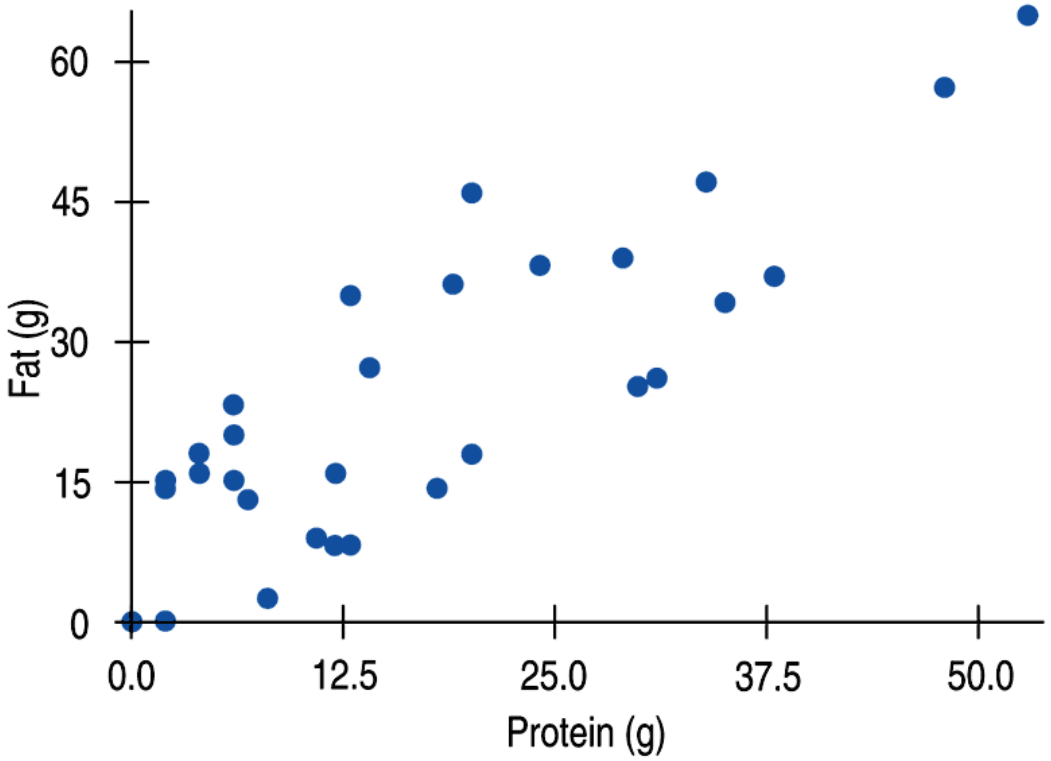
Regressão

Gordura x proteína

- O diagrama de dispersão apresenta o total de *gordura* versus *proteína* em uma pesquisa com trinta sanduíches.



O modelo linear

- A correlação no exemplo é de 0.83.
- Podemos dizer mais sobre uma relação linear entre duas variáveis quantitativas com um modelo.
- Um modelo simplifica a realidade e ajuda-nos a entender padrões e relações existentes.

O modelo linear

- O modelo linear é apenas uma equação de uma linha reta através dos dados.
  - Os pontos no diagrama não estão totalmente alinhados, mas uma linha estreita pode sumarizar o padrão geral.
  - O modelo linear pode ajudar-nos a entender como os valores estão associados.

Resíduos

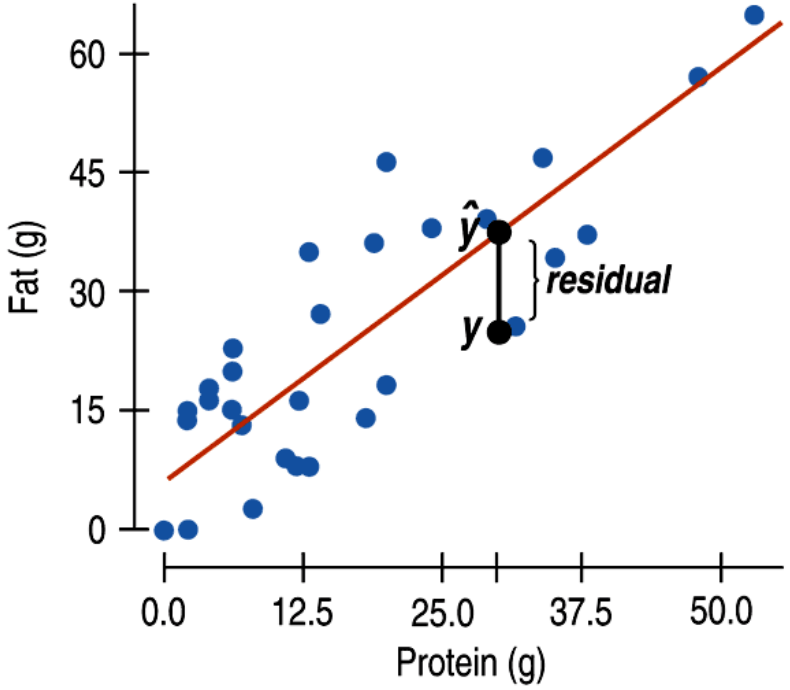
- O modelo não será perfeito, independentemente da linha que seja traçada.
- Alguns pontos estarão acima da linha, outros estarão abaixo.
- A diferença entre o valor observado e o valor predito é chamado de resíduo.

O "melhor ajuste" dos mínimos quadrados

- Alguns resíduos são negativos, outro são positivos; em média, um cancela o outro.
- Similar para o que fizemos para os desvio-padrão, elevamos os desvios ao quadrado e somamos.
- Quanto menor a soma, melhor é o ajuste.
- A linha de melhor ajuste é a a linha dos mínimos quadrados.

Resíduos

- Um resíduo negativo significa que o valor predito é muito grande (superestimado).
- Um resíduo negativo significa que que o valor predito é muito pequeno (subestimado).
- Na figura o valor estimado é 36 g, enquanto o verdadeiro valor de gordura é 25 g, então o resíduo é – 11 g de gordura.



A linha da regressão

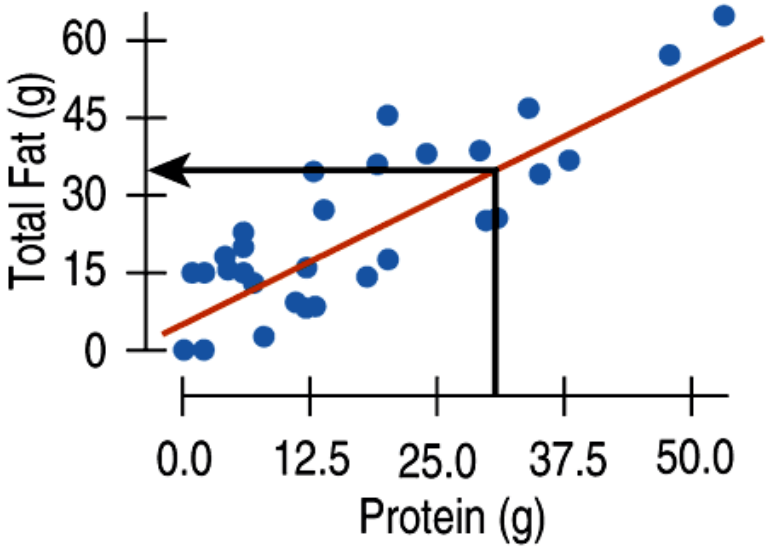
$$y = \alpha + \beta x$$

- β e α são a inclinação e o intercepto da linha.
- β é a inclinação, que iguala a mudança em y, com o aumento de uma unidade em x.
- α é o intercepto-y, que nos diz onde a linha atravessa (intercepta) o eixo-y.

Gordura x proteína

- A linha de regressão do banco pode ser vista no gráfico ao lado.
  - A equação é:

$$\widehat{fat} = 6.8 + 0.97\,protein.$$



O conteúdo de gordura predito para sanduíche com 30 g de proteína é 6.8 + 0.97(30) = 35.9 gramas de gordura.

Exemplo: nível de desenvolvimento e emissões de CO2

- *y* = emissões de dióxido de carbono
  (per capita, em metric tons)

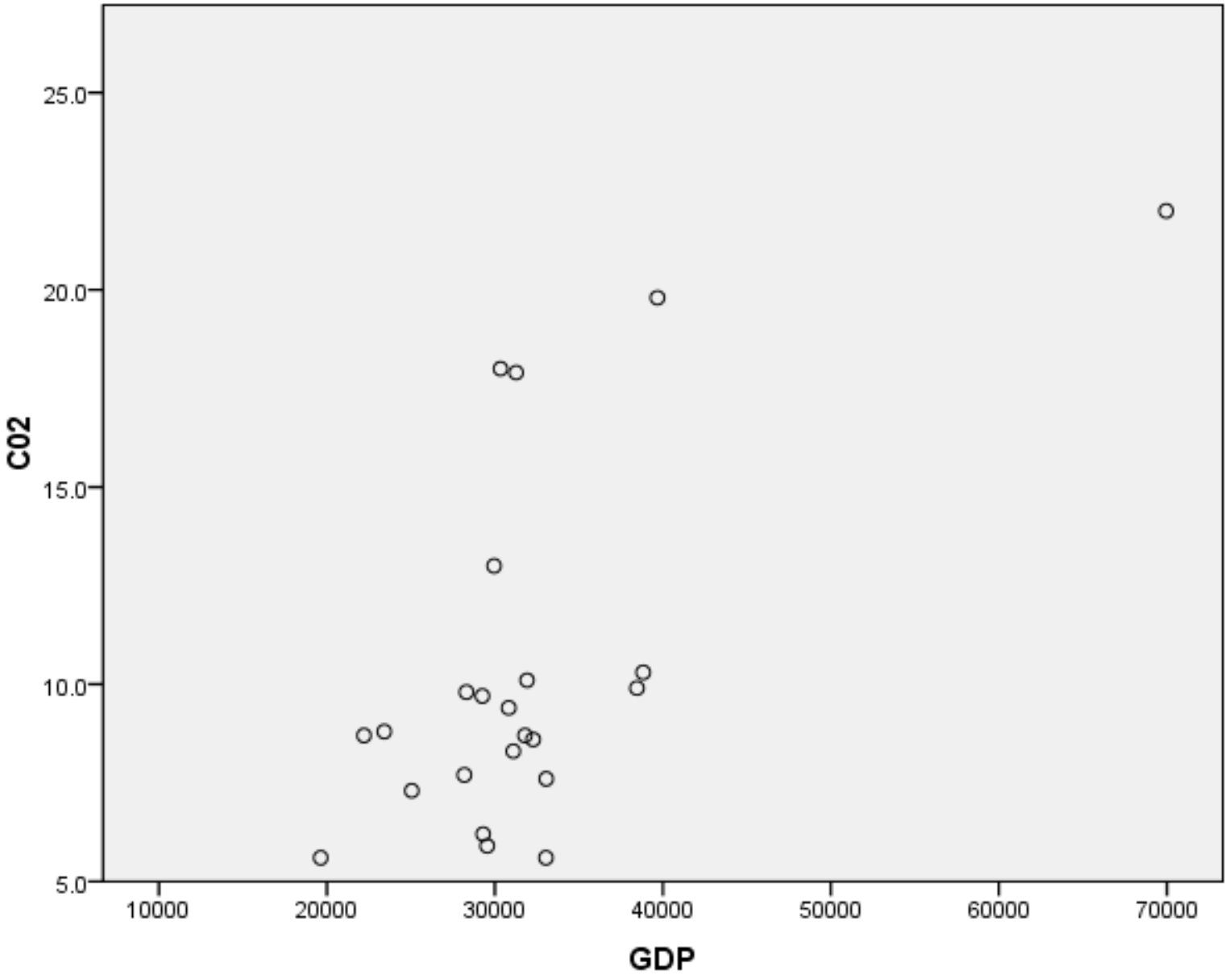
varia de 5.6 em Portugal e 22.0 em Luxemburgo (U.S. = 19.8)

média = 10.4 desvio padrão = 4.6

- *x* = produto interno bruto; gross domestic product
  (GDP, em milhares de dólares per capita)

varia de 19.6 em Portugal para 70.0 in Luxemburgo (U.S. = 39.7)

média = 32.1 desvio padrão = 9.6


C02
25.0
20.0
15.0
10.0
5.0
10000
20000
30000
40000
50000
60000
70000
GDP

A equação da relação entre x e y

$$y = 0.42 + 0.31x$$

- Quando $x = 0$, o nível de CO2 predito é $y = 0.42 + 0.31x = 0.42 + 0.31(0) = 0.42$ (irrelevante, porque o valor de nenhum GDP está próximo de 0)
- Quando $x = 39.7$ (valor para U.S.A), o nível do CO2 predito é $y = 0.42 + 0.31(39.7) = 12.7$ (real= 19.8 for U.S.)
- Para cada aumento de mil dólares em renda em GDP capita, o CO2 predito aumenta por 0.31 metric tons per capita
- Mas, a equação linear é apenas uma aproximação. A correlação entre *x* e *y* para as nações é 0.64, não 1.0

Resíduos

- O modelo linear pressupõe que a relação entre as duas variáveis é uma linha perfeita. Os resíduos são parte dos dados que não são modelados.

Dados = Modelo + Resíduo

ou

Resíduo = Dados – Modelo

R^2— o coeficiente de determinação

- Se a correlação fosse 1.0 e o modelo predissesse os valores da gordura perfeitamente os resíduos seriam zero e não teriam variação.
- Como a correlação é de 0.83, não temos perfeição.
- Nós podemos determinar quanto da variação é derivada do modelo, quanto é determinada pelos resíduos.

R^2— o coeficiente de determinação

- Todas as análises de regressão incluem o R^2 (pronuncia-se "R-quadrado"). Um R^2 de 0 significa que a nada da variância dos dados está no modelo; tudo está nos resíduos.
- Quando interpretar um modelo de regressão é fundamental saber o significado do R^2.
  - No exemplo dos sanduiches, 69% da variação da gordura total é derivada da variação no conteúdo da proteína.

Reportar o R^2

- Além da inclinação e do intercepto é fundamental reportar o R^2 o que permite que os leitores possam julgar o sucesso da regressão em se ajustar aos dados.
- Estatística é sobre variação e o R^2 mede o sucesso do modelo de regressão em termos de fração da variação de y explicada pela regressão.

Premissas para a Regressão

- Variáveis quantitativas:
  - Regressão pode apenas ser aplicada com duas variáveis quantitativas (e não duas variáveis categóricas)
- Linearidade:
  - O modelo pressupõe que a relação entre as variáveis seja linear.
  - Um diagrama de dispersão mostrará se esta premissa é razoável.

Premissas para a Regressão

- Atenção ao outlier:
  - Um outlier pode mudar dramaticamente o modelo da regressão.
  - Outliers podem inclusive mudar o sinal da linha de regressão.