Regressão Linear

Ronald Targino Nojosa

DEMA-UFC

Notas de Aula - Parte 1
Versão Parcial

1 Regressão Linear Simples
- Introdução
- Declaração do Modelo
- Características do modelo
- Interpretação dos parâmetros
- Estimação da função de regressão
- Resíduos

Regressão Linear Simples

Introdução

- **Análise de Regressão**: técnica estatística para investigar e modelar o relacionamento funcional entre variáveis

- Origem
  - Termo "reversion": Sir Francis Galton, palestra em 1877
  - Termo "regression": Sir Francis Galton, artigo de 1885[1]
  - Galton estudou a relação entre as alturas de pais e filhos:

Altura	
Pais	Filhos
muito altos	em média, mais baixos
muito baixos	em média, mais altos

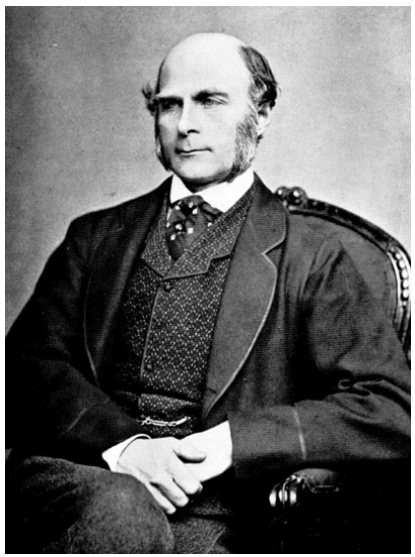
  - Descreveu o fenômeno como **regressão à mediocridade**
  - Galton[2] era eugenista, assim como Karl Pearson e Ronald Fisher

[1] 1886 para alguns autores

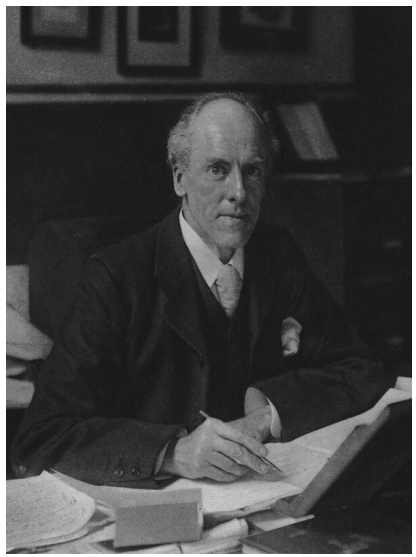
[2] o termo eugenia foi definido por Galton em 1883

- Primeiras referências
  - Galton, F. (1885). Regression towards mediocrity in hereditary stature, Journal of the Anthropological Institute, 15, 1886, 246-263.
  - Galton, F. (1885). Family likeness in stature. Procedures of Royal Society of the London, 40, 42–72.
  - Pearson, K. and Lee, A. (1903). On the laws of Inheritance. Biometrika, 2, 357–462

- Cabeções



Francis Galton
(1822-1911)



Karl Pearson
(1857-1936)

Ronald Fisher
(1890-1962)

- Aplicações
  - Economia
  - Biologia
  - Engenharia
  - Ciência de dados
  - Agronomia e diversas outras
- Terminologia
  - Variável(is) resposta(s) (dependente, explicada, saída)
  - Variável(is) explicativa(s) (independente, regressora, entrada)
  - As variáveis podem ser qualitativas ou quantitativas
- Modelo: depende do tipo e da quantidade de variáveis
- Modelo Matemático *versus* Modelo Estatístico

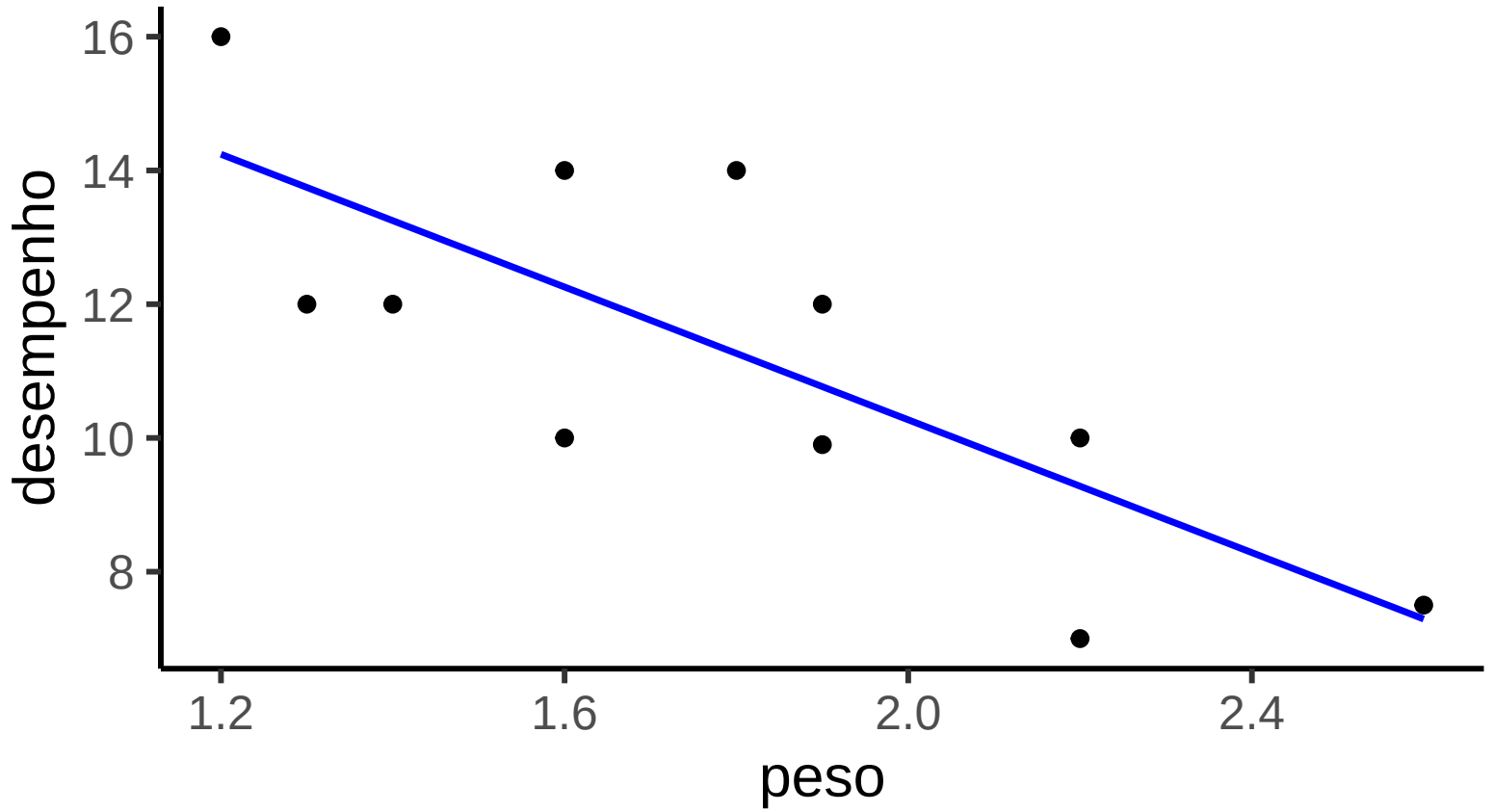
Exemplo 1

Qual a relação entre peso e desempenho de carros?

Tabela 1 - Peso (t) e desempenho (km/l) para uma amostra de 11 carros

peso	1.2	1.3	1.4	**1.6**	**1.6**	1.8	**1.9**	**1.9**	**2.2**	**2.2**	2.6
desemp.	16	12	12	14	10	14	12	9.9	10	7	7.5

Peso (t) e desempenho (km/l) para uma amostra de 11 carros

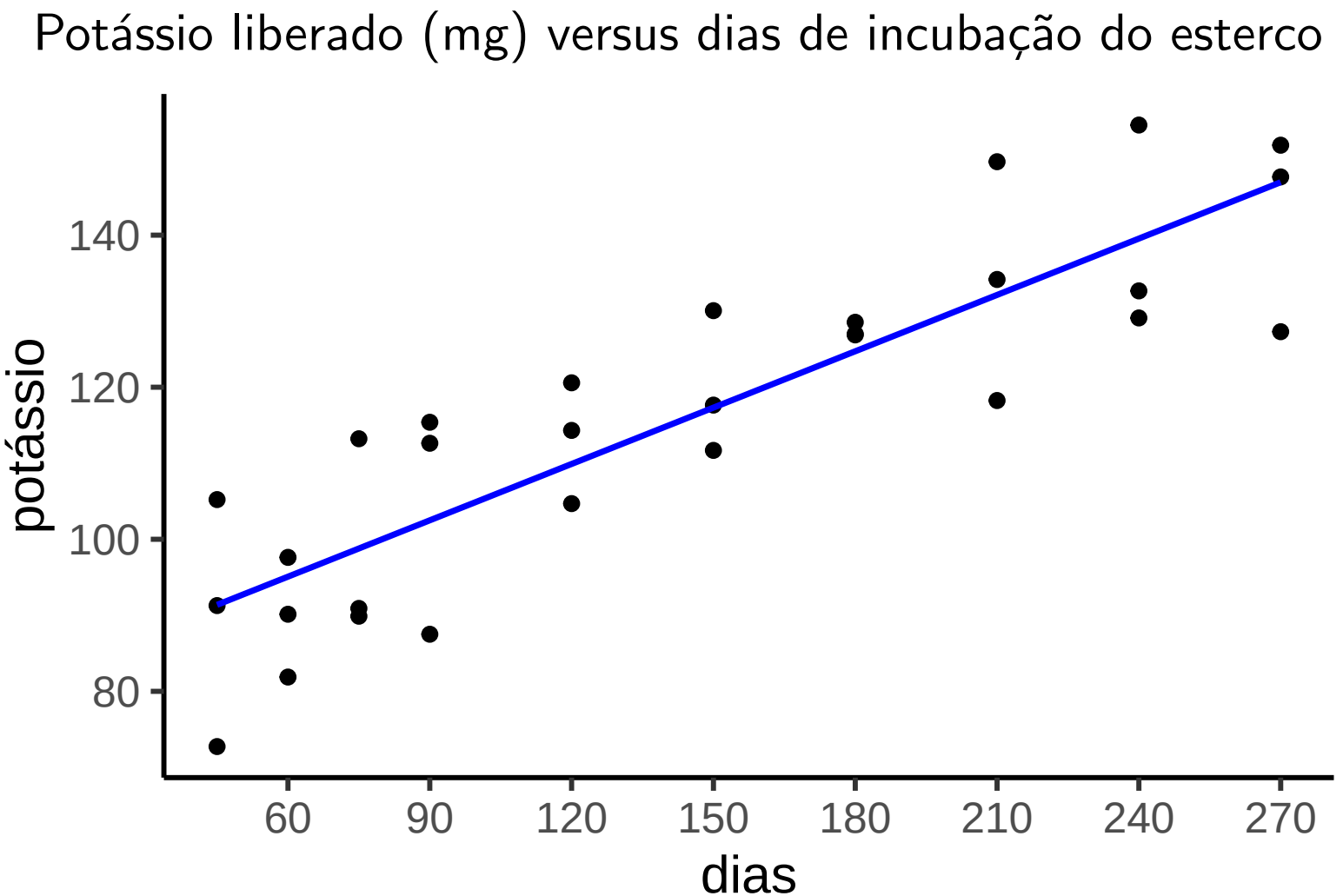


Exemplo 2 - caso 1

Como se dá o conteúdo liberado de potássio (mg) em função do tempo (dias) de incubação de esterco de codorna?

Tabela 2 - Potássio liberado (mg) versus dias de incubação do esterco

potássio	91.3	105.2	72.7	81.9	97.6	...	127.3
dias	45	45	45	60	60	...	270

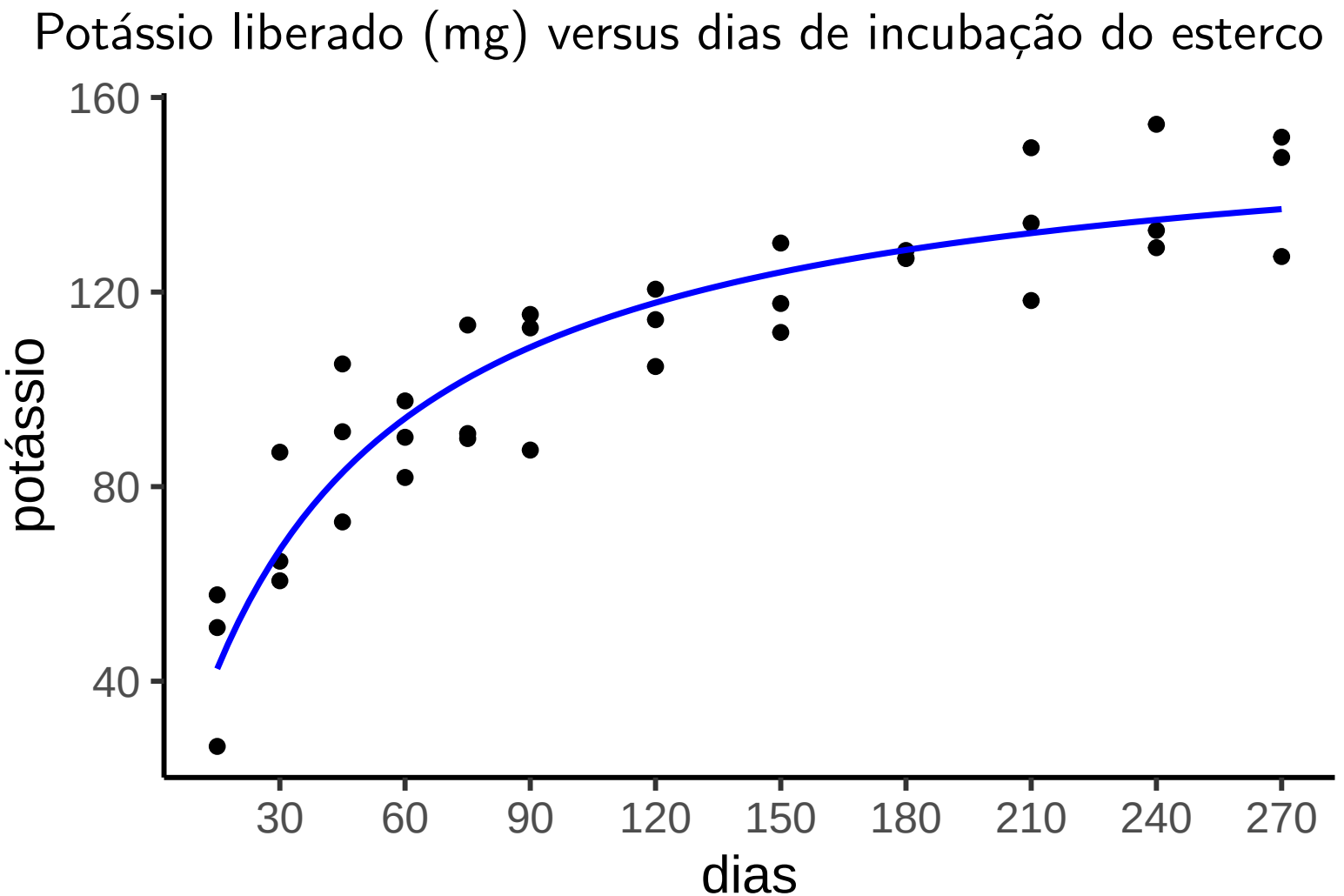

Potássio liberado (mg) versus dias de incubação do esterco
potássio
140
120
100
80
60
90
120
150
180
210
240
270
dias

Exemplo 2 - caso 2

Como se dá o conteúdo liberado de potássio (mg) em função do tempo (dias) de incubação de esterco de codorna?

Tabela 3 - Potássio liberado (mg) versus dias de incubação do esterco

potássio	51.0	57.7	26.6	60.7	87.1	64.7	91.3	...	127.3
dias	15	15	15	30	30	30	45	...	270

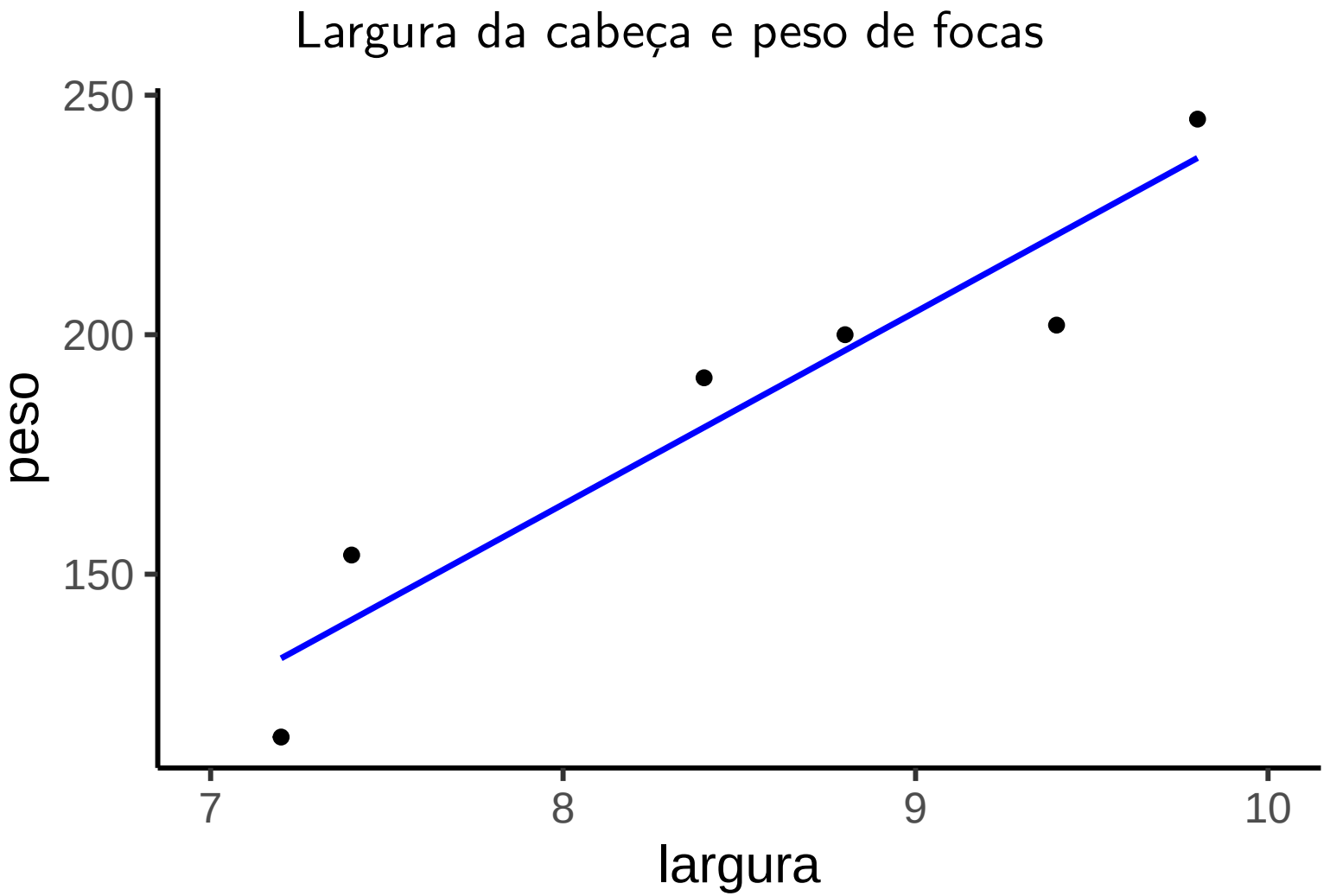

Potássio liberado (mg) versus dias de incubação do esterco
potássio
160
120
80
40
30
60
90
120
150
180
210
240
270
dias

Exemplo 3

Como estimar, usando fotografias, o peso de focas a partir da largura da cabeça?

Tabela 4 - Largura da cabeça e peso de focas

largura	7.2	7.4	9.8	9.4	8.8	8.4
peso	116	154	245	202	200	191


Largura da cabeça e peso de focas
250
200
150
peso
7
8
9
10
largura

Observações:

- Em geral, para variáveis (resposta e explicativa) com correlação significativa, teremos bom ajuste com o modelo de regressão linear simples
- Relação estatística entre variáveis não implica causa e efeito
  - conclusão não pode se basear apenas na amostra considerada
  - conclusão com base em teoria ou conhecimento empírico
  - Exemplo: despesas de consumo pessoal e renda pessoal disponível (invocar a teoria econômica); ganho de peso e consumo de calorias (estudos observacionais e experimentais)

Exercícios: Pesquisem, leiam, estudem, registrem, informe as referências dos seus registros:

- Coeficiente de correlação linear de Pearson: aplicação, suposições, cálculo, interpretações
- Correlação espúria
- Critério da parcimônia
- Cuidados com previsões e extrapolações

Declaração do Modelo

Declaração do Modelo

Dados n pares de valores de duas variáveis X e Y e admitindo que Y é uma função linear de X, podemos estabeler uma Regressão Linear Simples (RLS) cujo modelo estatístico é

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1,\ 2, \ldots,\ n, \tag{1}$$

em que

- β_0 e β_1 são parâmetros
- β_0: coeficiente linear
- β_1: coeficiente angular
- Y_i: variável resposta
- X_i: variável explicativa
- ϵ_i: erro aleatório (**fonte de variação**; variável latente)

Observações:

- $Y_1, Y_2, \ldots, Y_n$ não são variáveis iid
- Podemos escrever:
  - $Y_i = \mu(X_i; \boldsymbol{\beta}) + \epsilon_i$ (modelo aditivo)
  - $Y_i = \mu(X_i; \boldsymbol{\beta}) \times \epsilon_i$ (modelo multiplicativo)
- $\mu(X_i; \boldsymbol{\beta})$:
  - denominada de **função de regressão**
  - função que descreve a forma funcional entre Y e X
  - a definição da forma funcional e da fonte de variação (aditiva ou multiplicatica) dependem de vários fatores
  - A função de regressão do modelo (1) é *simples*, *linear nos parâmetros e nas variáveis preditoras* e de *primeira ordem*[3].

[3] a ordem refere-se ao valor da maior potência da variável preditora

Suposições para o MRLS

1. A função de regressão é linear nos parâmetros[4].
2. Os valores de X são fixos
3. $E(\epsilon_i) = 0$
4. Homocedasticidade: $V(\epsilon_i) = E(\epsilon_i^2) = \sigma^2$
5. Fontes de variação não correlacionadas: $Cov(\epsilon_i, \epsilon_j) \underset{\forall i \neq j}{=} E(\epsilon_i \epsilon_j) = 0$
6. $\epsilon_i \sim N(0, \sigma^2)$ [5]
7. $Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$ [6]

Nota: $E(\epsilon_i) = E(\epsilon_i / X_i)$,
$V(\epsilon_i) = V(\epsilon_i / X_i) = V(Y_i / X_i)$, $\forall i = 1, 2, \ldots, n$.
As suposições 6 e 7 serão necessárias para IC e TH.

[4] para cada unid. de mudança na var. explicativa, a mudança correspondente na var. resposta é constante

[5] sob normalidade multivariada e as suposições (3) a (5), os erros são *iid*

[6] sob a suposição (6) e sendo Y_i uma função linear de ϵ_i, os Y_i's são v.a. normais independentes

Características do modelo

Características do modelo

Modelo: $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad i = 1,\ 2, \ldots,\ n.$

- A resposta Y_i é a soma de dois componentes: o termo constante e o aleatório (ϵ_i)
- Sob as suposições, temos:

$$E(Y_i) = E(Y_i/X_i) = E(\beta_0 + \beta_1 X_i + \epsilon_i) = \beta_0 + \beta_1 X_i$$

$$V(Y_i) = V(Y_i/X_i) = E[(Y_i - E(Y_i))^2] = E[\epsilon_i^2] = \sigma^2$$

- A **função de regressão** do modelo é

$$E(Y_i) = E(Y_i/X_i) = \beta_0 + \beta_1 X_i$$

.

- Estimaremos $E(Y_i)$ usando os métodos de mínimos quadrados (MMQ) e de máxima verossimilhança (MMV).

Interpretação dos parâmetros

Modelo: $Y = \beta_0 + \beta_1 X + \epsilon$

Função de regressão: $E(Y/X) = \beta_0 + \beta_1 X$

Interpretação dos parâmetros

- β_0: valor esperado[7] de Y para X=0

$$E(Y/X = 0) = \beta_0$$

- β_1: variação na esperança de Y para a variação de 1 unidade de X

$$\beta_1 = E(Y/X = x + 1) - E(Y/X = x)$$

[7] valor sem interpretação prática quando não fizer sentido considerar X=0 ou quando $0 \in [min\{X_1, \ldots, X_n\}, máx\{X_1, \ldots, X_n\}]$. Nestes casos, a centralização da variável X permitirá interpretar β_0.

Estimação da função de regressão

Método dos Mínimos Quadrados (MMQ)

O método consiste em estimar os parâmetros β_0 e β_1 através da minimização da soma dos quadrados dos erros:

$$S = \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} [Y_i - (\beta_0 + \beta_1 X_i)]^2 \tag{2}$$

Os **estimadores de mínimos quadrados (EMQ)**, $\hat{\beta}_0$ e $\hat{\beta}_1$, dos parâmetros β_0 e β_1 são dados pela solução do sistema de equações:

$$\frac{\partial S}{\partial \beta_0}\Big|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0$$

$$\frac{\partial S}{\partial \beta_1}\Big|_{\beta_0=\hat{\beta}_0, \beta_1=\hat{\beta}_1} = -2 \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) X_i = 0$$

Segue:

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \quad (3)$$

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = 0 \quad (4)$$

As equações (3) e (4) são chamadas **equações normais**[8]. Veja "The Geometry of Last Squares", Drapper and Smith (1988, chap.20).

[8] normal se refere ao conceito de perpendicularidade/ortogonalidade

Estimação da função de regressão

De (3), temos:

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{5}$$

Usando (5) em (4), temos:

$$\sum_{i=1}^{n} X_i Y_i - \hat{\beta}_0 \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0 \Rightarrow$$

$$\sum_{i=1}^{n} X_i Y_i - (\overline{Y} - \hat{\beta}_1 \overline{X}) \sum_{i=1}^{n} X_i - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0 \Rightarrow$$

$$\sum_{i=1}^{n} X_i Y_i - n\overline{Y}\ \overline{X} + n\hat{\beta}_1 \overline{X}^2 - \hat{\beta}_1 \sum_{i=1}^{n} X_i^2 = 0 \Rightarrow$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{Y}\ \overline{X}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} \tag{6}$$

Exercícios

Notação para somatórios

$$S_{YY} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$$

$$S_{XX} = \sum_{i=1}^{n} (X_i - \overline{X})^2$$

$$S_{XY} = \sum_{i=1}^{n} (X_i - \overline{X})(Y_i - \overline{Y})$$

Exercícios 1.1.

1. *Mostre que:*
   a. $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$
   b. *A equação (6)* equação *pode ser escrita como*
   $$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$
   c. $\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y}) = \sum_{i=1}^{n}(X_i - \overline{X})Y_i$
   d. $\sum_{i=1}^{n}(X_i - \overline{X})^2 = \sum_{i=1}^{n}(X_i - \overline{X})X_i$

Exercícios 1.2.

1. *Mostre que:*
   a. $\hat{\beta}_0$ *pode ser escrito como uma combinação linear dos* Y_i, $i = 1,\ 2,\ \ldots,\ n$, *ou seja,* $\hat{\beta}_0 = \sum_{i=1}^{n} v_i Y_i$, *em que* $v_i = \dfrac{1}{n} - \dfrac{(X_i - \overline{X})\overline{X}}{S_{XX}}$.
   b. $\hat{\beta}_1$ *pode ser escrito como uma combinação linear dos* Y_i, $i = 1,\ 2,\ \ldots,\ n$, *ou seja,* $\hat{\beta}_1 = \sum_{i=1}^{n} w_i Y_i$, *em que* $w_i = \dfrac{(X_i - \overline{X})}{S_{XX}}$.
   c. $\sum_{i=1}^{n} v_i = 1$ *e* $\sum_{i=1}^{n} v_i X_i = 0$
   d. $\sum_{i=1}^{n} v_i = 0$ *e* $\sum_{i=1}^{n} w_i X_i = 1$
2. *Para* $x_1, x_2, \ldots, x_n$, *quaisquer números reais, prove*[a] *que* $\min_a \sum_{i=1}^{n}(x_i - a)^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$.

[a] dica: $(x_i - a) = (x_i - a + \overline{x} - \overline{x})$

Resultados importantes

i. Se U e R são variáveis aleatórias e a, b, c e d, constante:

- $Cov(U,R) = E[(U-E(U))(R-E(R))] = E(UR) - E(U)E(R)$
- $Cov(aU+c, bR+d) = abCov(U,R)$
- $Cov(U,U) = E(U^2) - [E(U)]^2 = V(U)$
- $V(aU \pm bR) = a^2V(U) + b^2V(R) \pm 2abCov(U,R)$
- $Cov(U-R, U+R) = Cov(U,U) + Cov(U,R) - Cov(R,U) - Cov(R,R) = V(U) - V(R)$

ii. $Cov(\sum_{i=1}^{n} a_i U_i, \sum_{i=1}^{n} b_i U_i) = \sum_{i=1}^{n} a_i b_i V(U_i)$, a_i e b_i constantes, U_i variável aleatória, $i = 1,\ 2, \ldots,\ n$.

Exercícios 1.3.

1. *Mostre que:*
   a. $E(\hat{\beta}_1) = \beta_1$
   b. $V(\hat{\beta}_1) = \dfrac{\sigma^2}{S_{XX}}$
   c. $Cov(\overline{Y}, \hat{\beta}_1) = 0$
   d. $E(\hat{\beta}_0) = \beta_0$
   e. $V(\hat{\beta}_0) = \dfrac{\sigma^2}{n} + \dfrac{\overline{X}^2\sigma^2}{S_{XX}} = \sigma^2 \dfrac{\sum_{i=1}^n X_i^2}{nS_{XX}} = \dfrac{\sum_{i=1}^n X_i^2}{n} V(\hat{\beta}_1)$
   f. $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\overline{X} V(\hat{\beta}_1)$
   g. $E(\hat{Y}_i) = \beta_0 + \beta_1 X_i \quad e \quad V(\hat{Y}_i) = \sigma^2 \left[\dfrac{1}{n} + \dfrac{(X_i - \overline{X})^2}{S_{XX}}\right]$

Estimadores para

coeficiente linear: $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$

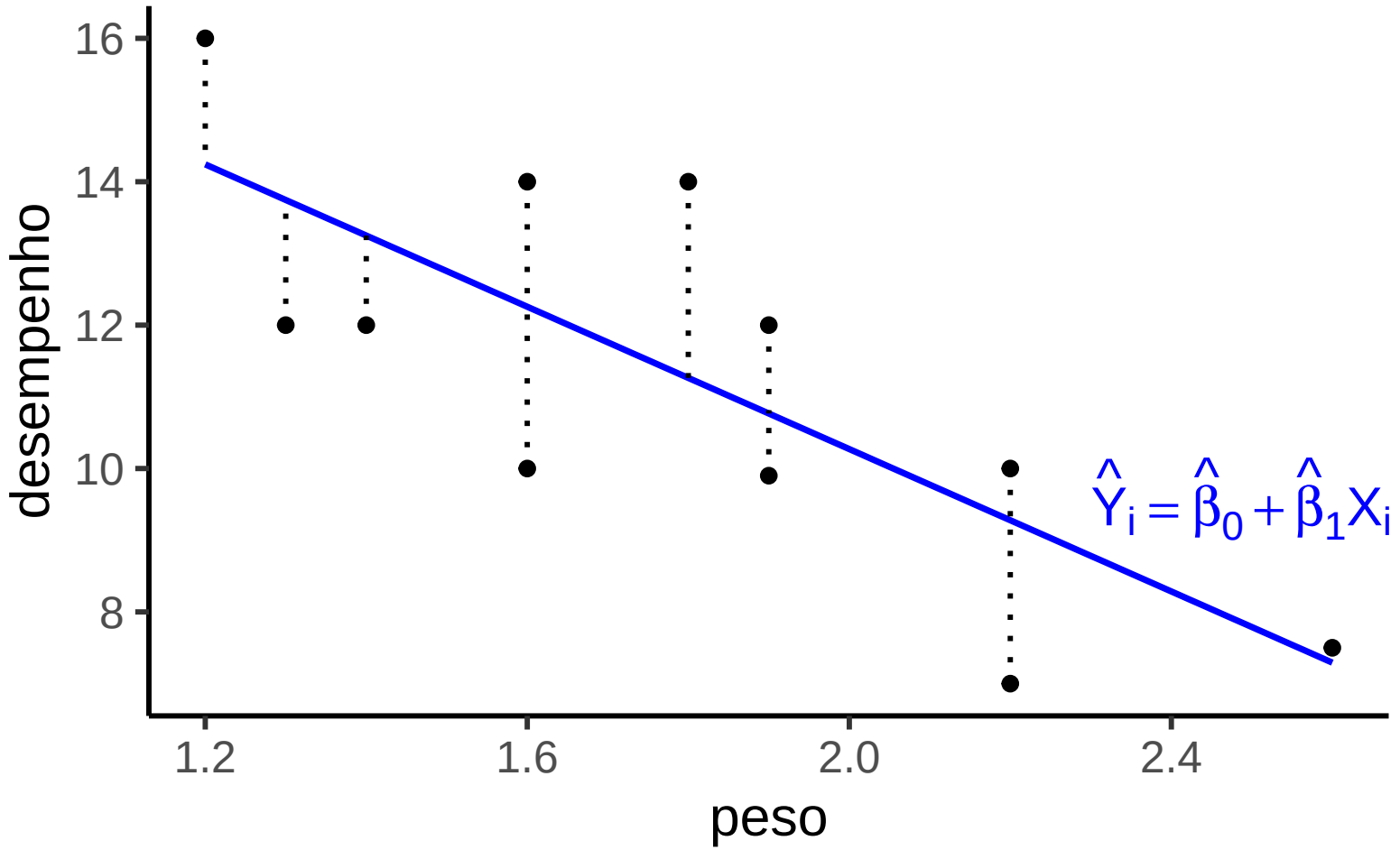
coeficiente angular: $\hat{\beta}_1 = \dfrac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \dfrac{S_{XY}}{S_{XX}}$

função de regressão[9]: $\hat{Y}_i = \widehat{E(Y_i)} = \widehat{E(Y_i/X_i)} = \hat{\beta}_0 + \hat{\beta}_1 X_i$

[9] modelo de RLS ajustado; **reta ajustada**

Estimação da função de regressão

Ilustração: A reta ajustada pelo MMQ é uma tal que a soma do quadrado de todas as discrepâncias verticais[10] seja mínima.



[10] **resíduos ordinários**; diferenças entre os valores observados e os respectivos valores ajustados (esperados) para Y

Propriedades dos EMQ

i. $\hat{\beta}_0$ e $\hat{\beta}_1$ são combinações lineares dos Y_i, $i = 1, 2, \ldots, n$:

$$\hat{\beta}_0 = \sum_{i=1}^{n} v_i Y_i \quad \text{e} \quad \hat{\beta}_1 = \sum_{i=1}^{n} w_i Y_i$$

ii. $E(\hat{\beta}_0) = \beta_0 \quad \text{e} \quad V(\hat{\beta}_0) = \dfrac{\sigma^2}{n} + \dfrac{\overline{X}^2 \sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

iii. $E(\hat{\beta}_1) = \beta_1 \quad \text{e} \quad V(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$

iv. $Cov(\hat{\beta}_0, \hat{\beta}_1) = -\overline{X} V(\hat{\beta}_1)$

Teorema de Gauss-Markov: Para o MRLS, considerando suas suposições de 1 a 5, os estimadores de mínimos quadrados são não viesados e têm variância mínima quando comparados com todos os outros estimadores não viesados que são combinações lineares dos $Y_i's$. Dizemos que são os melhores estimadores lineares não viesados (*Best Linear Unbiased Estimator - BLUE*).

Exercícios 1.4.

1. *Mostre que* $\hat{Y}_i = \overline{Y} + \hat{\beta}_1(X_i - \overline{X})$.
2. *Verifique que a reta de regressão ajustada contém o centro de gravidade (centróide) dos dados, isto é, o ponto* $(\overline{X}, \overline{Y})$ *está na reta de regressão.*
3. *Mostre que os estimadores de mínimos quadrados,* $\hat{\beta}_0$ *e* $\hat{\beta}_1$*, minimizam S* Equação (2) *:*
   a. *com base na matriz Hessina*[a] *de S*
   b. *com base*[b] *no exercício 2 dos Exercícios 1.2*

---

[a] matriz de derivadas de segunda ordem; veja condições para identificar o mínimio de funções de 2 variáveis

[b] dica: faça $[Y_i - (\beta_0 + \beta_1 X_i)] = [(Y_i - \beta_1 Xi) - \beta_0] = (z_i - \beta_0)$. Portanto, para qualquer valor β_1 fixado, o valor de β_0 que minimiza S é $\overline{z}$. Em S, substitua β_0 por $\overline{z}$ e verifique que $S = S_{YY} - 2\beta_1 S_{XY} + \beta_1^2 S_{XX}$. Mostre que o valor de β_1 que minimiza S é S_{XY}/S_{XX}.

Resíduos

Resíduos

A diferença entre o valor observado Y_i e o correspondente valor ajustado $\hat{Y}_i$ é chamado de **resíduo ordinário**:

$$e_i = Y_i - \hat{Y}_i, \quad i = 1, 2, \ldots, n.$$

Os resíduos serão importantes para investigar a adequação do modelo.

Resultados importantes

Das equações normais (3) e (4) temos:

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i) = 0 \Rightarrow \sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n} e_i = 0 \qquad (7)$$

$$\sum_{i=1}^{n}(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)X_i = 0 \Rightarrow \sum_{i=1}^{n}(Y_i - \hat{Y}_i)X_i = \sum_{i=1}^{n} X_i e_i = 0 \qquad (8)$$

De (7) e (8), concluímos que

$$\sum_{i=1}^{n} \hat{Y}_i e_i = 0. \qquad (9)$$

Exercícios 1.5.

1. *Mostre que* $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$.

2. *Verifique que a média dos valores observados é igual a média dos valores estimados:* $\overline{Y} = \overline{\hat{Y}}$.

O **resíduo**

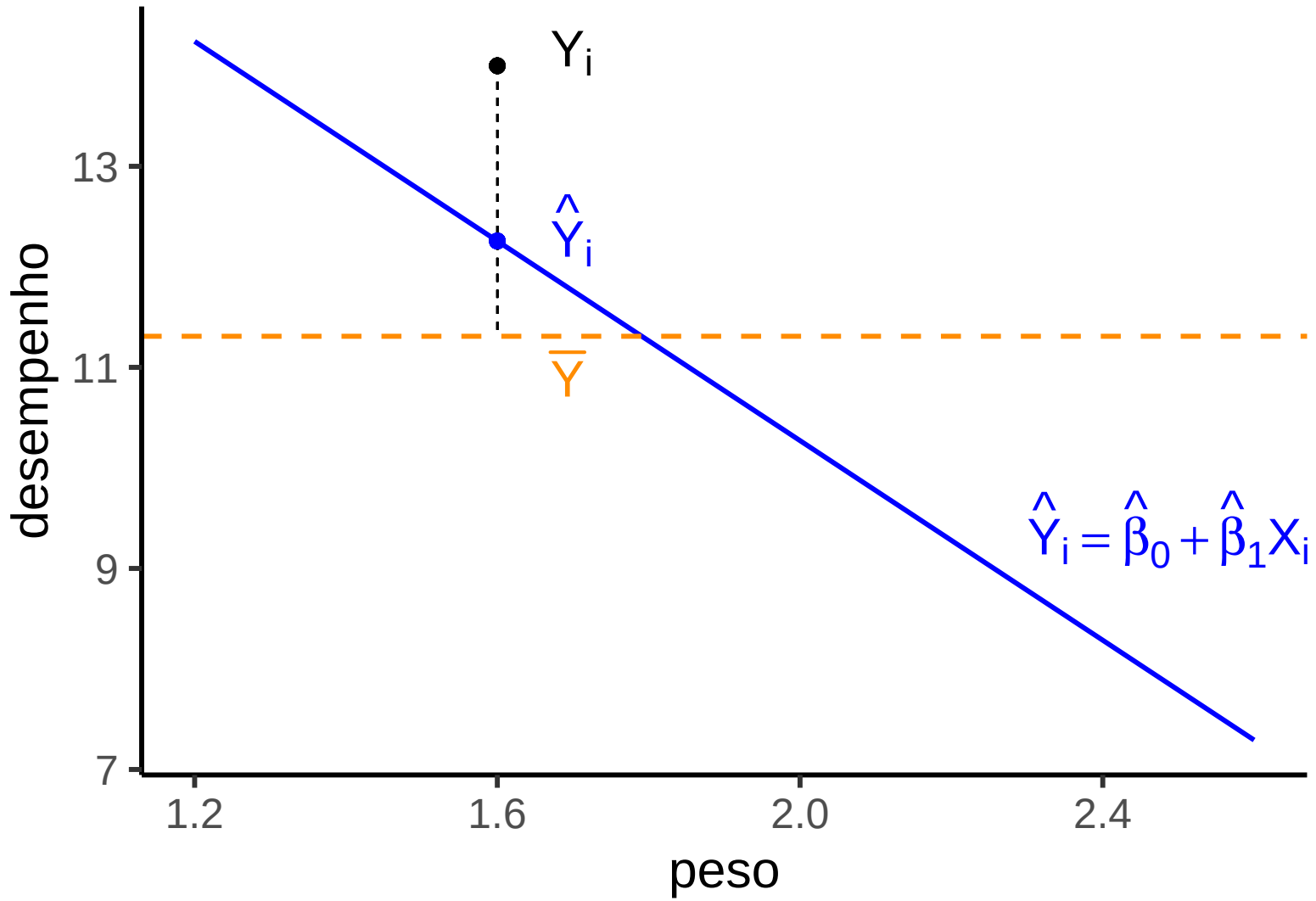
$$e_i = Y_i - \hat{Y}_i$$

pode ser visto pela diferença entre duas quantidades: o desvio de Y_i em relação à $\overline{Y}$ e o desvio de $\hat{Y}_i$ em relação a $\overline{Y}$, isto é,

$$Y_i - \hat{Y}_i = (Y_i - \overline{Y}) - (\hat{Y}_i - \overline{Y}). \qquad (10)$$

Resíduos

Peso (t) e desempenho (km/l) para uma amostra de 11 carros



Resíduos

Comentário: Se desconsideramos a relação entre Y e X, podemos[11] predizer valores de Y pela média aritmética de suas observações. Entretanto, se X afeta Y, os resíduos em relação a média aritmética serão, em geral, maiores do que em relação ao valor estimado pela reta de regressão.

Gráficos

[11] sendo $Y_1, Y_2, \ldots, Y_n$ iid, com $V(Y_i) = \sigma^2 < \infty$ e $E(Y_i) = \mu \in R$, $\overline{Y}$ é ENVVUM de μ