Introdução aos modelos de regressão normais lineares

Prof. Caio Azevedo

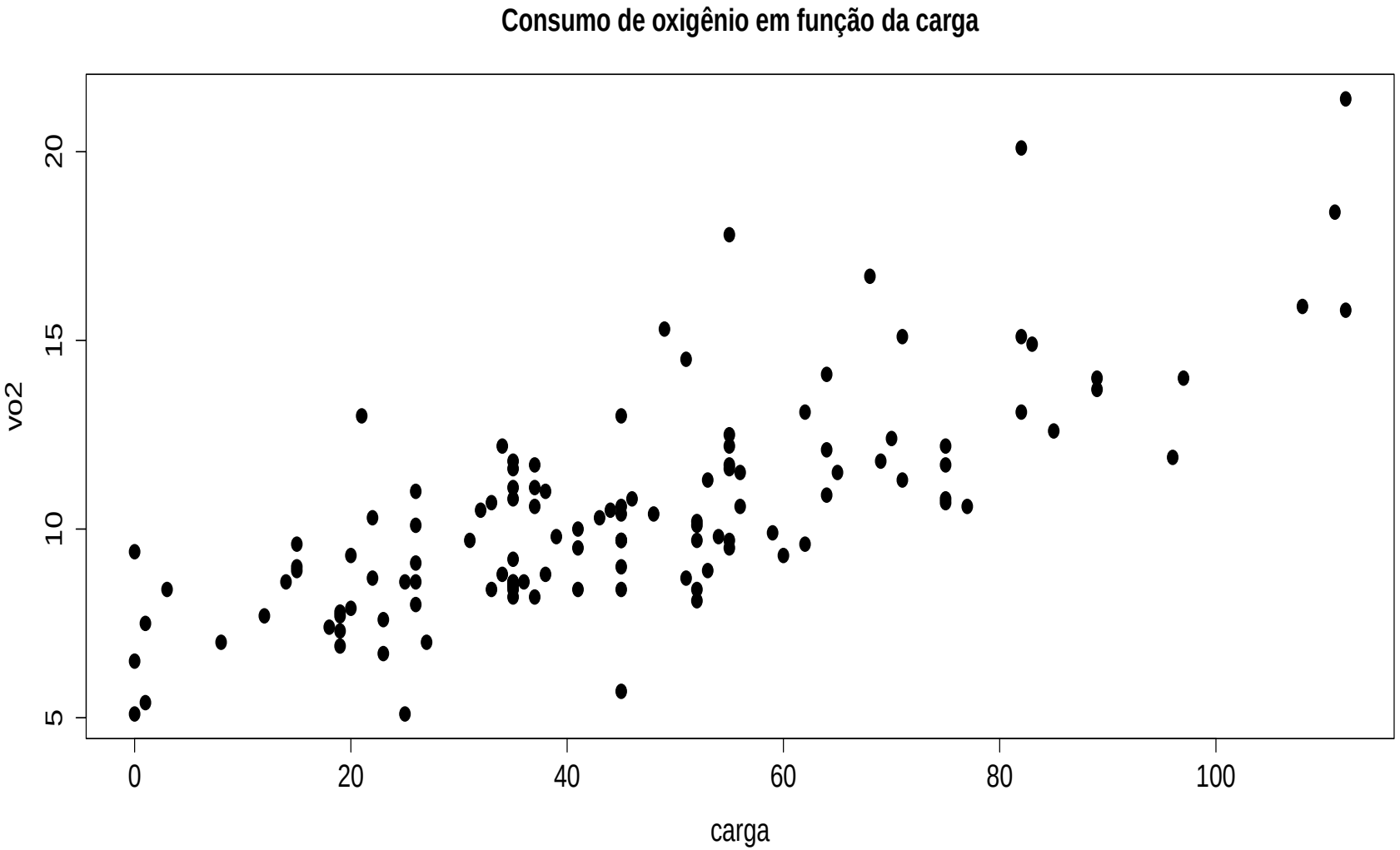
Exemplo 1: Teste de esforço cardiopulmonar

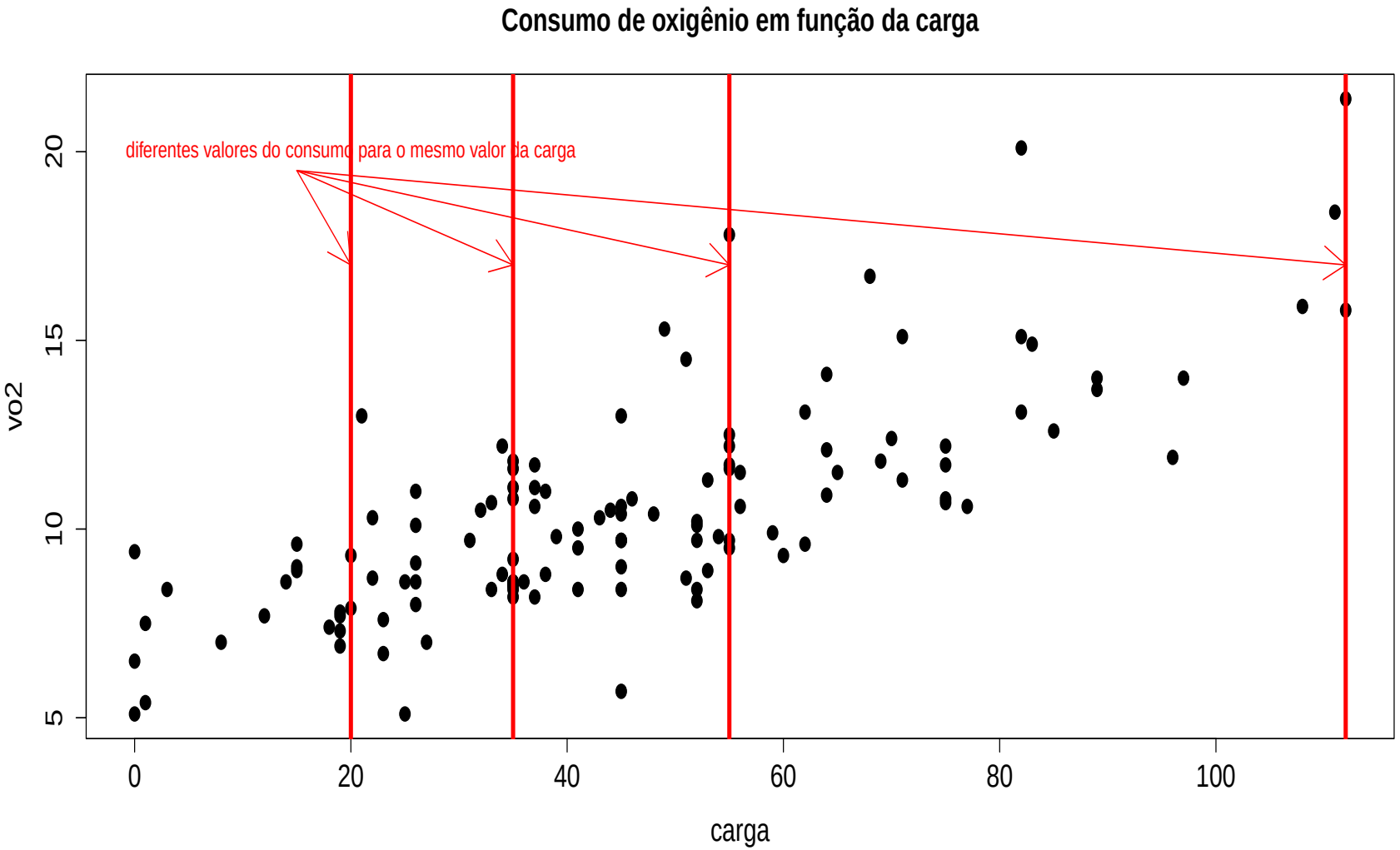
- Considere o estudo sobre teste de esforço cardiopulmonar em pacientes com insufiência cardíaca realizado no InCor da Faculdade de Medicina da USP pela Dra. Ana Fonseca Braga.
- Um dos objetivos do estudo é comparar os grupos formados pelas diferentes etiologias cardíacas quanto às respostas respiratórias 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 (ignorando as etiologias cardíacas, num primeiro momento)?.

Dados

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

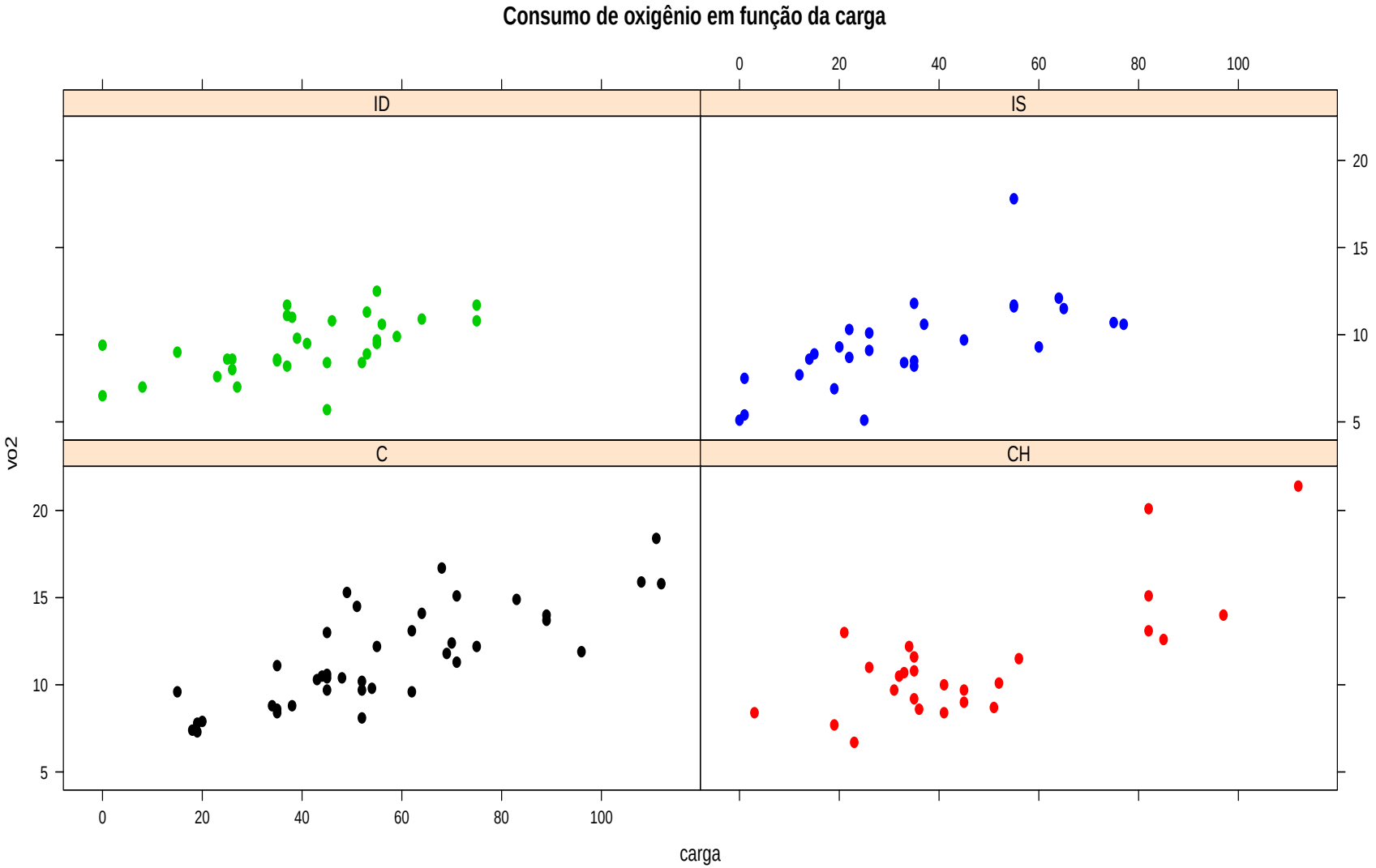
- Existe uma relação entre as duas variáveis?

- 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

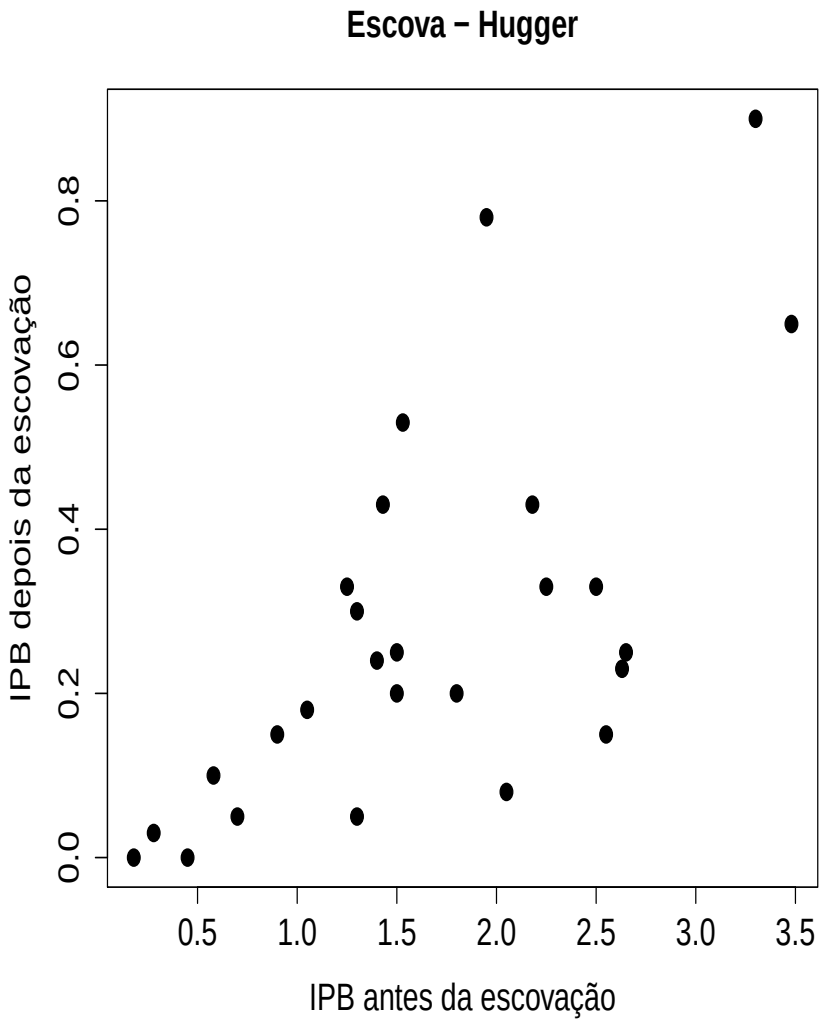
- Considere o seguinte estudo na área de Odontopediatria.
- O objetivo é comparar duas escovas de dente (convencional e experimental, chamada de "hugger") com respeito à redução de um índice de placa bacteriana (IPB) em crianças de ambos os sexos em idade pré-escolar.
- Os valores obtidos correspondem 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.

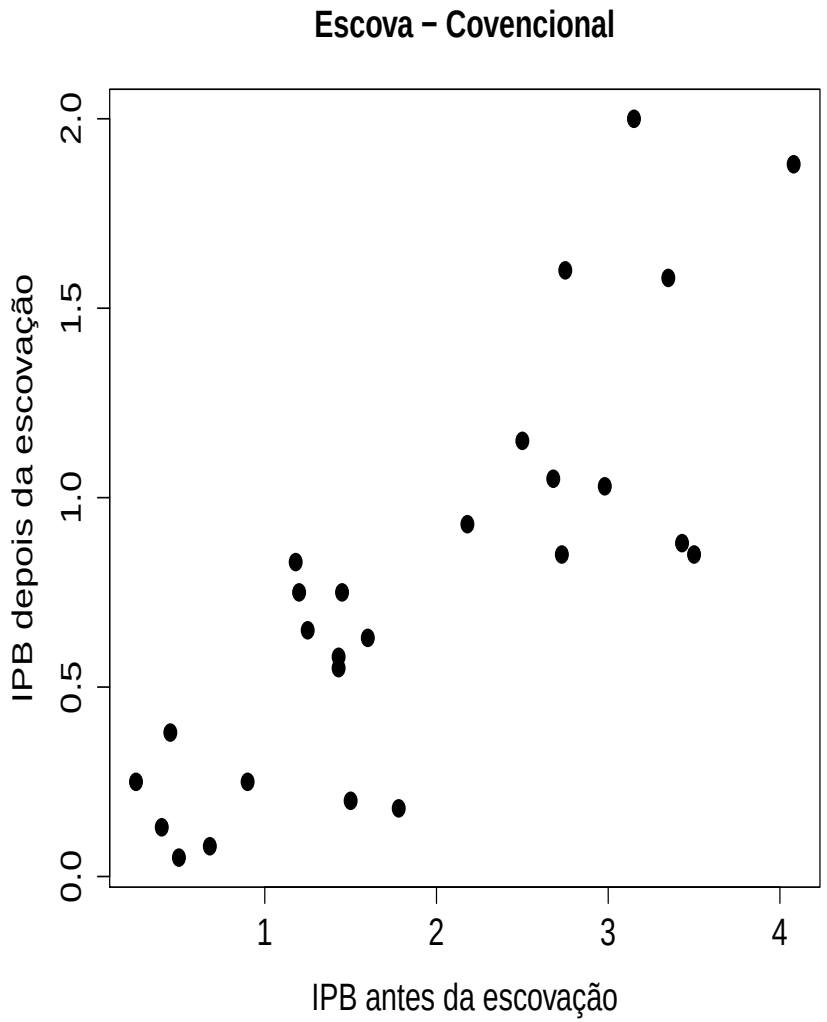
Dados

	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

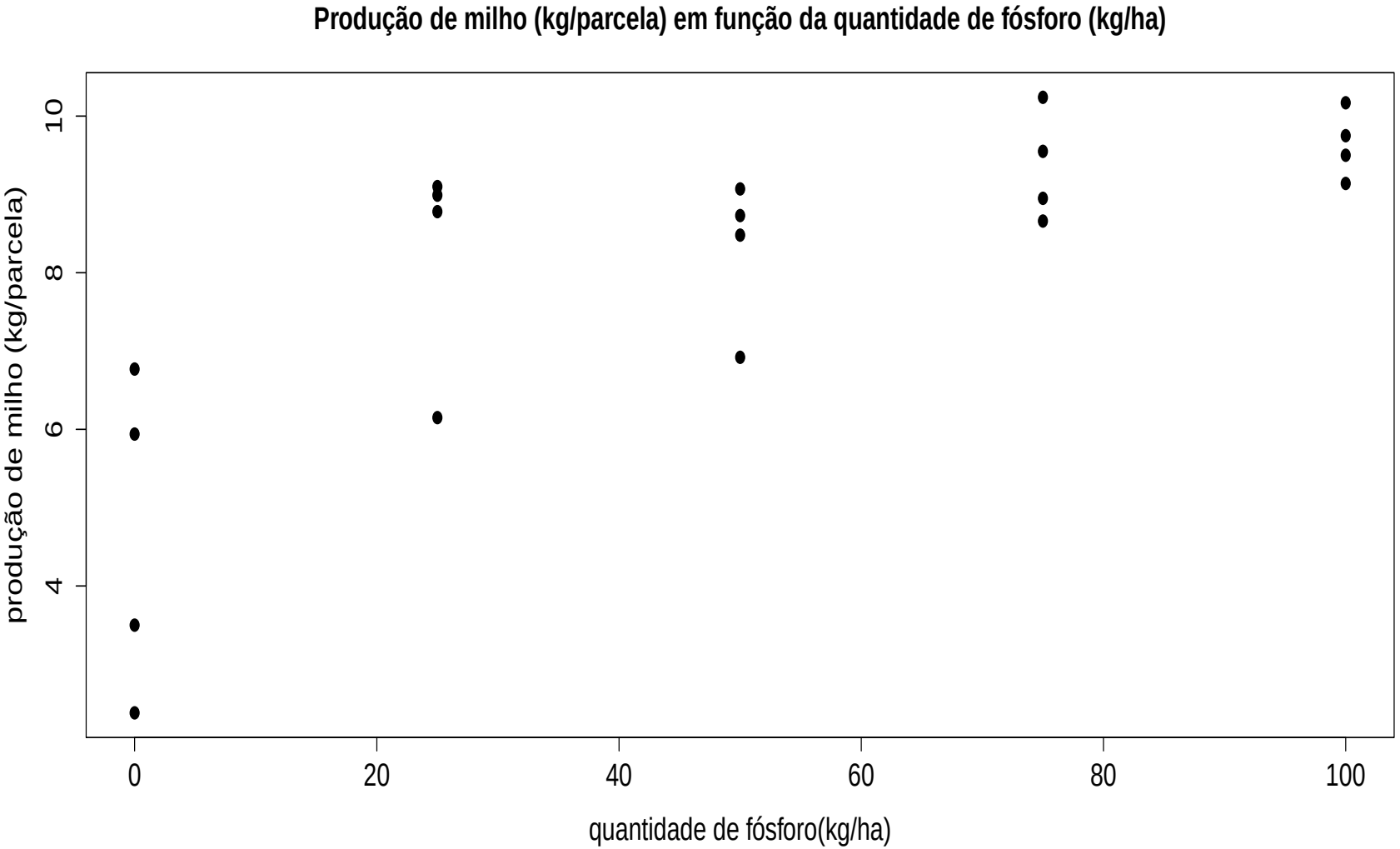
- Como utilizar os IPB's antes e depois ?
- Deve-se considerar a variável sexo?
- O fato de sempre se utilizar o tipo de escova experimental primeiramente pode ter influenciado os resultados?
- Medidas repetidas: cada criança é avaliada duas vezes. Possível existência de dependência entre as observações.


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 níves (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 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

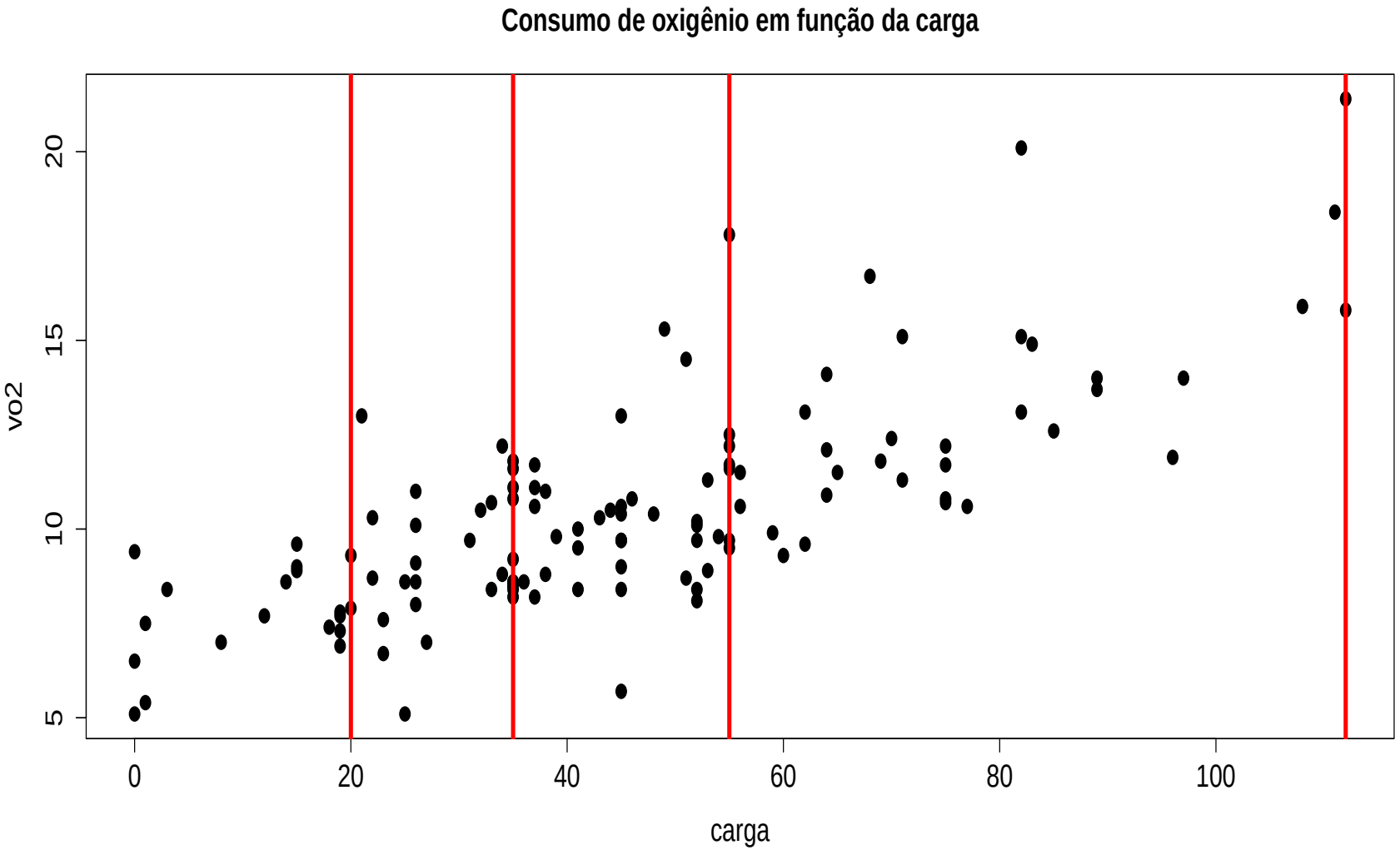
$$Y_i = \beta_0 + \beta_1 x_i + \xi_i, i = 1, ..., 124$$

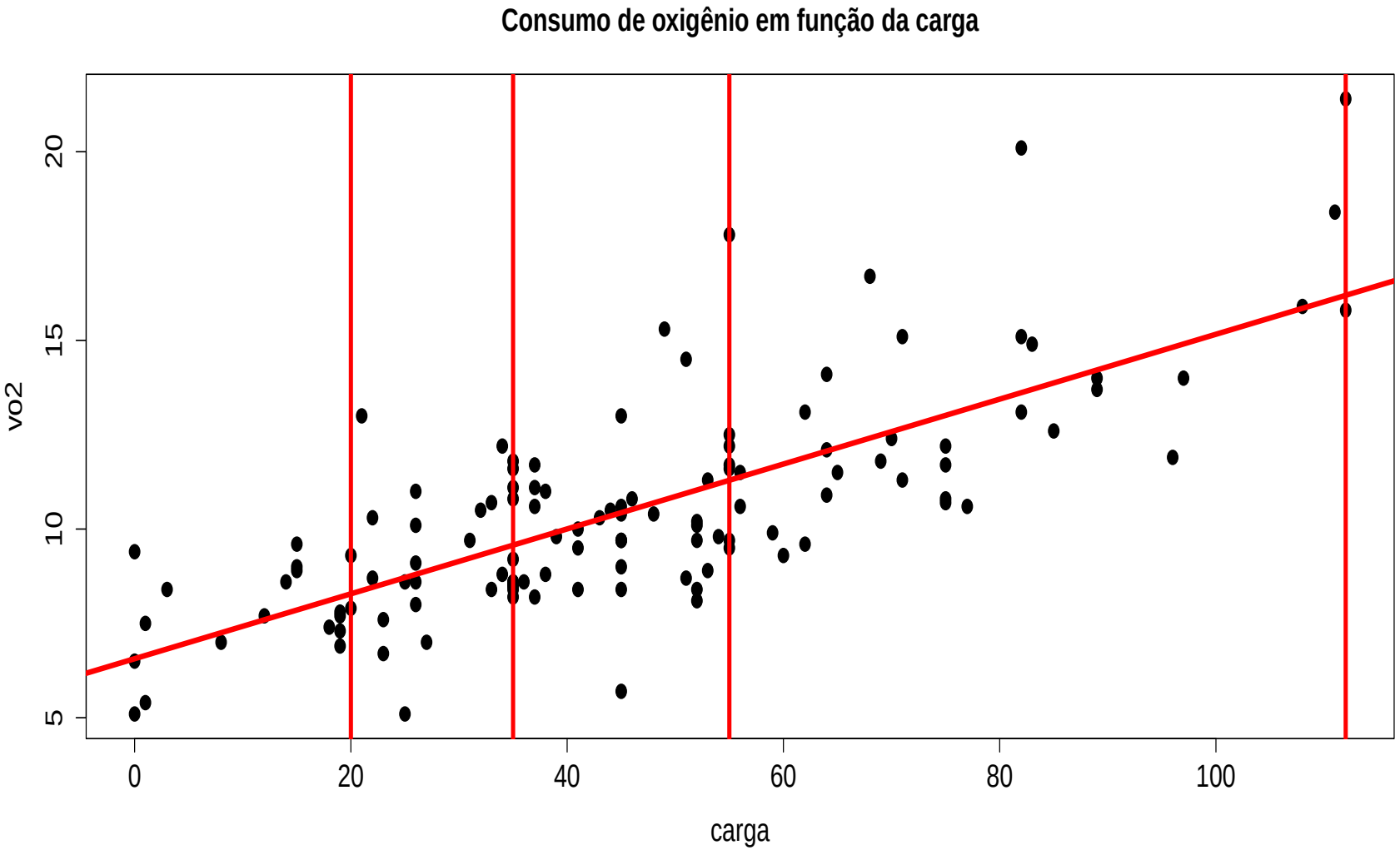
- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $(\beta_0, \beta_1, \sigma^2)'$: parâmetros desconhecidos.
- x_i: carga à que o paciente i foi submetido (conhecida e não aleatória).
- Parte sistemática: $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i$.
- Parte aleatória: ξ_i.
- O modelo acima implica que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

$$Y_{ij} = \beta_{0i} + \beta_{1i}(x_{ij} - \overline{x}) + \xi_{ij}, i = 1, 2(\text{tipo de escova}); j = 1, ..., 26(\text{criança}),$$

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

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

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

- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $(\beta_{01}, \beta_{02}, \beta_{11}, \beta_{12}, \sigma^2)'$: parâmetros desconhecidos.
- x_i : quantidade de fósforo ministrada a i-ésima parcela.
- Parte sistemática: $\mathcal{E}(Y_i) = \beta_{0i} + \beta_1 x_i + \beta_2 x_i^2$.
- 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 o máximo do valor esperado da produção de milho.

Notação matricial para o MNL

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\xi}$$

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \mathbf{X} = \begin{bmatrix} x_{11} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix}$$

- Suposição: $\boldsymbol{\xi} \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ (vetor de erros).
- O índice n da variável resposta é geral e pode representar combinações de índices.

Continuação

- $\mathbf{X}$ é a matriz de plajenamento (ou delineamento) que define a parte sistemática do modelo (conhecida e não aleatória).
- $\mathbf{Y}$ é o vetor associado à variável resposta. Assim, temos que $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$
- Depois dos dados coletados teremos um conjunto $\mathbf{y} = (y_1, ..., y_n)'$ de observações.

Exemplo 1

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{124} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_{124} \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{124} \end{bmatrix}$$

Exemplo 2

$$\mathbf{Y} = \begin{bmatrix} Y_{11} \\ \vdots \\ Y_{1(26)} \\ Y_{21} \\ \vdots \\ Y_{2(26)} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & (x_{11} - \overline{x}) & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & (x_{1(26)} - \overline{x}) & 0 & 0 \\ 0 & 0 & 1 & (x_{21} - \overline{x}) \\ \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & (x_{2(26)} - \overline{x}) \end{bmatrix},$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{01} \\ \beta_{11} \\ \beta_{02} \\ \beta_{12} \end{bmatrix}, \boldsymbol{\xi} = \begin{bmatrix} \xi_{11} \\ \vdots \\ \xi_{1(26)} \\ \xi_{21} \\ \vdots \\ \xi_{2(26)} \end{bmatrix}$$

Exemplo 3

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_{20} \end{bmatrix}, \mathbf{X} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_{20} & x_{20}^2 \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}, \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_{20} \end{bmatrix}$$

Estimação dos parâmetros

- Estimador usual para $\boldsymbol{\beta}$: mínimos quadrados ordinários (MQO).
- Objetivo: obter $\boldsymbol{\beta}$ que minimiza $Q(\boldsymbol{\beta}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})$. Em geral, $\boldsymbol{\beta} \in \mathcal{R}^p$. Assim, para efetuar a minimização, podemos resolver o sistema de equações definido por $\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ (chamada de equações normais).
- Logo, temos que resolver o seguinte sistema:

$$\left.\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \mathbf{0}$$

- Por outro lado, temos que:

$$\begin{aligned}
\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \frac{\partial}{\partial \boldsymbol{\beta}}(\mathbf{Y}'\mathbf{Y} - 2\mathbf{Y}'\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) = -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\
&\rightarrow \left.\frac{\partial Q(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|_{\boldsymbol{\beta}=\widehat{\boldsymbol{\beta}}} = \mathbf{0} \rightarrow -2\mathbf{X}'\mathbf{Y} + 2\mathbf{X}'\mathbf{X}\widehat{\boldsymbol{\beta}} = \mathbf{0} \qquad (1) \\
&\rightarrow \widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}
\end{aligned}$$

desde que $\mathbf{X}'\mathbf{X}$ seja inversível. Como $n >>> p$, tal inversibilidade ocorrerá se, o somente se, a matriz $\mathbf{X}$ tiver posto coluna completo.

- Isto, por sua vez, ocorre quando o modelo está identificado (não está superparametrizado) e/ou quando não há covariáveis que sejam combinações lineares de outras.

- O sistema de equações definido por (1) é consistente, ou seja, apresenta pelo menos uma solução.
- A justificativa não formal para isso é relativamente simples:
  - Se $\mathbf{X}'\mathbf{X}$ for inversível ($rank(\mathbf{X}) = p$), a solução única.
  - Se $\mathbf{X}'\mathbf{X}$ for não inversível ($rank(\mathbf{X}) < p$), podemos considerar alguma inversa generalizada de $\mathbf{X}'\mathbf{X}$. Neste caso, o sistema pode apresentar infinitas soluções e as funções estimáveis passam a ter uma importância maior do que os parâmetros isoladamente.
  - No último caso, uma solução é dada por $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-}\mathbf{X}'\mathbf{Y}$.
- Em geral, vamos trabalhar com modelos em que a solução é única.

Propriedades do Estimador de MQO (exercício)

- Uma vez que $\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{Y}$, $\mathbf{Y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2\mathbf{I}_n)$ e pelas propriedades associados à vetores aleatórios e a distribuição normal multivariada, temos que:
  - $\mathcal{E}(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathcal{E}(\mathbf{Y}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} = \boldsymbol{\beta}$. (não viciado).
  - $Cov(\widehat{\boldsymbol{\beta}}) = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'Cov(\mathbf{Y})\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$.
  - $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\mathbf{X}'\mathbf{X})^{-1})$ (normalidade).
  - $\widehat{\beta}_j \sim N_p(\beta_j, \sigma^2\psi_j)$, ψ_j é o j-ésimo elemento da diagonal principal da matriz $\mathbf{X}'\mathbf{X}^{-1}$.
  - Observação: sob a suposição de normalidade, o estimador de MQO coincide com o estimador de MV (máxima verossimilhança).

Estimador de σ^2

- Sob normalidade, o estimador de máxima verosimilhança de σ^2 é dado por

$$\widehat{\sigma}^2_{MV} = \frac{1}{n}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

  o qual é viciado.

- Na prática considera-se o seguinte estimador:

$$\widehat{\sigma}^2 = \frac{1}{n-p}\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)'\left(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\beta}}\right)$$

- O qual é não-viciado. Além disso, pode-se provar que $\widehat{\boldsymbol{\beta}} \perp \widehat{\sigma}^2$ e $\frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-p)}$ (exercício).