Introdução aos modelos de regressão normais lineares

Prof. Caio Azevedo

Introdução

- Estatística: área do conhecimento/Ciência que trata de metodologias (matemáticas) apropriadas para se coletar, organizar e analisar dados.
- A Estatística é uma ferramenta muito importante na resolução de problemas levantados pelas nas diversas áreas: Biologia, Psicometria, Educação, Medicina, Física, Computação entre outras.
- É importante que o Estatístico participe de todas as etapas de um estudo (pesquisa/consultoria).

Etapas para a resolução de um problema

1. Determinação do problema/objeto de estudo.
2. Determinação dos objetivos específicos.
3. Determinação do tamanho da amostra.
4. Execução do levantamento dos dados: entrevistas, experimento, coleta de dados etc.
5. Análise Descritiva.
6. Análise Inferencial (Modelos de regressão).
7. Conclusões e elaboração dos relatórios/artigos/trabalhos pertinentes.

Pode-se retornar a pontos anteriores ou mesmo avançar (pulando alguns), consoante a necessidade.

Introdução

- Foram vistas no Bacharelado, até o momento, diversas ferramentes de análise: descritiva, probabilística e inferencial.
- Estudaremos como analisar a influência de uma ou mais variáiveis (variáveis explicativas, covariáveis, variáveis explanatórias) em uma variável de interesse (variável resposta ou resposta).
- Nos focaremos nos modelos de regressão normal linear homocedásticos (simples/múltipla).

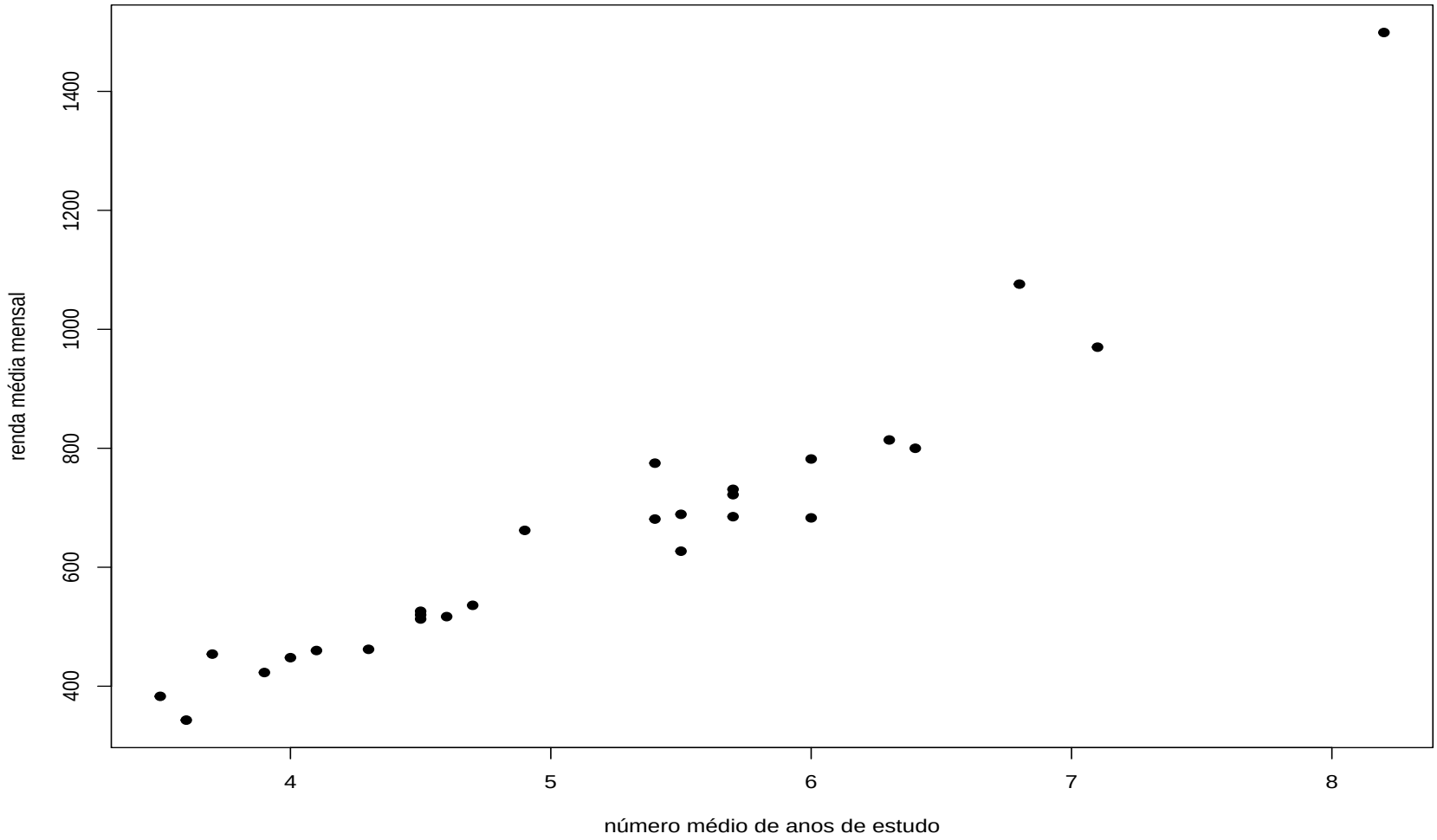
Exemplo introdutório: estudo entre renda e escolaridade

- O conjunto de dados foi extraído do censo do IBGE de 2000, apresenta para cada unidade da federação o número médio de anos de estudo (anos de estudo) e a renda média mensal em reais (renda) do chefe ou chefes do domicílio.
- Esses dados estão também armazenados no arquivo censo.txt

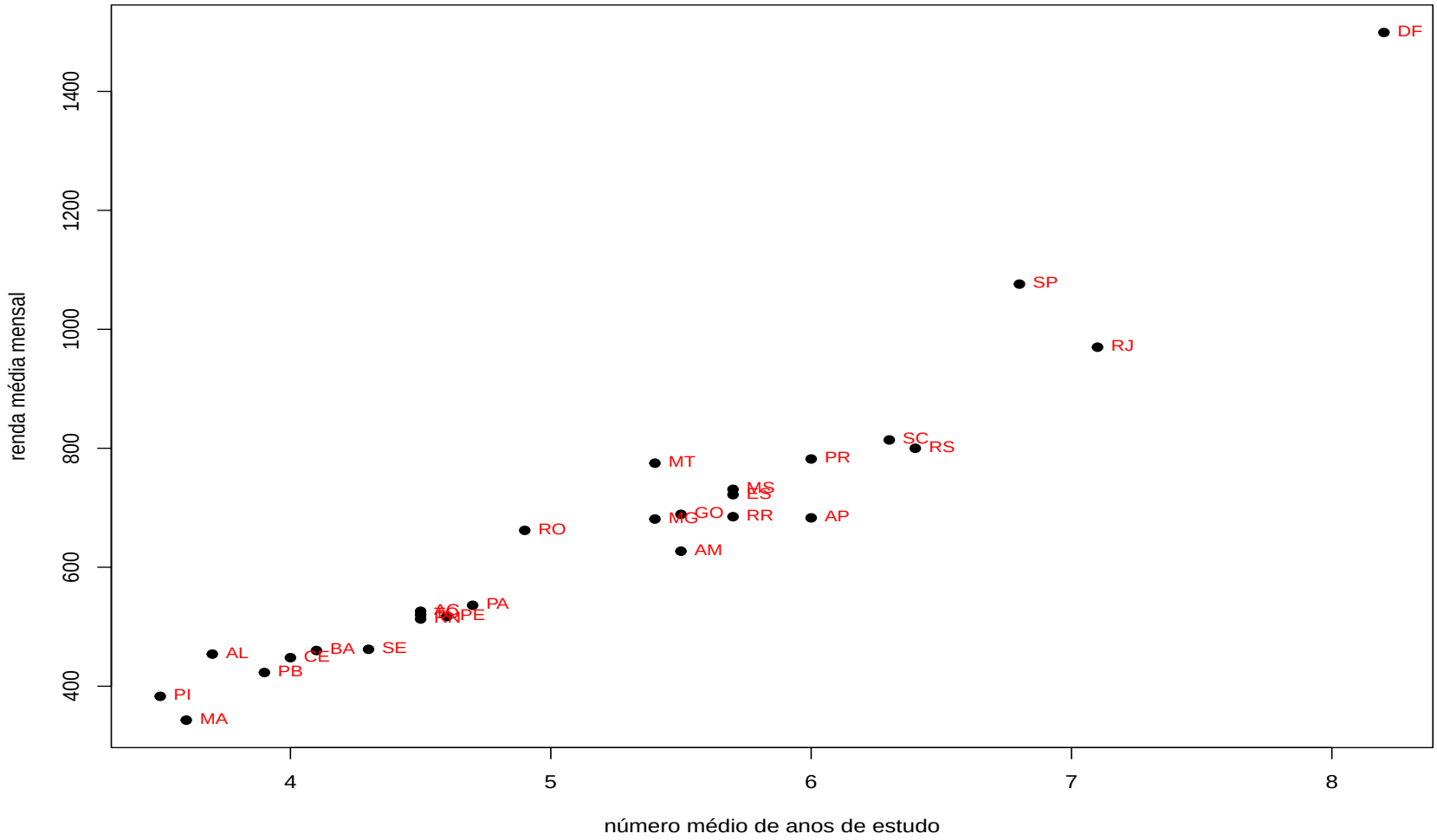
Exemplo introdutório: bancos de dados

ID	UF	anos de estudo	renda
1	RR	5,70	685
2	AP	6,00	683
3	AC	4,50	526
⋮	⋮	⋮	⋮
24	MT	5,40	775
25	GO	5,50	689
26	MS	5,70	731
27	DF	8,20	1499

Exemplo introdutório: gráfico de dispersão



Exemplo introdutório: gráfico de dispersão



Exemplo introdutório: estudo entre renda e escolaridade

- Como esperado, a correlação estimada é elevada e positiva. Mas isso não implica numa relação de causa e efeito, necessariamente.
- Ou seja, não é o fato de uma pessoa (chefes de família) ter muitos anos de estudo que a leva a ter uma renda elevada e vice-versa.
- Devem existir outros fatores (ocupação, condição financeira dos antepassados-herança, idade, local de residência, quantidade de chefes de família) que influenciem ambas as variáveis e que as façam estar positivamente relacionadas.
- Além disso, correlações altas (positivas ou negativas) podem ser espúrias (sem sentido).

Exemplo 0: altura e peso de homens e mulheres (espaços em branco) missing

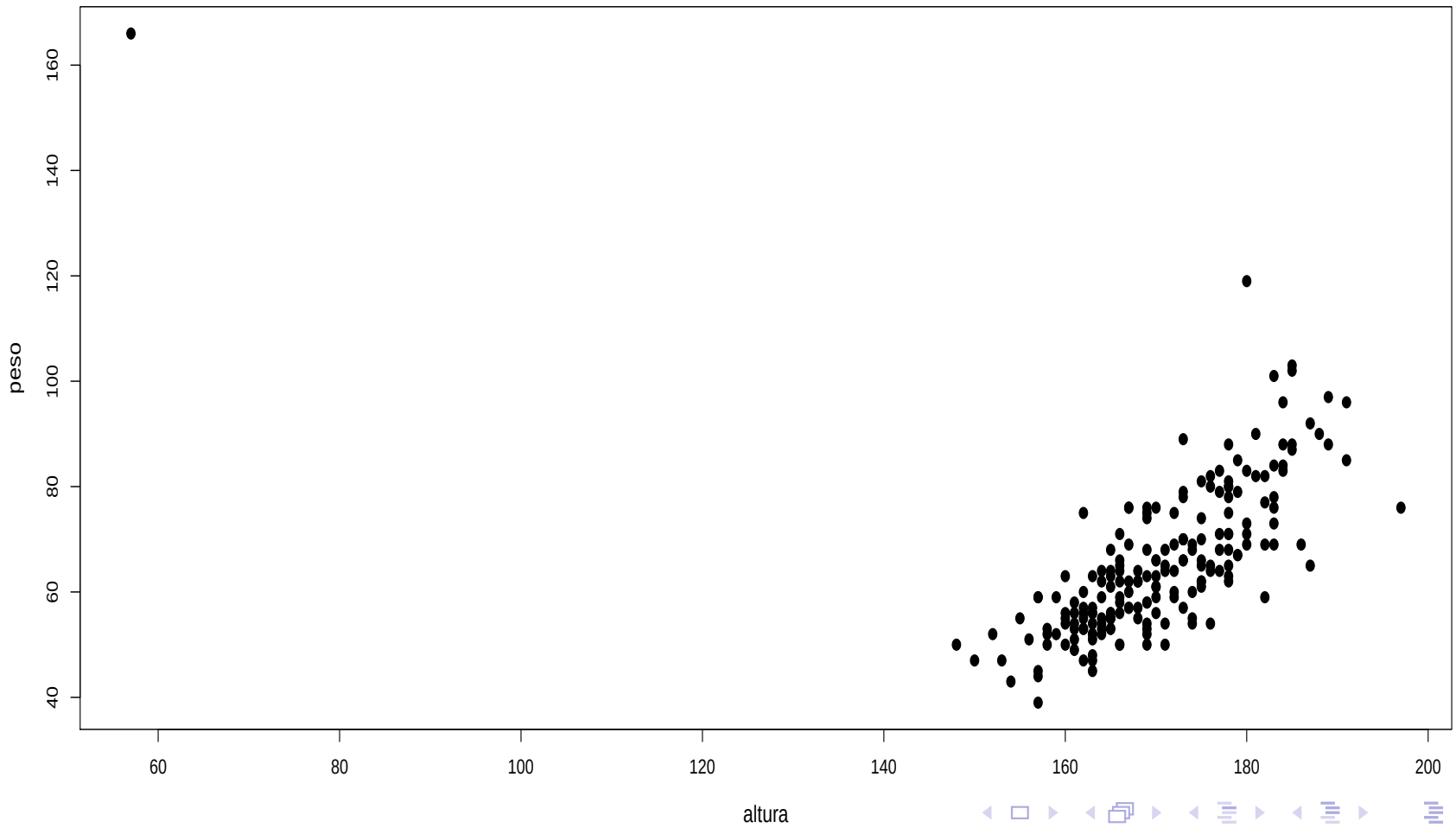
- Os dados correspondem aos pesos (em kg) e alturas (em cm) medidos e informados, de 200 indivíduos.
- O sexo de cada um também foi coletado, sendo 112 mulheres e 88 homens.
- Este conjunto de dados está disponível em no R no pacote "car" sob o nome "Davis".
- Consideraremos os valores medidos.

Exemplo 0: bancos de dados

	sex	weight	height	repwt	repht
1	M	77	182	77	180
2	F	58	161	51	159
3	F	53	161	54	158
⋮	⋮	⋮	⋮	⋮	⋮
47	M	73	180		
48	F	49	161		
199	M	90	181	91	178
200	M	79	177	81	178

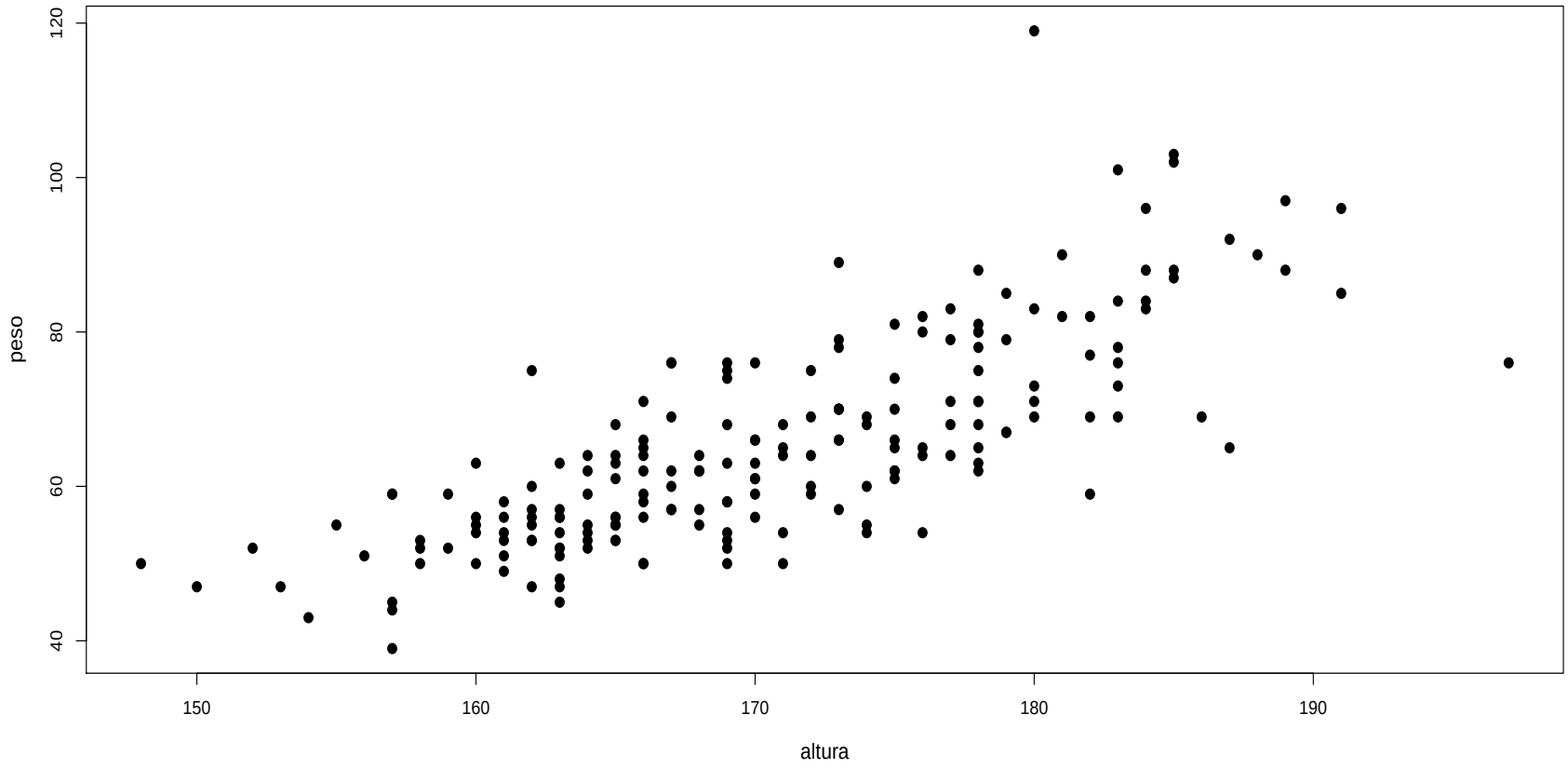
"rep" significa "reportado".

Exemplo 0: gráfico de dispersão



Exemplo 0: altura e peso de homens e mulheres

($\widetilde{\rho} = 0,7707$) sem a observação discrepante



Exemplo 0: altura e peso de homens e mulheres

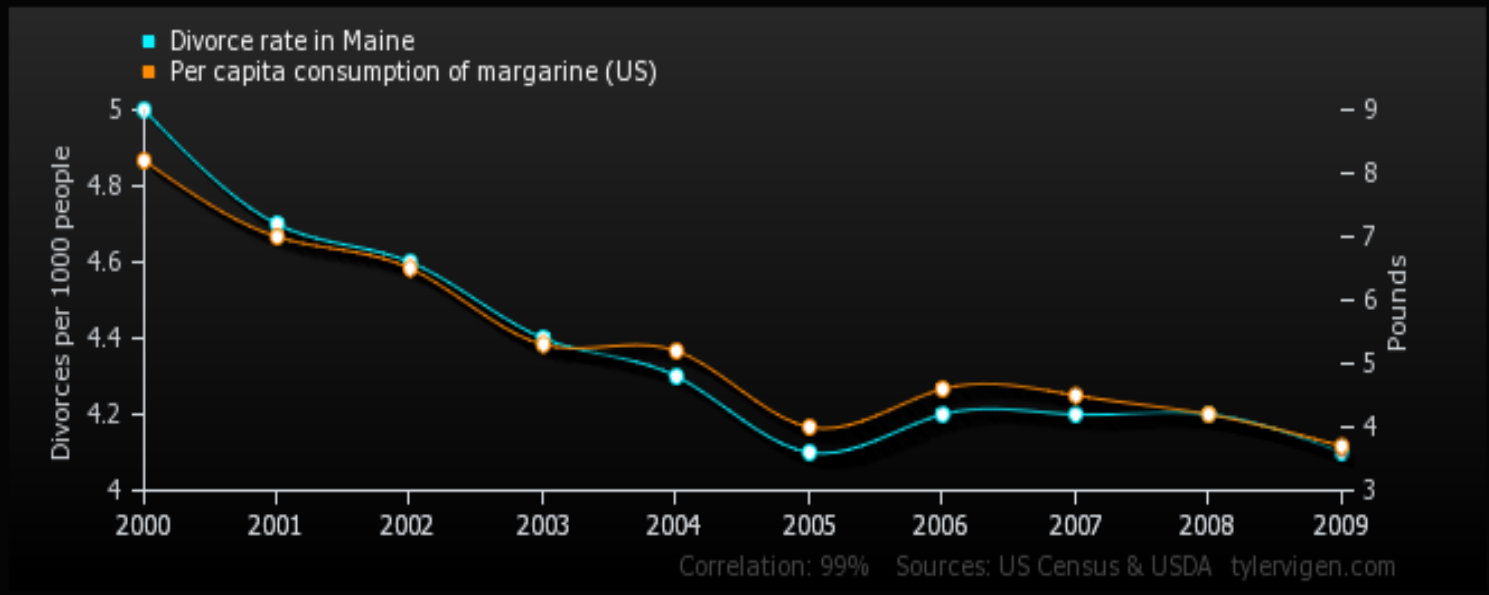
- Como esperado, a correlação estimada é elevada e positiva, mas isso não implica numa relação de causa e efeito, necessariamente.
- Ou seja, não é o fato de uma pessoa ser alta que a faz ter uma peso elevado e vice-versa.
- Devem existir outros fatores (genética, qualidade de vida, alimentação, fatores ambientais) que influenciem ambas as variáveis e que as façam estar positivamente relacionadas.
- Além disso, correlações altas (positivas ou negativas) podem ser espúrias (sem sentido).

Correlação

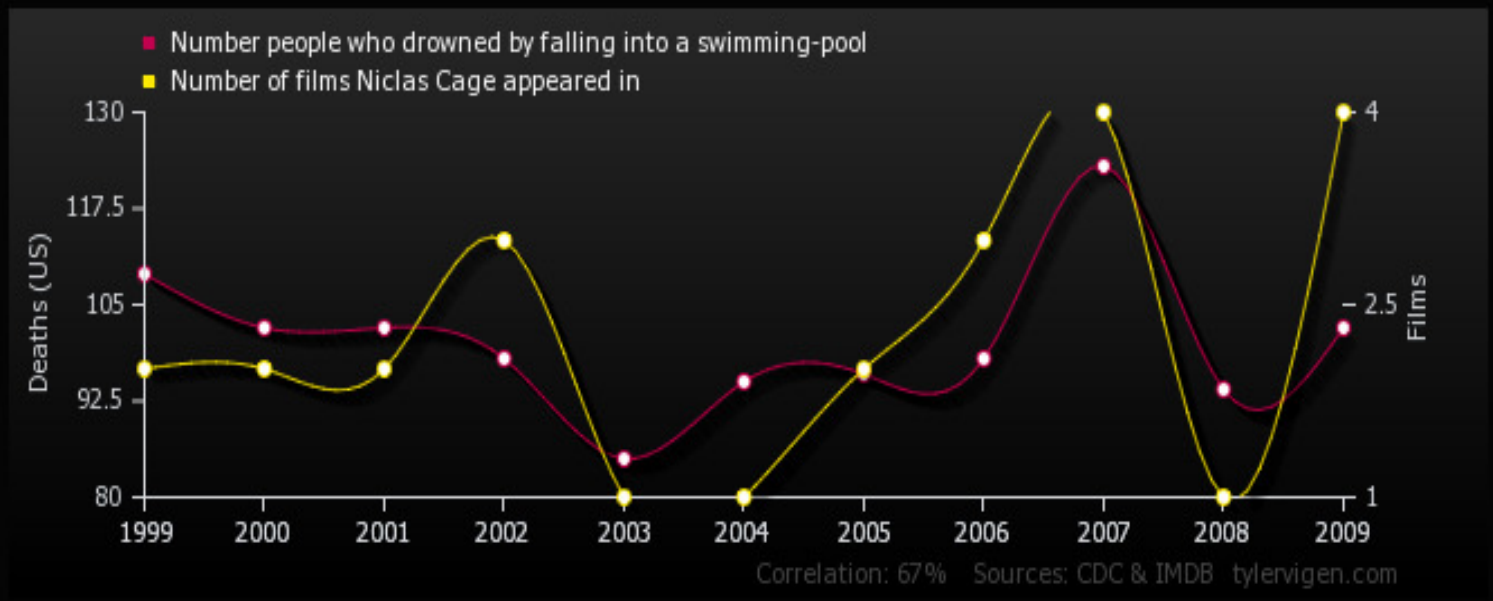
- Os dois gráficos a seguir foram extraídos do site

  `http://www.fastcodesign.com/3030529/infographic-of-the-day/hilarious-graphs-prove-that-correlation-isnt-causation`

Número de divórcios em Maine $\times$ Consumo per capita de margarina (EUA)



Número de pessoas que se afogaram em piscinas × número de filmes em que o Nicolas Cage apareceu



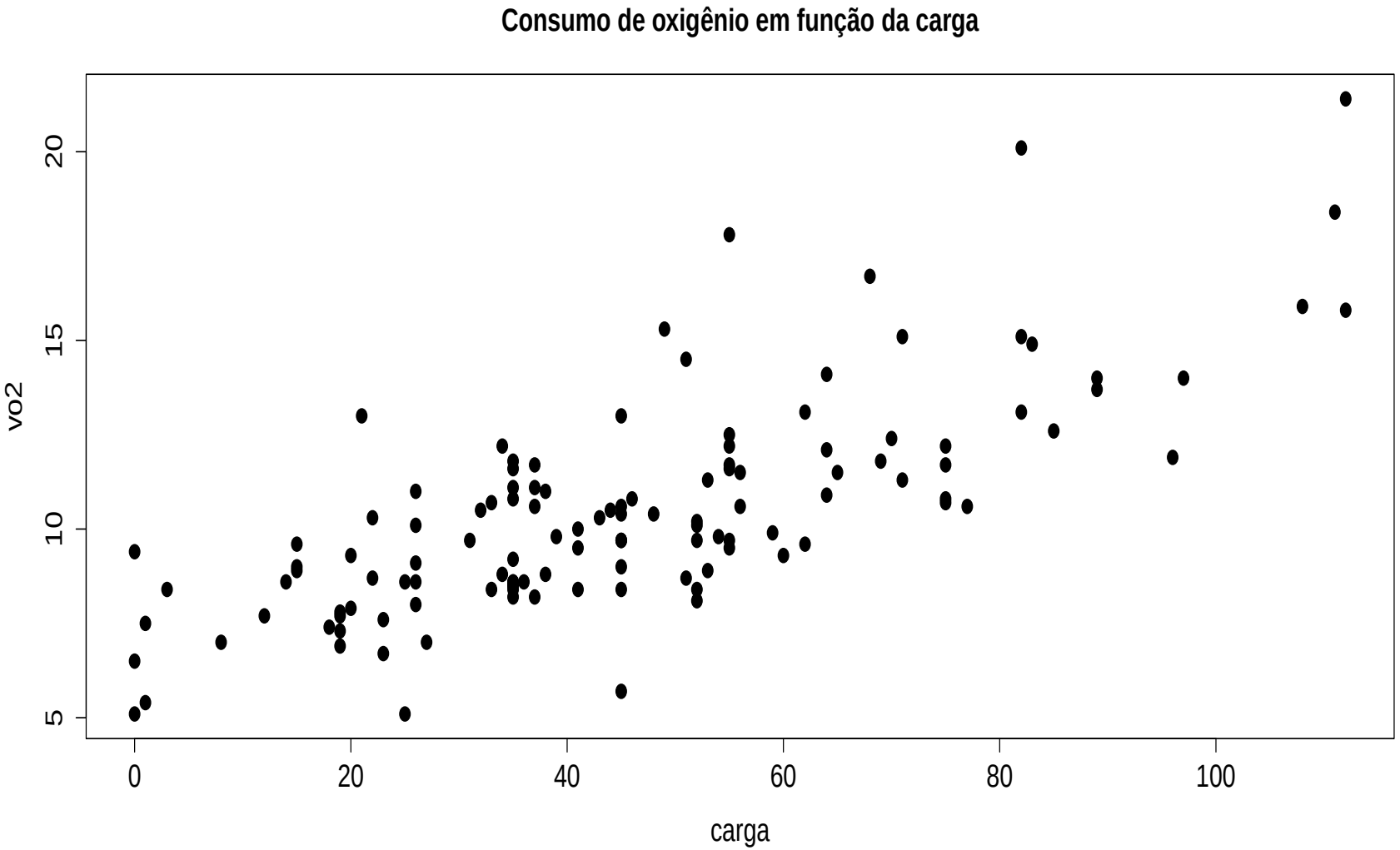
Exemplo 1: Teste de esforço cardiopulmonar

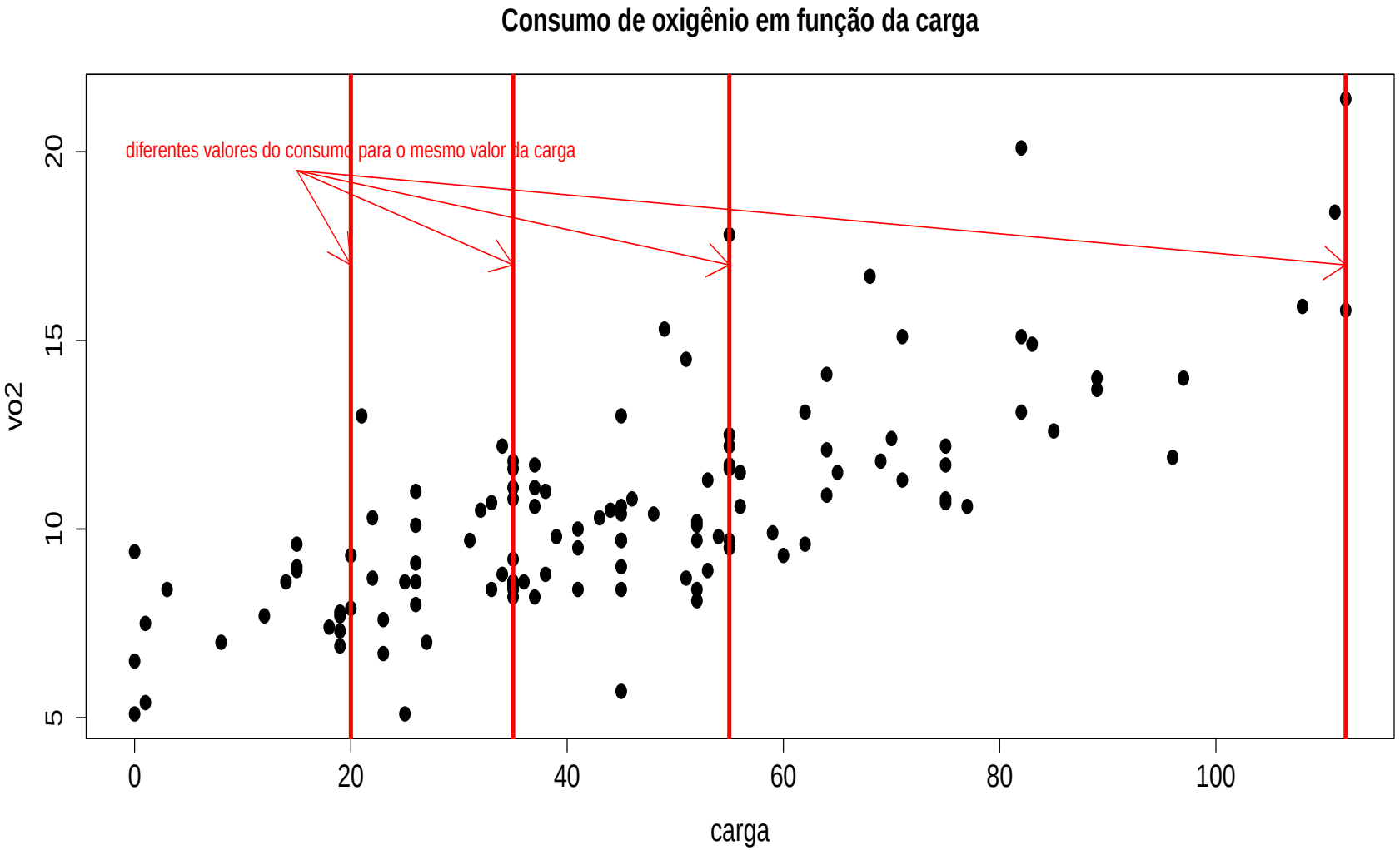
- Considere o estudo sobre teste de esforço cardiopulmonar em pacientes com insufiência cardíaca realizado no InCor da Faculdade de Medicina da USP pela Dra. Ana Fonseca Braga.
- Um dos objetivos do estudo é comparar os grupos formados pelas diferentes etiologias cardíacas quanto às respostas respiratórias e metabólicas obtidas do teste de esforço cardiopulmonar.
- Outro objetivo do estudo é saber se alguma das características observadas (ou combinação delas) pode ser utilizada como fator prognóstico de óbito.
- Os dados podem ser encontrados em http://www.ime.usp.br/∼jmsinger/doku.php?id=start.

- Etiologias : CH: chagásicos, ID: idiopáticos, IS: isquêmicos, C: controle.
- Considere que o objetivo é explicar a variação do consumo de oxigênio no limiar anaeróbio ($ml/(kg.min)$) em função da carga utilizada na esteira ergométrica para pacientes com diferentes etiologias (causas) de insuficiência cardíaca.
- A grosso modo o Limiar Anaeróbio é um ponto (limite), de divisão entre metabolismo essencialmente aeróbio e metabolismo essencialmente anaeróbio.
- Aeróbio (com a utilização de oxigênio) ; anaeróbio (sem a utilização de oxigênio).
- Como responder à pergunta de interesse?

Dados (constantes no site sob o nome Braga1998.txt)

ID	Etiologia	Carga	VO2
1	CH	41	10,0
2	CH	56	11,5
3	ID	8	7,0
4	ID	53	8,9
⋮	⋮	⋮	
7	ID	0	6,5
⋮	⋮	⋮	
123	C	64	14,1
124	C	70	12,4


Consumo de oxigênio em função da carga
vo2
5
10
15
20
0
20
40
60
80
100
carga


Consumo de oxigênio em função da carga
diferentes valores do consumo para o mesmo valor da carga
vo2
carga
5
10
15
20
0
20
40
60
80
100

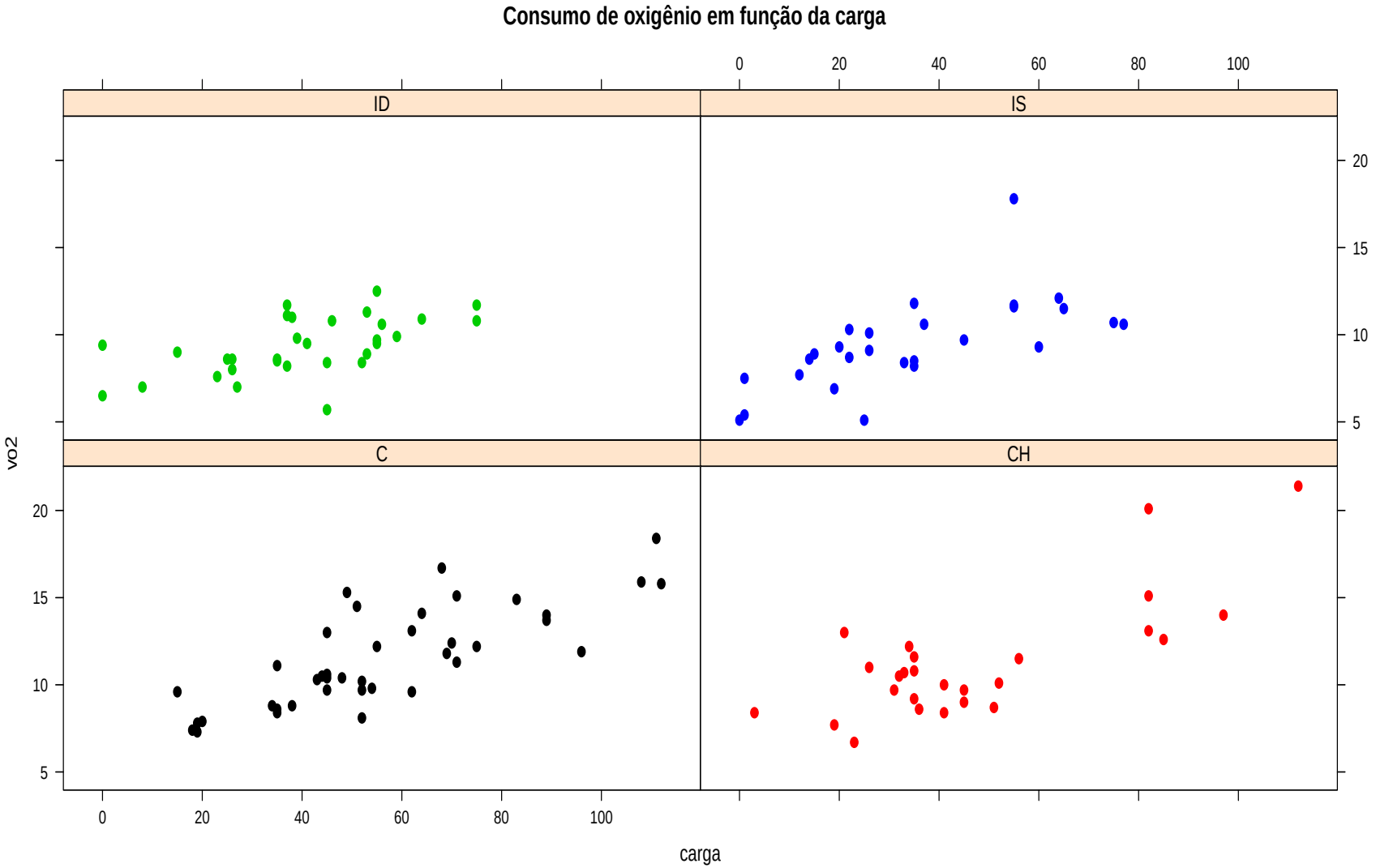
- Existe uma relação entre as duas variáveis?

- Existe uma relação entre as duas variáveis?
- De que tipo?

- Existe uma relação entre as duas variáveis?
- De que tipo?
- O fato de que quanto maior o valor da carga maior, maior o valor do consumo de oxigênio, implica numa relação de causa e efeito?

- Existe uma relação entre as duas variáveis?
- De que tipo?
- O fato de que quanto maior o valor da carga maior, maior o valor do consumo de oxigênio, implica numa relação de causa e efeito?
- Há outros fatores biológicos (hereditariedade, outras doenças), comportamentais (dieta, prática de exercícios, remédios) e ambientais (poluição, clima), que, verdadeiramente, ditariam os valores dessas duas variáveis para cada indivíduo?

- Existe uma relação entre as duas variáveis?
- De que tipo?
- O fato de que quanto maior o valor da carga maior, maior o valor do consumo de oxigênio, implica numa relação de causa e efeito?
- Há outros fatores biológicos (hereditariedade, outras doenças), comportamentais (dieta, prática de exercícios, remédios) e ambientais (poluição, clima), que, verdadeiramente, ditariam os valores dessas duas variáveis para cada indivíduo?
- O que significa dizer: para um dado valor da carga, o comportamento do consumo de oxigênio é aleatório e que pode ser modelado "apropriadamente" por uma estrutura probabilística (paramétrica)?


Consumo de oxigênio em função da carga
ID
IS
C
CH
vo2
carga
0
20
40
60
80
100
5
10
15
20

- É importante levar em consideração as diferentes etiologias?
- Se sim, como considerá-las na análise?
- Há interesse em comparar a influência da carga no consumo de oxigênio entre as diferentes etiologias cardíacas ?

Exemplo 2: Estudo da eficácia de escovas de dentes

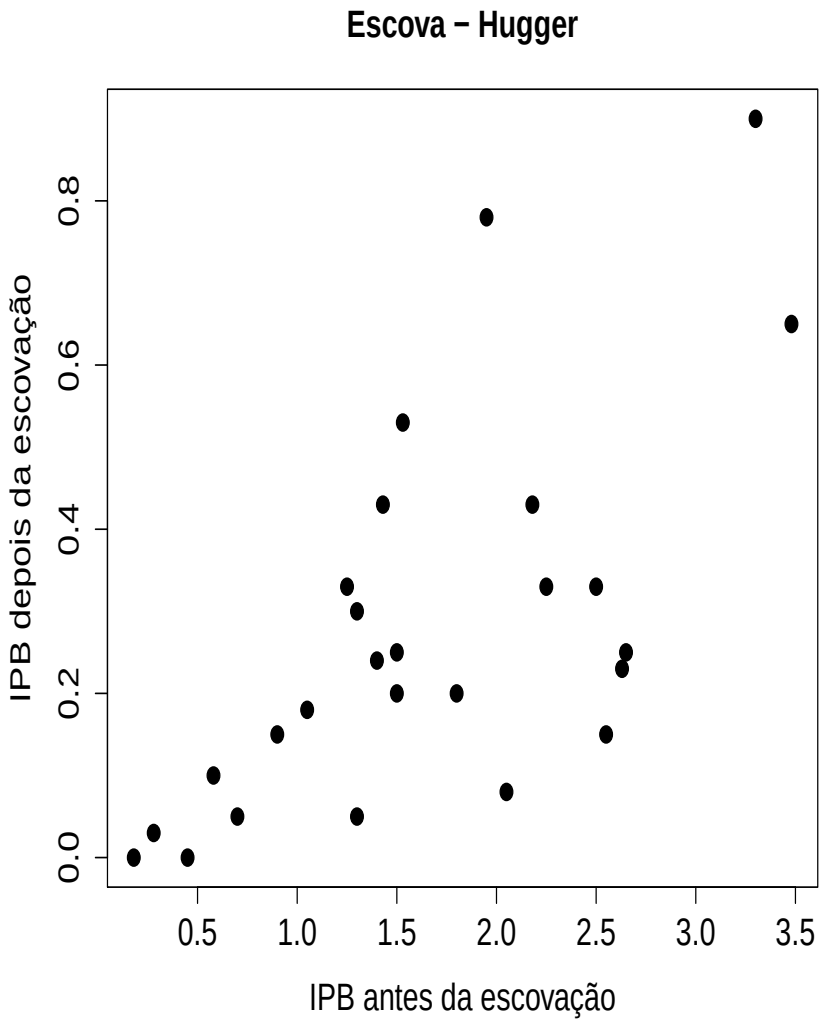
- Considere o seguinte estudo na área de Odontopediatria.
- O objetivo é comparar duas escovas de dente (convencional e experimental, chamada de "hugger") com respeito à redução de um índice de placa bacteriana (IPB) em crianças de ambos os sexos em idade pré-escolar.
- Os valores obtidos correspondem aos IPB's medidos em alguns dentes antes e depois da escovação dental de 14 crianças do sexo feminino e 12 do sexo masculino. Cada criança utilizou cada um dos tipos de escova sendo sempre a experimental, a primeira. O tipo de escova tende a ser melhor quanto maior for sua "capacidade de remoção" da placa bacteriana.

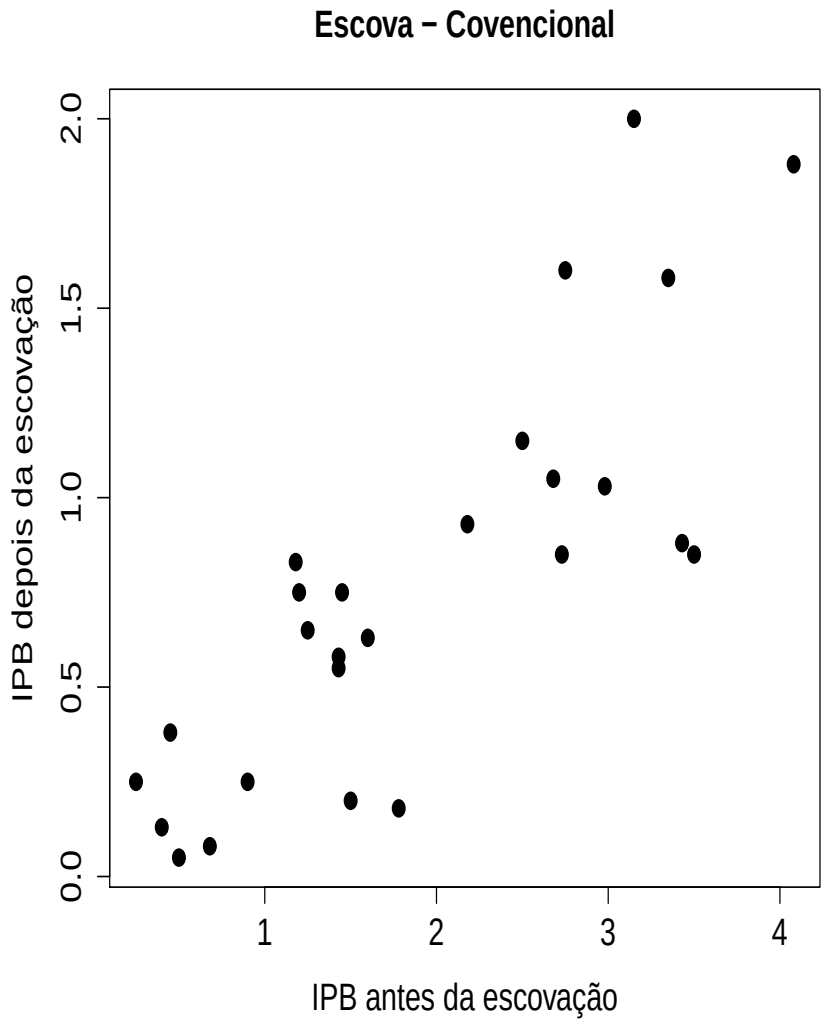
Dados

	Tipo de escova			
	Hugger		Convencional	
Criança	**Sexo**	**Antes**	**Depois**	**Antes**	**Depois**
1	F	2,18	0,43	1,2	0,75
2	F	2,05	0,08	1,43	0,55
⋮	⋮	⋮	⋮	⋮	⋮
25	M	1,3	0,05	2,73	0,85
26	M	2,65	0,25	3,43	0,88

Exemplo 2: Estudo da eficácia de escovas de dentes

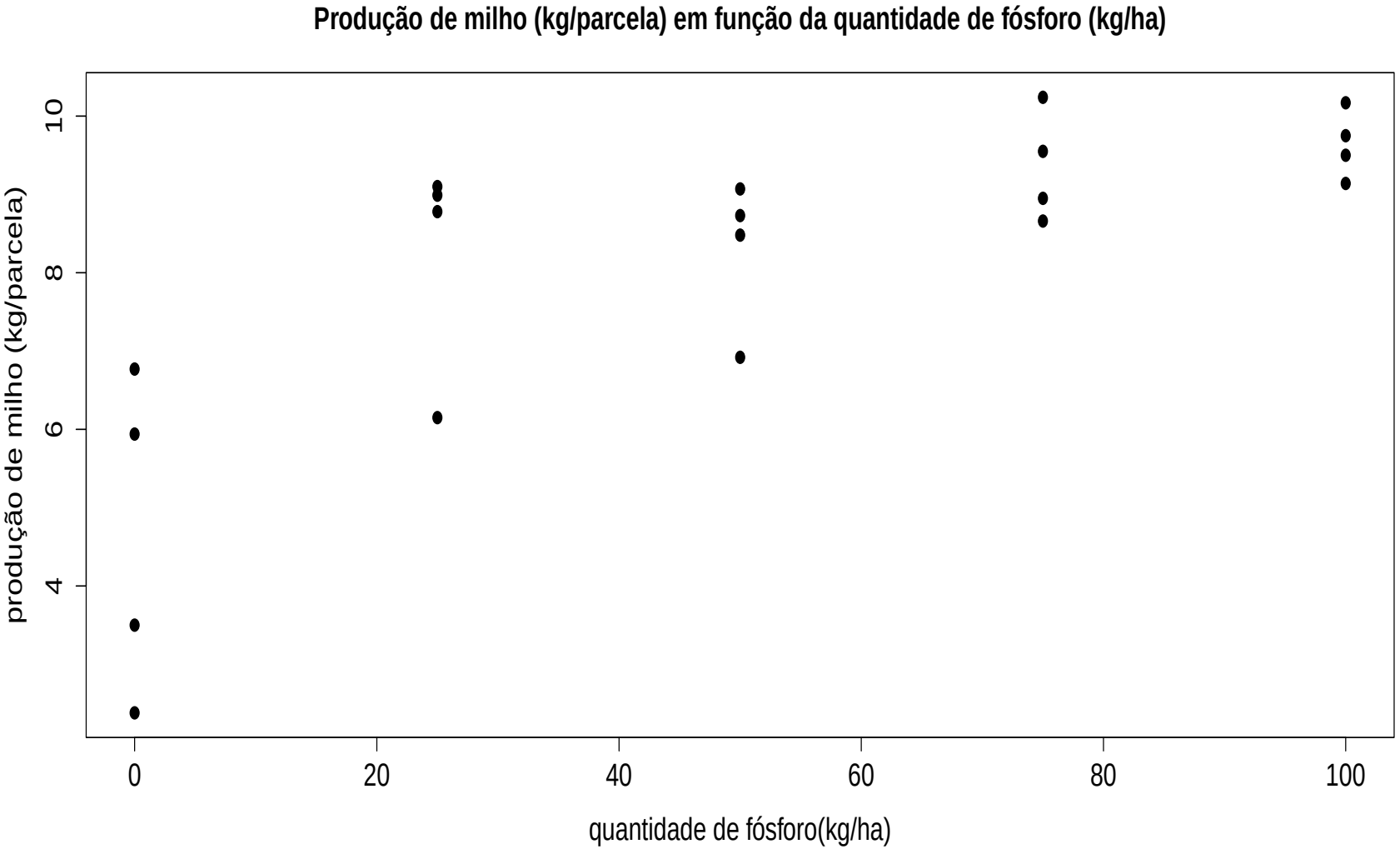
- Como utilizar os IPB's antes e depois ?
- Deve-se considerar a variável sexo?
- O fato de sempre se utilizar o tipo de escova experimental primeiramente pode ter influenciado os resultados?
- Medidas repetidas: cada criança é avaliada duas vezes. Possível existência de dependência entre as observações.


Escova – Hugger
IPB depois da escovação
IPB antes da escovação
0.0
0.2
0.4
0.6
0.8
0.5
1.0
1.5
2.0
2.5
3.0
3.5


Escova – Covencional
IPB depois da escovação
IPB antes da escovação
0.0
0.5
1.0
1.5
2.0
1
2
3
4

Exemplo 3: efeito do fósforo na produção de milho

- Tem-se o interesse em se saber se a quantidade (kg/ha) de fósforo existente (administrada) no solo afeta a produção de milho (de uma certa variedade) kg/parcela.
- Fator: quantidade de fósforo, k $= 5$ níveis (0,25,50,75,100), $n_i = 4, i = 1, 2, 3, 4$ repetições por tratamento (quantidade de fósforo administrada).
- Procedimento: 20 porções de terras, chamadas de parcelas (em condições semelhantes) foram consideradas e cada uma delas recebeu uma determinada quantidade de fósforo, de modo aleatório (completamente casualizado).


Produção de milho (kg/parcela) em função da quantidade de fósforo (kg/ha)
produção de milho (kg/parcela)
4
6
8
10
0
20
40
60
80
100
quantidade de fósforo(kg/ha)

Exemplo 3: efeito do fósforo na produção de milho

- Aparentemente, há uma "tendência crescente" na produção de milho em função da quantidade de fósforo (até certo valor).
- Contudo, provavelmente, depois de uma certa quantidade de fósforo, a produção tenderá a diminuir.
- Isso deve ser levado em consideração.

Modelagem

- Para todos os exemplos, podemos considerar algum tipo de modelagem estatística para responder às perguntas de interesse.
- Em nosso curso, consideraremos modelos lineares, em geral, normais e homocedásticos (variabilidade constante).
- A escolha de um modelo deve ser pautada: nos objetivos do experimento, nas características dos dados, em experiências anteriores (informações a priori) e na análise descritiva.

Cont.

- Tais modelos (de regressão, de planejamento ou de análise de covariância) podem ser decompostos em uma parte sistemática e uma parte aleatória.
- Todos eles podem ser acomodados em uma estrutura geral que estudaremos ao longo do semestre.
- Vamos discutir uma possibilidade para cada situação.

Exemplo 1: desconsiderando as etiologias cardíacas

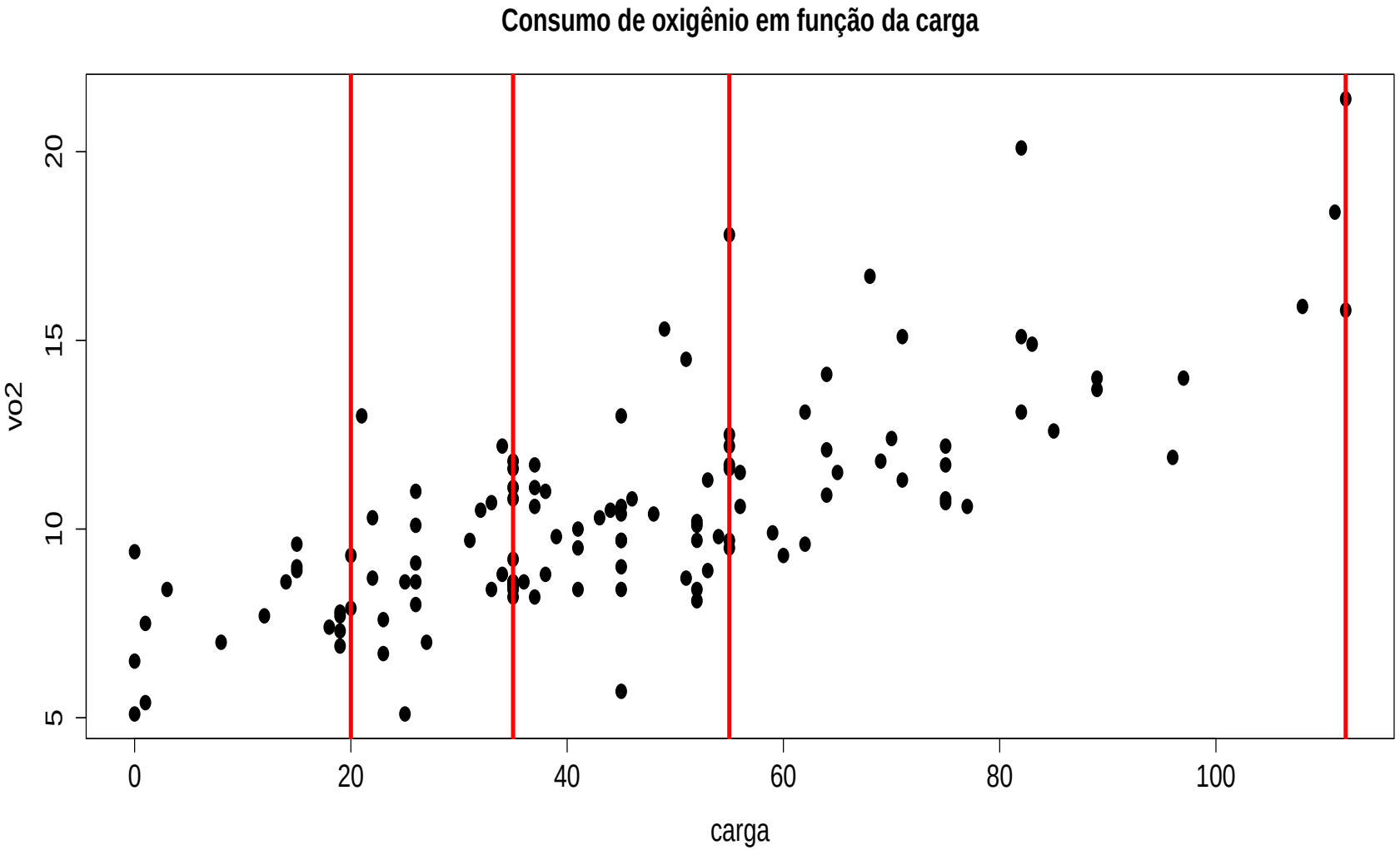
$$Y_i = \beta_0 + \beta_1 x_i + \xi_i, i = 1, ..., 124$$

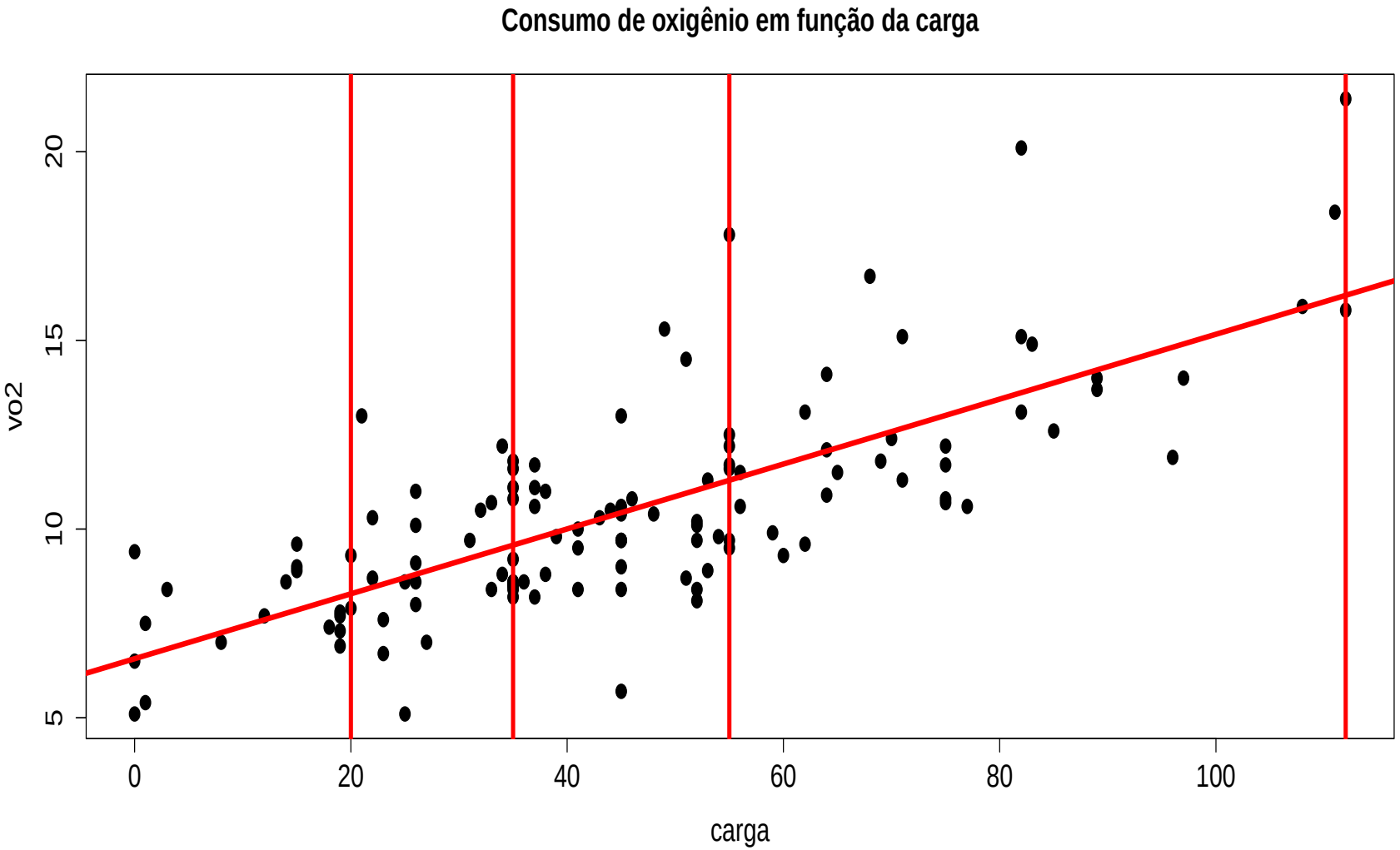
- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $(\beta_0, \beta_1, \sigma^2)'$: parâmetros desconhecidos.
- x_i: carga à que o paciente i foi submetido (conhecida e não aleatória).
- Parte sistemática: $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i$.
- Parte aleatória: ξ_i.
- O modelo acima implica que $Y_i \overset{ind.}{\sim} N(\beta_0 + \beta_1 x_i, \sigma^2)$, Y_i : valor do consumo de oxigênio do paciente i.

- β_1 : é o incremento (positivo ou negativo) esperado no consumo de oxigênio para o aumento de uma unidade na carga imposta.
- Se for possível observar $x_i = 0$, carga igual à 0, temos que:
  - β_0 : valor esperado do consumo de oxigênio para pacientes submetidos à uma carga igual à 0.
- Caso contrário, podemos considerar o seguinte modelo:

$$Y_i = \beta_0 + \beta_1(x_i - \overline{x}) + \xi_i, i = 1, ..., 124, \overline{x} = \frac{1}{124}\sum_{i=1}^{n} x_i.$$

- Neste caso, β_0 é o valor esperado do consumo de oxigênio para pacientes submetidos à uma carga igual à média amostral.


Consumo de oxigênio em função da carga
vo2
5
10
15
20
0
20
40
60
80
100
carga


Consumo de oxigênio em função da carga
vo2
carga
5
10
15
20
0
20
40
60
80
100

Exemplo 2: desconsiderando o sexo

$Y_{ij} = \beta_{0i} + \beta_{1i}(x_{ij} - \overline{x}) + \xi_{ij}, i = 1, 2$(tipo de escova, 1 - Hugger; 2 - Convencional)

$$\overline{x} = \frac{1}{52} \sum_{i=1}^{2} \sum_{j=1}^{26} x_{ij} = 1,76$$

- $\xi_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $(\beta_{01}, \beta_{02}, \beta_{11}, \beta_{12}, \sigma^2)'$: parâmetros desconhecidos.
- x_{ij}: IPB pré-escovação da criança j utilizando a escova do tipo i.
- Parte sistemática: $\mathcal{E}(Y_{ij}) = \beta_{0i} + \beta_{1i}(x_{ij} - \overline{x})$.
- Parte aleatória: ξ_{ij}.
- O modelo acima implica que $Y_{ij} \overset{ind.}{\sim} N(\beta_{0i} + \beta_{1i}(x_{ij} - \overline{x}), \sigma^2)$,

  Y_{ij} :(IPB pós - escovação) da criança j utilizando a escova do tipo i.

- β_{1i} : é o incremento (positivo ou negativo) esperado no IPB pós-escovação para o aumento em uma unidade no IPB pré-escovação quando se utiliza a escova i.
- β_{0i} é o valor esperado no IPB pós-escovação para crianças com IPB pré-escovação igual à $\overline{x}$ quando se utiliza a escova i.

Exemplo 3

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \xi_i, i = 1, 2, ..., 20$$

- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $(\beta_0, \beta_1, \beta_2, \sigma^2)'$: parâmetros desconhecidos.
- x_i : quantidade de fósforo ministrada a i-ésima parcela.
- Parte sistemática: $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i + \beta_2 x_i^2$.
- Parte aleatória: ξ_i.
- O modelo acima implica que $Y_i \overset{ind.}{\sim} N(\beta_0 + \beta_1 x_i + \beta_2 x_i^2, \sigma^2)$, Y_i : é produção de milho da i-ésima parcela.

- β_0 : valor esperado (média) da produção de milho quando a quantidade de fósforo aplicada é igual à 0.
- A interpretação isolada dos parâmetros β_1 e β_2 é complicada mas, podemos dizer que $\frac{-\beta_1}{2\beta_2}$ é a quantidade de fósforo que retornar a produção máxima (ou mínima (?)) esperada de milho.
- Neste caso, o valor esperado da produção máxima (ou mínima) de milho é dado por : $\mu_{max} = \frac{4\beta_0\beta_2 - \beta_1^2}{4\beta_2}$

Modelo de regressão normal linear simples homocedástico (MRNLSH)

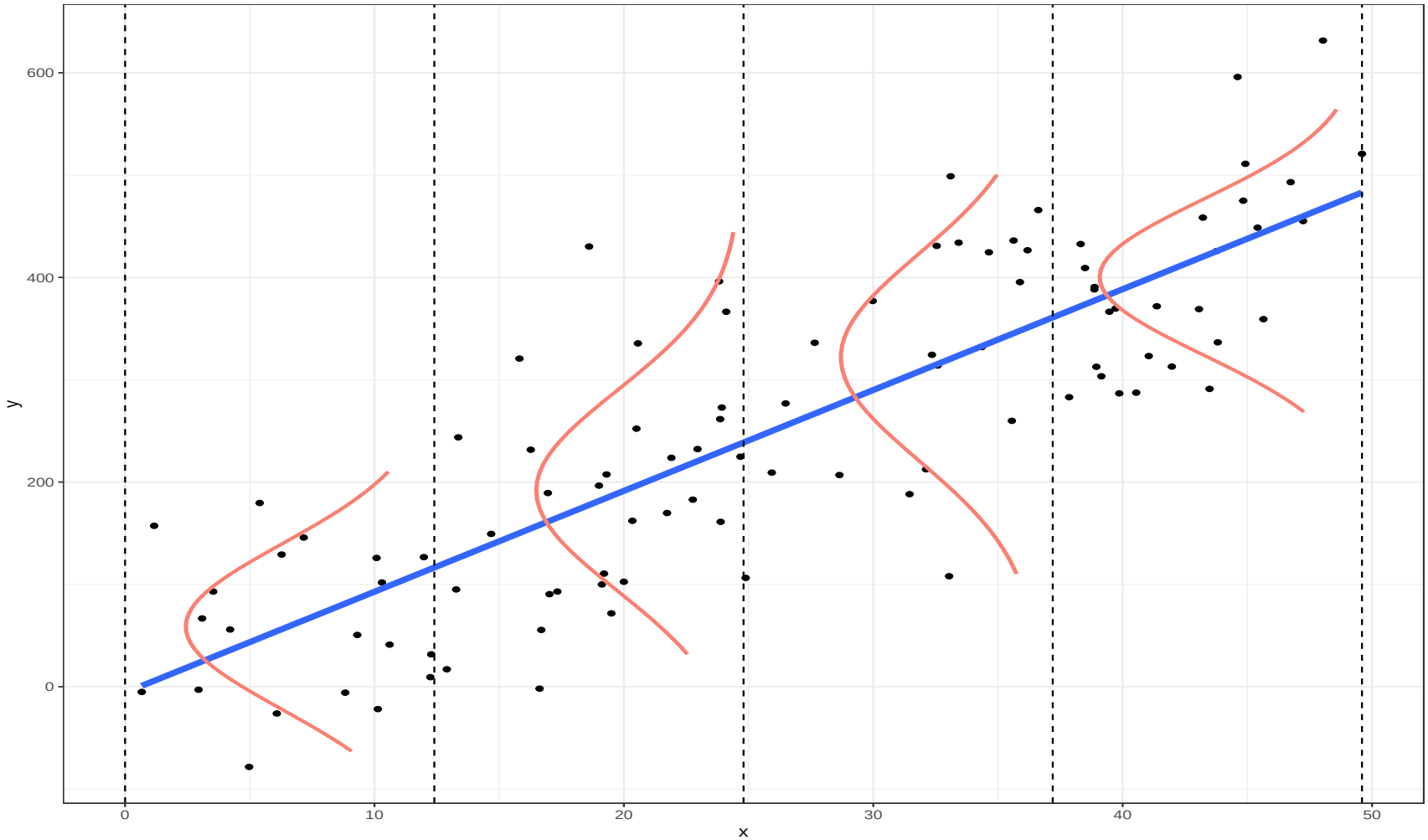
$$Y_i = \beta_0 + \beta_1 x_i + \xi_i, i = 1, ..., n$$

- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- Estimação de $\boldsymbol{\beta} = (\beta_0, \beta_1)'$ (por máxima verossimilhança: maximizar a verossimilhança).
- Mínimos quadrados ordinários (MQO). Minimizar

$$Q = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2,$$

em relação à $\boldsymbol{\beta}$.

Ilustração gráfica do MRNLSH



Estimação

- Resolver o sistema de equações a seguir (defina:

$S(\beta_i) = \frac{\partial Q}{\partial \beta_i}, i = 0, 1)$

$$\begin{cases} S(\widetilde{\beta}_0) = 0 \\ S(\widetilde{\beta}_1) = 0 \end{cases}$$

- Temos que $S(\beta_0) = -2\sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i)$ e

$S(\beta_1) = -2\sum_{i=1}^{n} x_i(y_i - \beta_0 - \beta_1 x_i)$

- Portanto

$$\begin{cases} \sum_{i=1}^{n}(y_i - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_i) = 0 \\ \sum_{i=1}^{n} x_i(y_i - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_i) = 0 \end{cases} \rightarrow$$

$$\begin{cases} \sum_{i=1}^{n} y_i - \widetilde{\beta}_0 n - \widetilde{\beta}_1 \sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} x_i y_i - \widetilde{\beta}_0 \sum_{i=1}^{n} x_i - \widetilde{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0 \end{cases}$$

- Logo

$$\begin{cases} \overline{y} - \widetilde{\beta}_0 - \widetilde{\beta}_1 \overline{x} = 0 \ (1) \\ \overline{yx} - \widetilde{\beta}_0 \overline{x} - \widetilde{\beta}_1 \overline{x}_2 = 0 \ (2) \end{cases} \tag{1}$$

em que $\overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$, $\overline{x}_j = \sum_{i=1}^{n} \frac{x_i^j}{n}, j = 1, 2$ e $\overline{yx} = \sum_{i=1}^{n} \frac{y_i x_i}{n}$.

Da Equação (1) do sistema (1) temos que

$$\widetilde{\beta}_0 = \overline{y} - \widetilde{\beta}_1 \overline{x}. \tag{2}$$

De (2) na Equação (2) do sistema (1), temos que

$$\overline{yx} - \overline{y}\,\overline{x} + \widetilde{\beta}_1 \overline{x}^2 - \widetilde{\beta}_1 \overline{x}_2 = 0 \rightarrow \widetilde{\beta}_1 = \frac{\overline{yx} - \overline{y}\,\overline{x}}{\overline{x}_2 - \overline{x}^2}$$

$$= \frac{1}{n(\overline{x}_2 - \overline{x}^2)} \left[\sum_{i=1}^{n} y_i (x_i - \overline{x}) \right] = \frac{1}{\sum_{i=1}^{n} (x_i - \overline{x})^2} \left[\sum_{i=1}^{n} y_i (x_i - \overline{x}) \right]$$

- Prove que $\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \overline{x}_2 - \overline{x}^2$
- Assim os estimadores de MQ de β_0 e β_1 são, respectivamente,

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1\overline{x} = \sum_{i=1}^{n}\left[Y_i\left(\frac{1}{n} - \overline{x}\frac{(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\right] \quad (3)$$

$$\widehat{\beta}_1 = \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\left[\sum_{i=1}^{n} Y_i(x_i - \overline{x})\right] \quad (4)$$

e as respectivas estimativas são dadas por

$$\widetilde{\beta}_0 = \overline{y} - \widetilde{\beta}_1\overline{x} = \sum_{i=1}^{n}\left[y_i\left(\frac{1}{n} - \overline{x}\frac{(x_i - \overline{x})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\right)\right]$$

$$\widetilde{\beta}_1 = \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\left[\sum_{i=1}^{n} y_i(x_i - \overline{x})\right]$$

- Defini-se o valore predito para a i-ésima observação (que coincide com o valor predito para sua esperança) como $\widehat{Y}_i = \widehat{\mathcal{E}(Y)}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_i$. Mais a frente, veremos como construir intervalos de confiança (esperança) e de previsão (valor individual).
- Note que o método de MQO não requer suposições para a distribuição dos erros. Exercício: prove que, sob as suposições consideradas (independência, normalidade e homocedasticidade dos erros) os estimatimadores de MQ de $\boldsymbol{\beta}$ coincidem com os de MV (máxima verossimilhança). Obtenha também o EMV de σ^2.
- Vamos obter as distribuições dos estimadores de MQ. Como ambos são combinações lineares de normais (veja (3) e (4)), então segue-se que $\widehat{\beta}_0 \sim N(\mathcal{E}(\widehat{\beta}_0), \mathcal{V}(\widehat{\beta}_0))$ e $\widehat{\beta}_1 \sim N(\mathcal{E}(\widehat{\beta}_1), \mathcal{V}(\widehat{\beta}_1))$.

- Temos que

$$\mathcal{E}(\widehat{\beta}_0) = \mathcal{E}(\overline{Y}) - \mathcal{E}(\widehat{\beta}_1)\overline{x} = \beta_0 + \beta_1\overline{x} - \mathcal{E}(\widehat{\beta}_1)\overline{x} \qquad (5)$$

- Por outro lado

$$\begin{aligned}\mathcal{E}(\widehat{\beta}_1) &= \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\left[\sum_{i=1}^{n}\mathcal{E}(Y_i)(x_i - \overline{x})\right] \\ &= \frac{1}{\sum_{i=1}^{n}(x_i - \overline{x})^2}\left[\sum_{i=1}^{n}(\beta_0 + \beta_1 x_i)(x_i - \overline{x})\right]\end{aligned}$$

- Continuando

$$
\begin{aligned}
\mathcal{E}(\widehat{\beta}_1) &= \frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\left[\beta_0\sum_{i=1}^{n}(x_i-\overline{x})+\beta_1\sum_{i=1}^{n}(x_i^2-x_i\overline{x})\right] \\
&= \frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\left[\beta_0\underbrace{\sum_{i=1}^{n}(x_i-\overline{x})}_{0}+\beta_1\underbrace{\sum_{i=1}^{n}(x_i^2-x_i\overline{x})}_{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right] \\
&= \frac{1}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\left[\beta_1\sum_{i=1}^{n}(x_i-\overline{x})^2\right]=\beta_1 \qquad (6)
\end{aligned}
$$

- De (6) em (5), vem que:

$$\mathcal{E}(\widehat{\beta}_0) = \beta_0 + \beta_1\overline{x} - \beta_1\overline{x} = \beta_0$$

- Portanto, ambos os estimadores são não viciados.
- Por outro lado

$$\begin{aligned}\mathcal{V}(\widehat{\beta}_1) &\overset{ind.}{=} \frac{1}{(\sum_{i=1}^{n}(x_i-\overline{x})^2)^2}\left[\sum_{i=1}^{n}\mathcal{V}(Y_i)(x_i-\overline{x})^2\right] \\ &= \sigma^2\frac{1}{(\sum_{i=1}^{n}(x_i-\overline{x})^2)^2}\left[\sum_{i=1}^{n}(x_i-\overline{x})^2\right] = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\end{aligned}$$

- Além disso,

$$\begin{aligned}\mathcal{V}(\widehat{\beta}_0) &= \mathcal{V}(\overline{Y}) + \overline{x}^2\mathcal{V}(\widehat{\beta}_1) - 2\text{Cov}(\overline{Y}, \widehat{\beta}_1\overline{x}) \\ &= \frac{\sigma^2}{n} + \overline{x}^2\frac{\sigma^2}{\sum_{i=1}^n (x_i - \overline{x})^2} - 2\overline{x}\text{Cov}(\overline{Y}, \widehat{\beta}_1)\end{aligned}$$

Mas, note que $\widehat{\beta}_1 = \sum_{i=1}^n Y_i a_i$, em que

$$a_i = \frac{(x_i - \overline{x})}{\sum_{i=1}^n (x_i - \overline{x})^2} \tag{7}$$

com $\sum_{i=1}^n a_i = 0$

$$\begin{aligned}\text{Cov}(\overline{Y}, \overline{x}\widehat{\beta}_1) &= \overline{x}\text{Cov}\left(\frac{1}{n}\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i a_i\right) = \frac{\overline{x}}{n}\text{Cov}\left(\sum_{i=1}^n Y_i, \sum_{i=1}^n Y_i a_i\right) \\ &= \frac{\overline{x}}{n}\sum_{i=1}^n a_i \underbrace{\text{Cov}(Y_i, Y_i)}_{\mathcal{V}(Y_i)=\sigma^2} = \frac{\overline{x}\sigma^2}{n}\sum_{i=1}^n a_i = 0\end{aligned}$$

- Logo

$$\mathcal{V}(\widehat{\beta}_0) \quad = \quad \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \right]$$

- Portanto $\widehat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right]\right)$ e $\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\right)$. Como $\mathcal{V}(\widehat{\beta}_0) \to 0$ e $\mathcal{V}(\widehat{\beta}_1) \to 0$, ambos os estimadores são consistentes.
- A distribuição conjunta de $\widehat{\boldsymbol{\beta}}_0 = (\widehat{\beta}_0, \widehat{\beta}_1)$ (provaremos tal resultado mais a frente) é dada por:

$$N_2 \left[\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix}, \begin{pmatrix} \mathcal{V}(\widehat{\beta}_0) & \text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) \\ . & \mathcal{V}(\widehat{\beta}_1) \end{pmatrix} \right]$$

(normal bivariada)

- Em que

$$\begin{aligned} \text{Cov}(\widehat{\beta}_0, \widehat{\beta}_1) &= \text{Cov}(\overline{Y} - \widehat{\beta}_1 \overline{x}, \widehat{\beta}_1) = \underbrace{\text{Cov}(\overline{Y}, \widehat{\beta}_1)}_{0} - \overline{x}\mathcal{V}(\widehat{\beta}_1) \\ &= -\overline{x}\frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2} \end{aligned}$$

- Inferência: para construirmos intervalos de confiança (IC) bem como testes de hipótese (TH) precisaremos (é uma forma) de obter a distribuição exata ou assintótica de quantidades pivotais bem como de estatística de teste apropriadas.
- Contudo, notem que σ^2 é desconhecido. Devemos, portanto, utilizar um estimador apropriado para ele.

- Sugestão:

$$\widehat{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)^2$$

em que $\widetilde{\sigma}^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \widetilde{\beta}_0 - \widetilde{\beta}_1 x_i)^2$ é a respectiva estimativa.

- Este estimador é não viciado, consistente, $\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-2)}$. Além disso $\widehat{\beta}_0 \perp \widehat{\sigma}^2$, $\widehat{\beta}_1 \perp \widehat{\sigma}^2$ (provaremos mais adiante).
- Vamos provar que ele é não viciado. Temos que:

$$\begin{aligned}
\mathcal{E}(\widehat{\sigma}^2) &= \frac{1}{n-2}\sum_{i=1}^{n}\mathcal{E}\left(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i\right)^2 \\
&= \frac{1}{n-2}\left[\sum_{i=1}^{n}\mathcal{V}(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i) + \underbrace{\mathcal{E}^2(Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_i)}_{0}\right]
\end{aligned}$$

- Continuando (veja "a_i" na Equação (7)):

$$
\begin{aligned}
\mathcal{E}(\widehat{\sigma}^2) &= \frac{1}{n-2}\sum_{i=1}^{n}\Bigg[\mathcal{V}(Y_i)+\mathcal{V}(\widehat{\beta}_0)+\mathcal{V}(\widehat{\beta}_1)x_i^2-\mathrm{Cov}(Y_i,\widehat{\beta}_0) \\
&- x_i\mathrm{Cov}(Y_i,\widehat{\beta}_1)+x_i\mathrm{Cov}(\widehat{\beta}_0,\widehat{\beta}_1)\Bigg] \\
&= \frac{1}{n-2}\sum_{i=1}^{n}\Bigg[\sigma^2+\frac{\sigma^2}{n}+\frac{\overline{x}\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}+\frac{x_i^2\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}-\frac{2\sigma^2}{n} \\
&+ 2\overline{x}a_i\sigma^2-2x_ia_i\sigma^2-2x_i\overline{x}\frac{\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\Bigg] \\
&= \frac{1}{n-2}\Bigg[n\sigma^2-\sigma^2+\frac{n\overline{x}\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}+\frac{\sigma^2\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\
&- 2\frac{\sigma^2\sum_{i=1}^{n}x_i(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}-2n\frac{\overline{x}^2\sigma^2}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\Bigg]=\sigma^2
\end{aligned}
$$

- Voltando à questão da inferência com respeito ao vetor $\boldsymbol{\beta}$. Temos que (definindo $d = \frac{1}{n} + \frac{\overline{x}}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$)

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\sigma^2 d}} \sim N(0,1)\,; \quad \frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\sigma^2 / (\sum_{i=1}^{n}(x_i - \overline{x})^2)}} \sim N(0,1)$$

- Além disso, já vimos que $\frac{(n-2)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-2}$ e que é independente de $\widehat{\beta}_0$ e $\widehat{\beta}_1$, logo

$$\frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\sigma^2 d}} \Big/ \sqrt{\frac{(n-2)\widehat{\sigma}^2}{(n-2)\sigma^2}} = \frac{\widehat{\beta}_0 - \beta_0}{\sqrt{\widehat{\sigma}^2 d}} \sim t_{(n-2)}$$

- Analogamente, temos que $\frac{\widehat{\beta}_1 - \beta_1}{\sqrt{\widehat{\sigma}^2 / (\sum_{i=1}^{n}(x_i - \overline{x})^2)}} \sim t_{(n-2)}$

- Intervalos de confiança: considerando-se

  $P(X \leq t_{\frac{1+\gamma}{2}}) = \frac{1+\gamma}{2}, (X \sim t_{(n-p)})$, temos que $j = 0, 1$

$$IC(\beta_j, \gamma) = \left[\widehat{\beta}_j - t_{\frac{1+\gamma}{2}}\sqrt{\widehat{\sigma}^2\psi_j}; \widehat{\beta}_j + t_{\frac{1+\gamma}{2}}\sqrt{\widehat{\sigma}^2\psi_j}\right]$$

  em que $\psi_0 = d, \psi_1 = \frac{1}{(\sum_{i=1}^{n}(x_i - \overline{x})^2)}$.

- IC numérico:

$$IC(\beta_j, \gamma) = \left[\widetilde{\beta}_j - t_{\frac{1+\gamma}{2}}\sqrt{\widetilde{\sigma}^2\psi_j}; \widetilde{\beta}_j + t_{\frac{1+\gamma}{2}}\sqrt{\widetilde{\sigma}^2\psi_j}\right]$$

  em que $\widetilde{(.)}$ são as respectivas estimativas.

Testes de hipóteses

- Suponha que queremos testar $H_0 : \beta_j = \beta_{j0}$ vs $H_1 : \beta_j \neq \beta_{j0}$, para algum j, em que β_{j0} é um valor fixado, j=0,1.
- Estatística do teste $T_t = \frac{\widehat{\beta}_j - \beta_{j0}}{\sqrt{\widehat{\sigma}^2 \psi_j}}$, em que $\widehat{\beta}_j$.

Testes de hipóteses

- Sob H_0, $T_t \sim t_{(n-p)}$. Assim, rejeita-se H_0 se $|t_t| \geq t_c$, em que $t_t = \frac{\widetilde{\beta}_j - \beta_{j0}}{\sqrt{\widetilde{\sigma}^2 \psi_j}}$ e $P(X \geq t_c | H_0) = \alpha/2, X \sim t_{(n-p)}$.
- De modo equivalente, rejeita-se H_0 se p-valor $\leq \alpha$, em que p-valor $= 2P(X \geq |t_t| | H_0)$, $X \sim t_{(n-p)}$

Ajuste de modelos de regressão linear simples normais homocedásticos no R

- Função *lm*.
- Comando geral $lm(y \sim x1)$, y: variável resposta, x_1: variável explicativa.
- Modelo sem intercepto $lm(y \sim -1 + x1)$, y: variável resposta, x_1: variável explicativa.

Exemplo 1: sem considerar as etiologias cardíacas

$$Y_i = \beta_0 + \beta_1 x_i + \xi_i$$

Parâmetro	Estimativa	EP	IC(95%)	Estat. t	p-valor
β_0	6,563	0,356	[5,859 ; 7,268]	18,434	$<$0,0001
β_1	0,085	0,006	[0,072 ; 0,100]	12,516	$<$0,0001

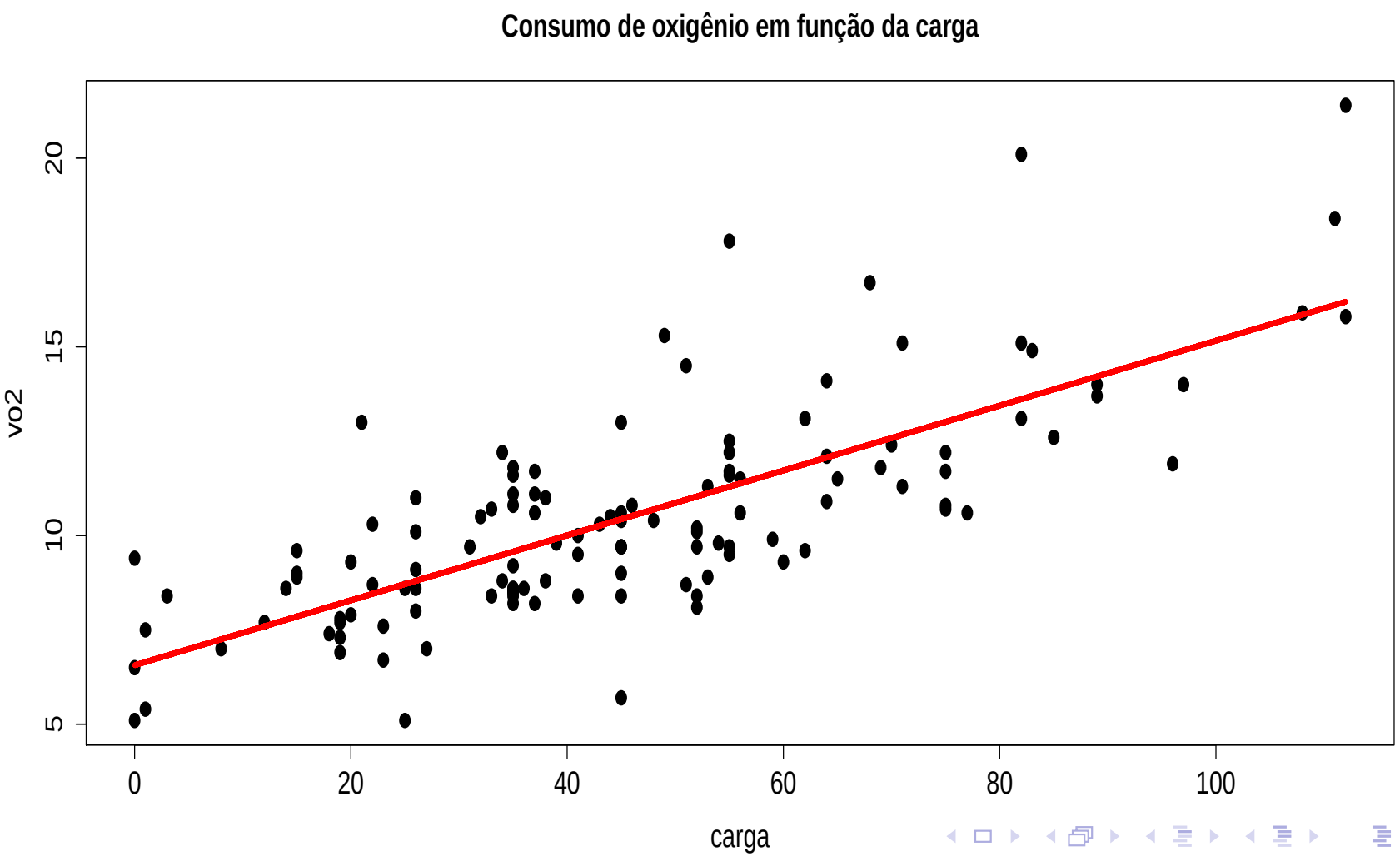
- Os dois parâmetros são diferentes de 0.
- A carga influencia positivamente o consumo de oxigênio.

Exemplo 1: sem considerar as etiologias cardíacas

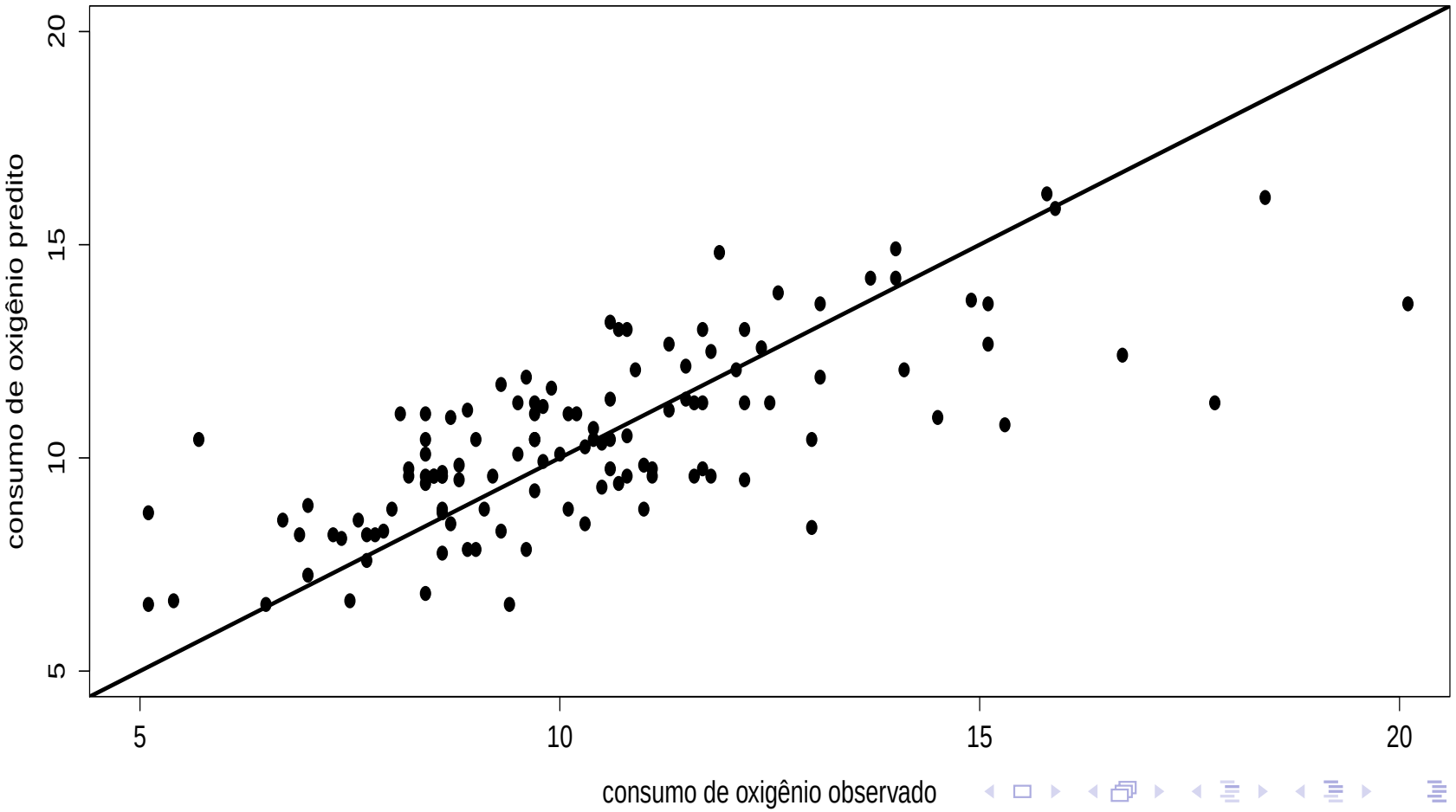
$Y_i = \beta_0 + \beta_1 x_i + \xi_i$ (cont.)

- O consumo de oxigênio para pacientes submetidos à carga 0 tende a se apresentar entre 5,859 e 7,268 ml/(kg.min).
- Por outro lado, o aumento esperado no consumo para o aumento em uma unidade da carga tende a se apresentar entre 0,072 e 0,100 ml/(kg.min).
- **A etapa de verificação de qualidade de ajuste do modelo, que deve preceder a sua utilização para fins inferenciais, será discutida posteriormente. Isso vale para todos os exemplos que veremos.**

Dispersão entre carga e consumo e reta ajustada



Consumos de oxigênio observado e predito modelo



Otimalidade dos estimadores

- Vamos provar que $\widehat{\beta}_0$ é o melhor (menor variância) estimador linear não viciado para β_0. A prova para $\widehat{\beta}_1$ é análoga e fica como exercício.
- Estamos restritos à classe dos estimadores $\widehat{\beta}_0 = \sum_{i=1}^{n} a_i Y_i$ com $a_i, i = 1, 2, ..., n$ não aleatórios, tais que

$$\begin{aligned} \mathcal{E}(\widehat{\beta}_0) &= \sum_{i=1}^{n} a_i \mathcal{E}(Y_i) = \sum_{i=1}^{n} a_i(\beta_0 + \beta_1 x_i) \\ &= \beta_0 \sum_{i=1}^{n} a_i + \beta_1 \sum_{i=1}^{n} a_i x_i = \beta_0 \end{aligned}$$

Otimalidade dos estimadores

- Isso implica que

$$\sum_{i=1}^{n} a_i = 1 \, ; \sum_{i=1}^{n} a_i x_i = 0 \tag{8}$$

- Além disso, $\mathcal{V}(\widehat{\beta}_0) = \sum_{i=1}^{n} a_i^2 \mathcal{V}(Y_i) = \sigma^2 \sum_{i=1}^{n} a_i^2$.
- Portanto, devemos minimizar $\sum_{i=1}^{n} a_i^2$ sujeito à (8). Utilizando a metodologia dos multiplicadores de Lagrange, isto equivale à minimizar (em a_i) a função:

$$L = \sum_{i=1}^{n} a_i^2 + \lambda_1 \left(\sum_{i=1}^{n} a_i - 1 \right) + \lambda_2 \left(\sum_{i=1}^{n} a_i x_i \right)$$

Otimalidade dos estimadores

- As derivadas de interesse são: $\frac{\partial L}{\partial a_i} = 2a_i + \lambda_1 + \lambda_2 x_i$, $\frac{\partial L}{\lambda_1} \sum_{i=1}^{n} a_i - 1$ e $\frac{\partial L}{\lambda_2} \sum_{i=1}^{n} a_i x_i$.
- Assim, obtemos o seguinte sistema de equações:

$$\begin{cases} 2\widetilde{a}_i + \widetilde{\lambda}_1 + \widetilde{\lambda}_2 x_i = 0\,(1), i = 1, 2, ..., n \\ \sum_{i=1}^{n} \widetilde{a}_i = 1\,(2) \\ \sum_{i=1}^{n} \widetilde{a}_i x_i = 0\,(3) \end{cases} \tag{9}$$

- Somando-se as "n" em (1) equações do sistema (9), vem que:

$$2\underbrace{\sum_{i=1}^{n}\widetilde{a}_i}_{1} + n\widetilde{\lambda}_1 + \widetilde{\lambda}_2\sum_{i=1}^{n}x_i = 0 \rightarrow n\widetilde{\lambda}_1 + \widetilde{\lambda}_2\sum_{i=1}^{n}x_i = -2 \quad (10)$$

- Multiplicando as "n" em (1) equações do sistema (9) por x_i e somando-as, vem que:

$$2\underbrace{\sum_{i=1}^{n}\widetilde{a}_i x_i}_{0} + \widetilde{\lambda}_1\sum_{i=1}^{n}x_i + \widetilde{\lambda}_2\sum_{i=1}^{n}x_i^2 = 0 \rightarrow \widetilde{\lambda}_1\sum_{i=1}^{n}x_i + \widetilde{\lambda}_2\sum_{i=1}^{n}x_i^2 = 0$$

$$\rightarrow \quad \widetilde{\lambda}_1 = -\frac{\widetilde{\lambda}_2\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}x_i} \quad (11)$$

- De (11) em (10), temos que:

$$-n\frac{\widetilde{\lambda}_2\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}x_i}+\widetilde{\lambda}_2\sum_{i=1}^{n}x_i=-2\rightarrow\widetilde{\lambda}_2\left[\sum_{i=1}^{n}x_i-n\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}x_i}\right]=-2$$

$$\rightarrow\quad\widetilde{\lambda}_2\left[n\overline{x}-\frac{\sum_{i=1}^{n}x_i^2}{\overline{x}}\right]=-2\rightarrow\widetilde{\lambda}_2=\frac{-2\overline{x}}{n\overline{x}^2-\sum_{i=1}^{n}x_i^2}$$

$$=\quad\frac{2\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}\qquad(12)$$

- De (12) em (11), vem que:

$$\widetilde{\lambda}_1=-\frac{\sum_{i=1}^{n}x_i^2}{\sum_{i=1}^{n}x_i}\frac{2\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2}=-\frac{\overline{x}\sum_{i=1}^{n}x_i^2}{n\sum_{i=1}^{n}(x_i-\overline{x})^2}\qquad(13)$$

- De (12) e (13) na Equação (1) do sistema (9), temos que:

$$
\begin{aligned}
2\widetilde{a}_i &= -\frac{\overline{x}\sum_{i=1}^{n} x_i^2}{n\sum_{i=1}^{n}(x_i-\overline{x})^2} + x_i\frac{2\overline{x}}{\sum_{i=1}^{n}(x_i-\overline{x})^2} \\
\rightarrow \quad \widetilde{a}_i &= \frac{\sum_{i=1}^{n} x_i^2/n - \overline{x}x_i}{\sum_{i=1}^{n}(x_i-\overline{x})^2} = \frac{1}{n} - \frac{\overline{x}(x_i-\overline{x})}{\sum_{i=1}^{n}(x_i-\overline{x})^2}.
\end{aligned}
$$

Veja a equação (3). Assim, o resultado está provado.