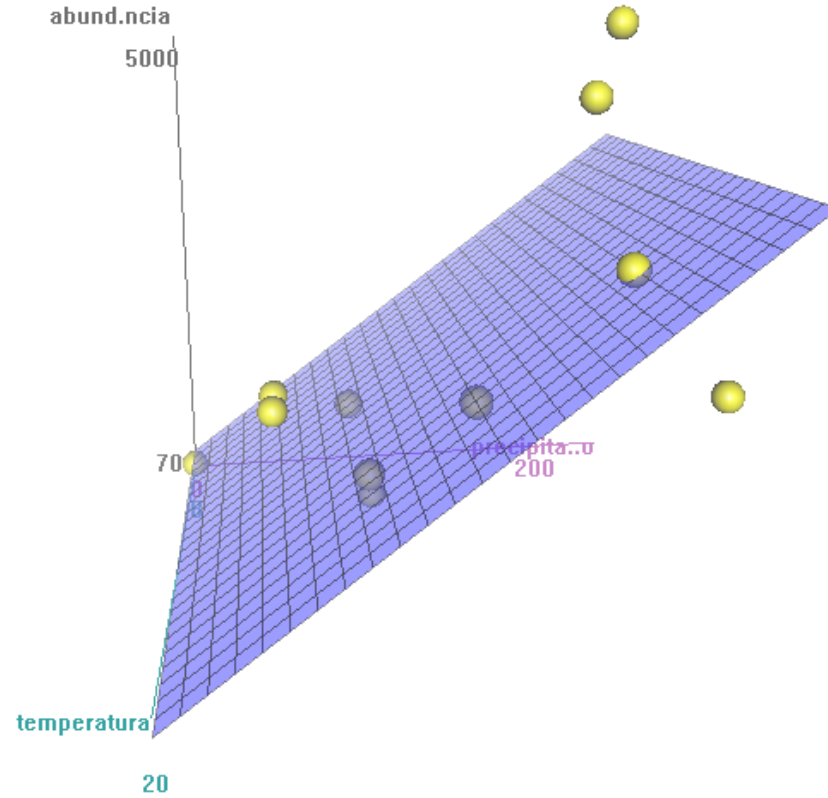
- Problema com multicolinearidade
 - Como diagnosticar? Variation Inflation Factor (VIF)

$$R^2 \text{ ajustado} = 1 - \frac{SS_{\text{Residual}}/[n - (p + 1)]}{SS_{\text{Total}}/(n - 1)}$$

- 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

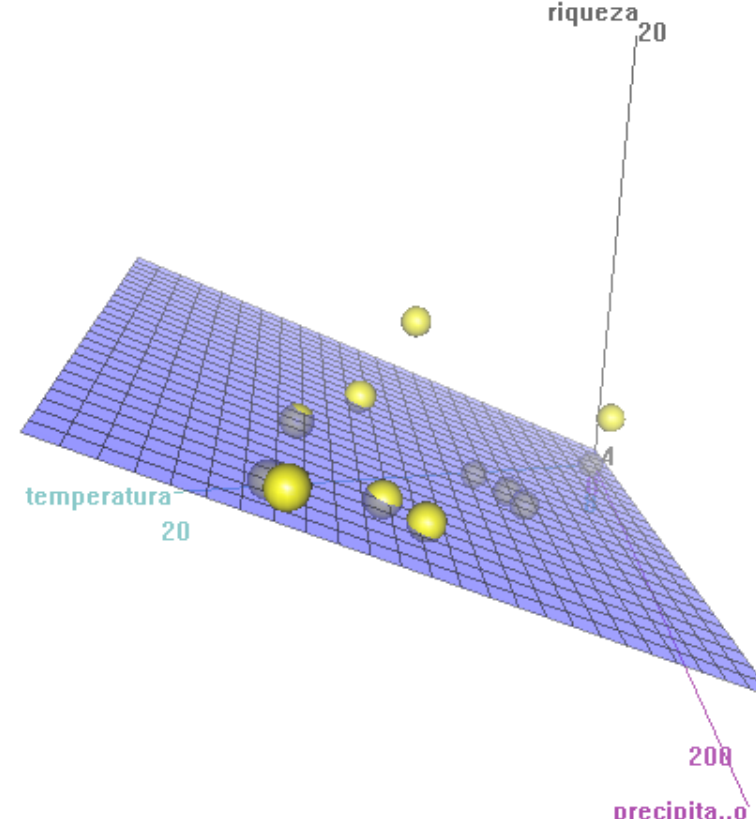
Resultados

Abundância



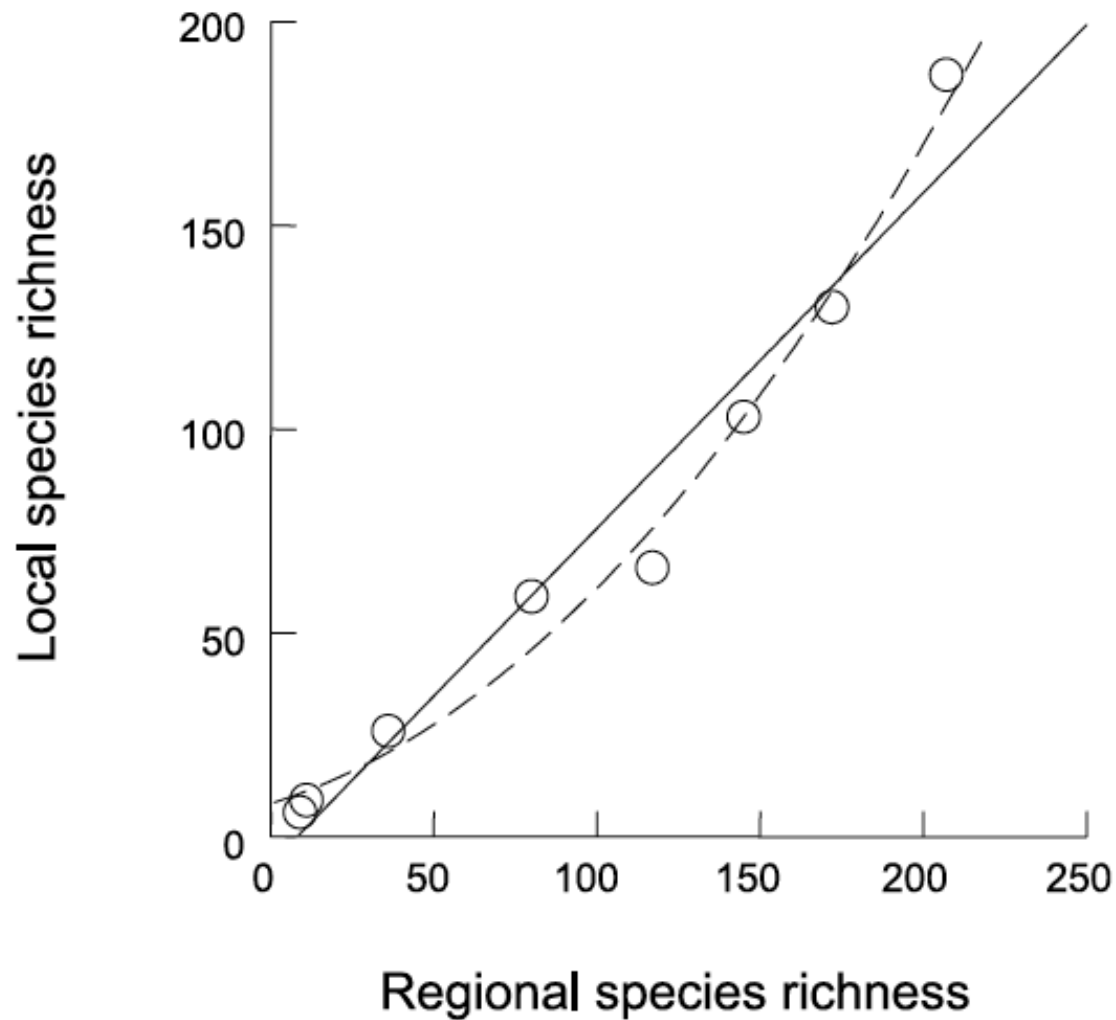
$(r^2=0.516; F_{(2,9)}=4.811;$
 $P=0.07$

Riqueza



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

Temperatura do ar
afetou a riqueza



Regressão polinomial

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

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.

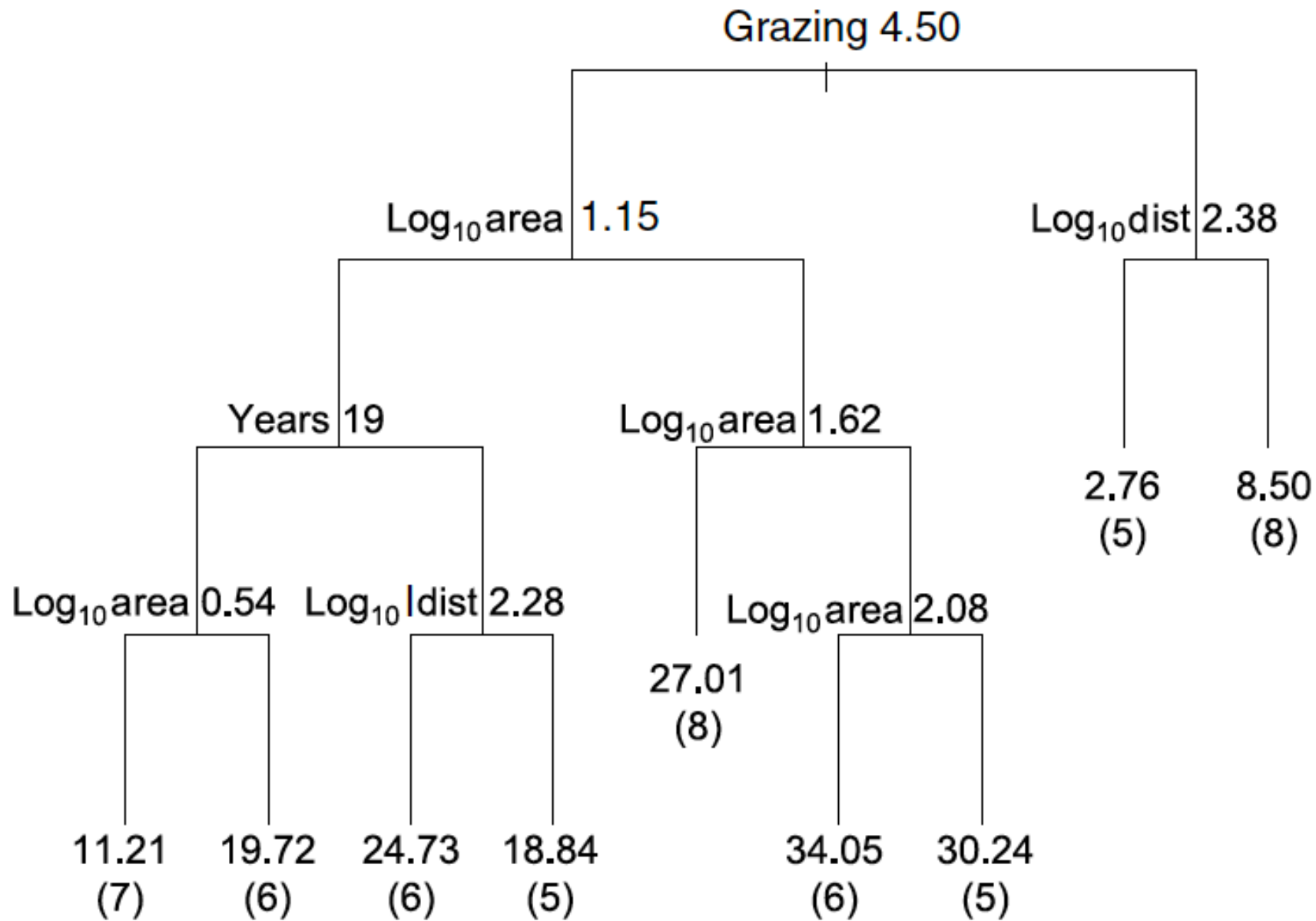
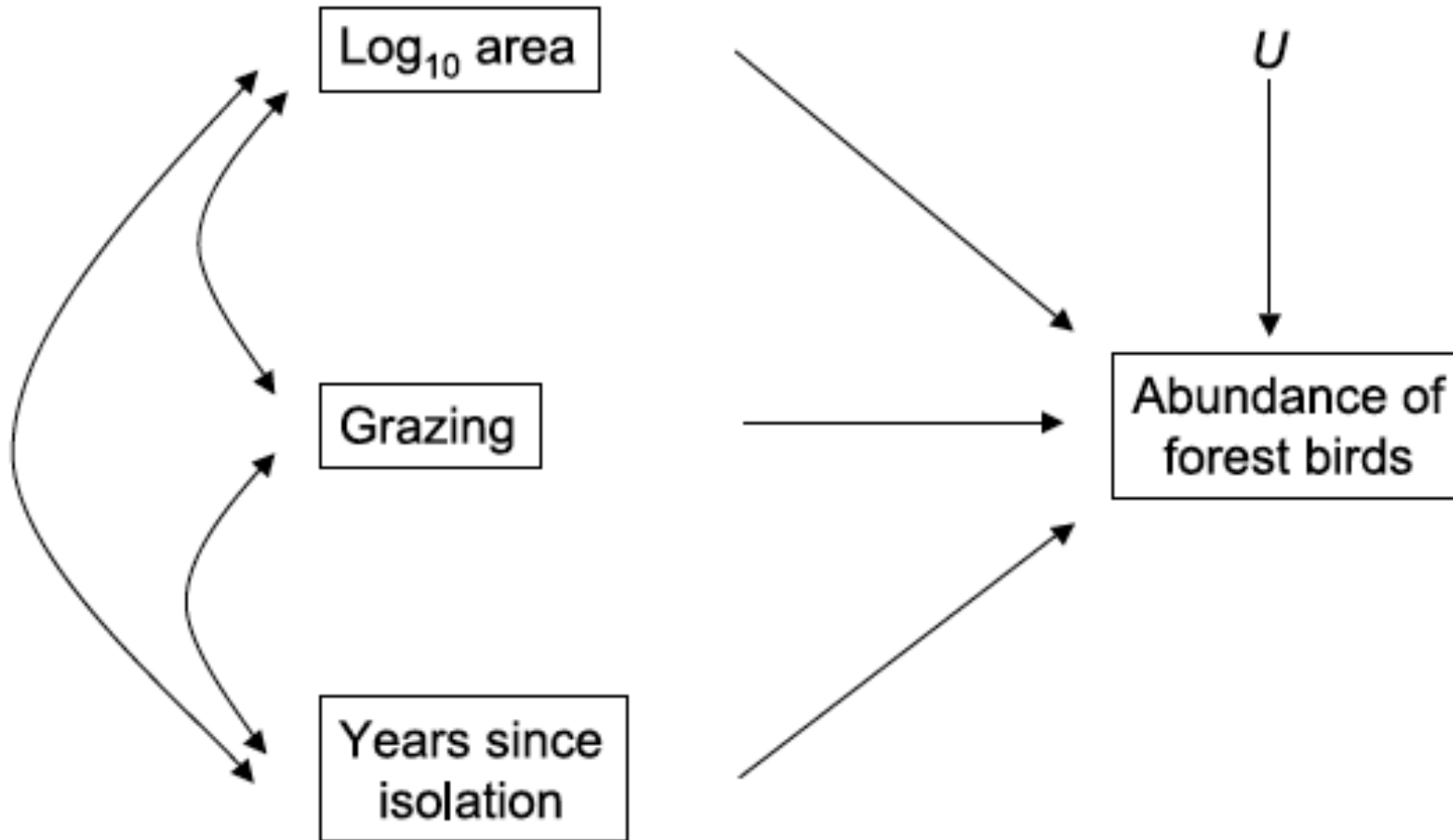


Figure 6.7 Regression tree based on the same data as in Figure 6.6, except that branching was continued to a lower level.

Árvore de regressão
 - Uma boa alternativa para regressão múltipla



Análise de rota
(Path analysis)

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).

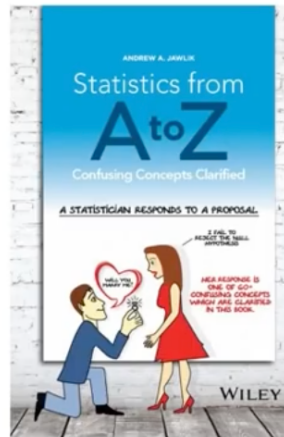
Channel: Statistics from A to Z – Confusing Concepts Clarified

ANOVA vs. Regression

Based on content from the book
Statistics from A to Z
-- Confusing Concepts Clarified

by Andrew A. Jawlik, published by Wiley

statisticsfromatoz.com



ANOVA vs Regression



Statistics from A to Z -- Confusing Concepts Clarified

Inscrito 367

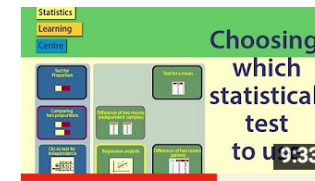
+ Adicionar a Compartilhar Mais

307 visualizações

7 0

Próximo

Reprodução automática



Choosing which statistical test to use - statistics help
Statistics Learning Centre
422.410 visualizações

https://www.youtube.com/watch?v=aG1G7_Ysl38