Breve revisão sobre o MRNLH (Modelo de regressão normal linear homocedástico)

Prof. Caio Azevedo

Exemplo 5: 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(m) ser utilizada(s) como fator prognóstico de óbito.
- Os dados (em sua íntegra) podem ser encontrados aqui.

Cont.

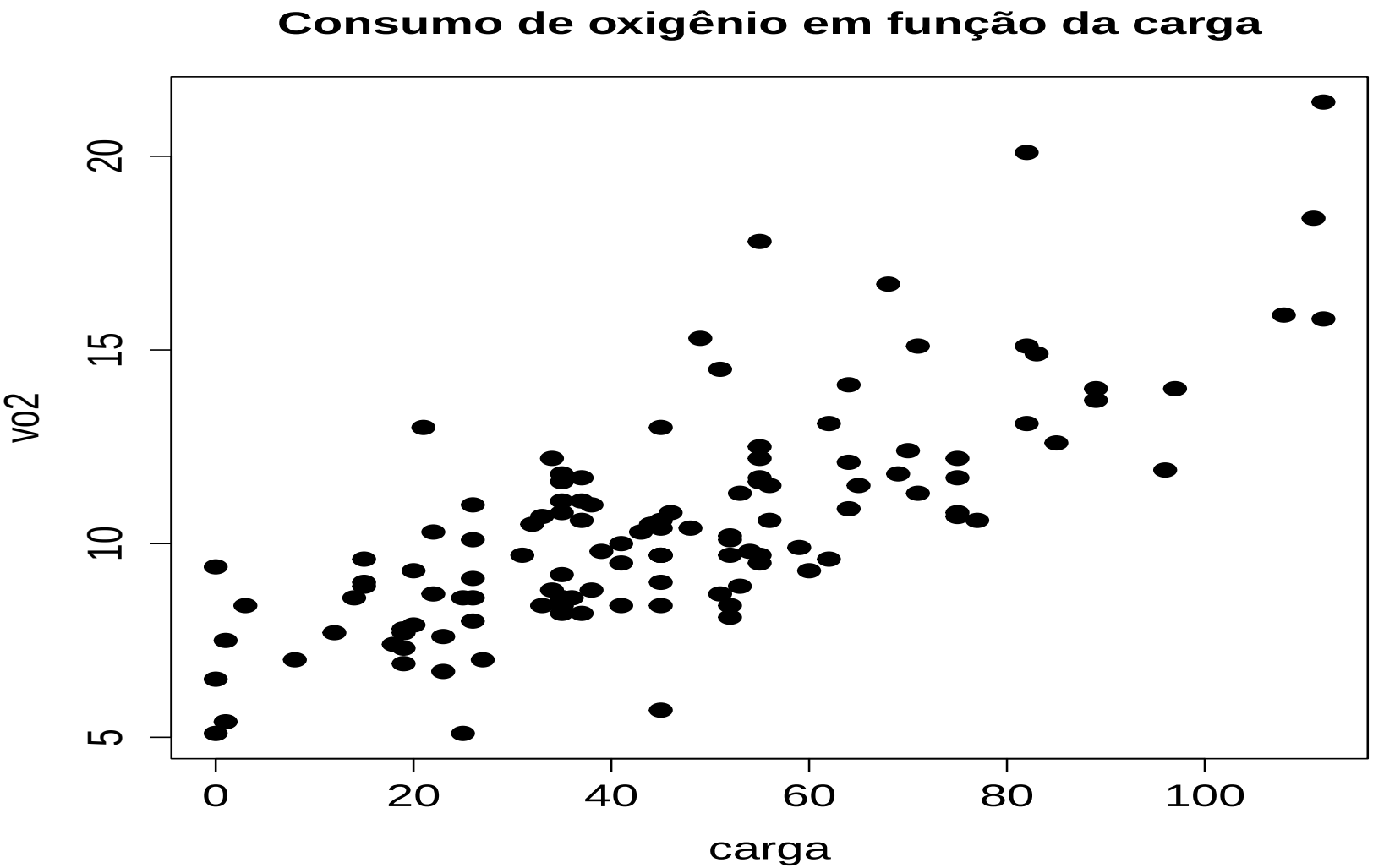
- Etiologias : CH: chagásicos, ID: idiopáticos, IS: isquêmicos, C: controle.
- Considere que o objetivo é explicar a variação do consumo de oxigênio no limiar anaeróbio (ml/kg/min - mililitros por quilograma de peso por minuto), em função da carga utilizada na esteira ergométrica para pacientes com diferentes 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.

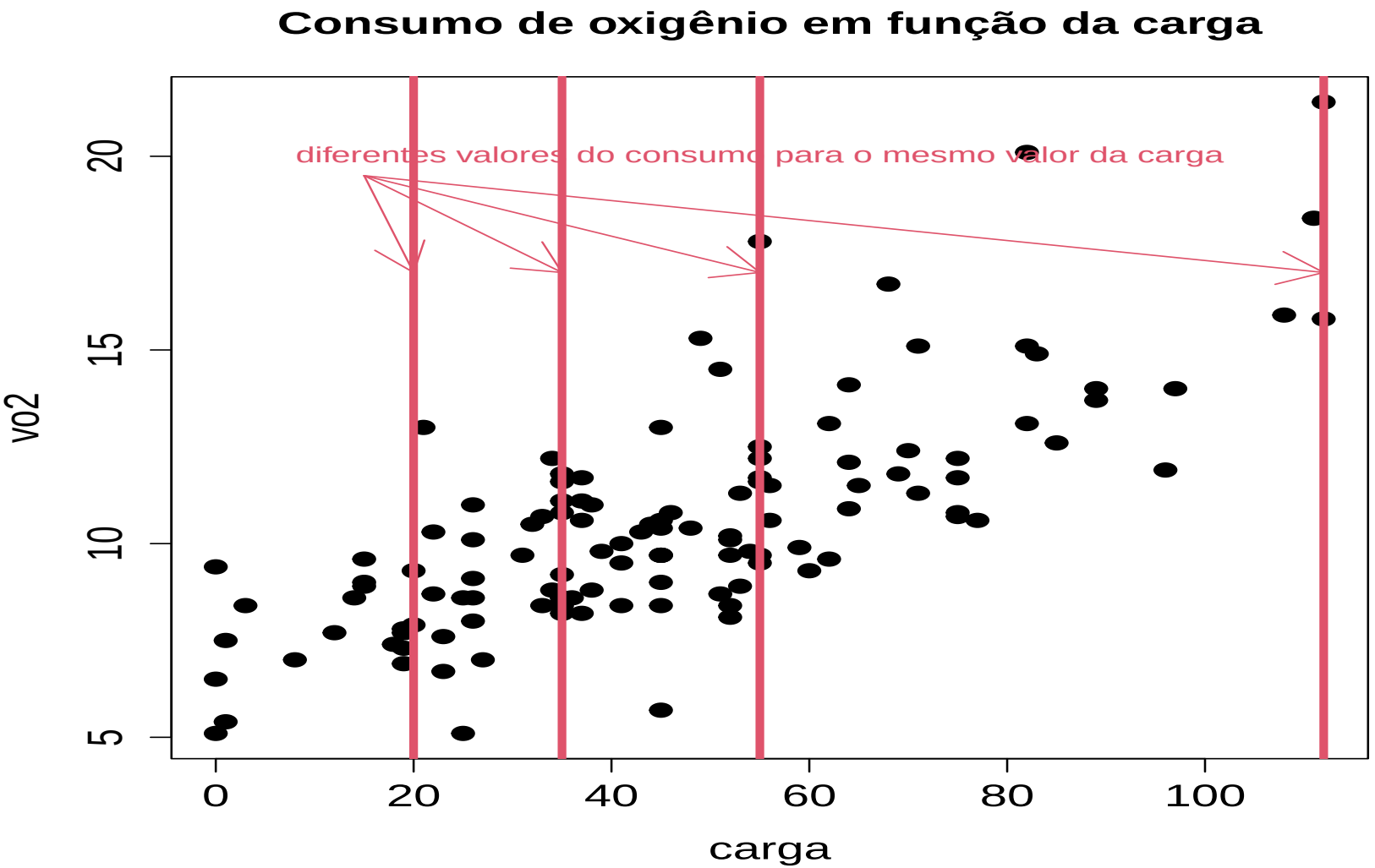
Cont.

- Aeróbio (com a utilização de oxigênio) ; anaeróbio (sem a utilização de oxigênio).
- Como responder a pergunta de interesse (ignorando as etiologias cardíacas, num primeiro momento)?.
- Detalhes adicionais aos que apresentaremos poderm ser encontrados aqui e aqui.

Dados

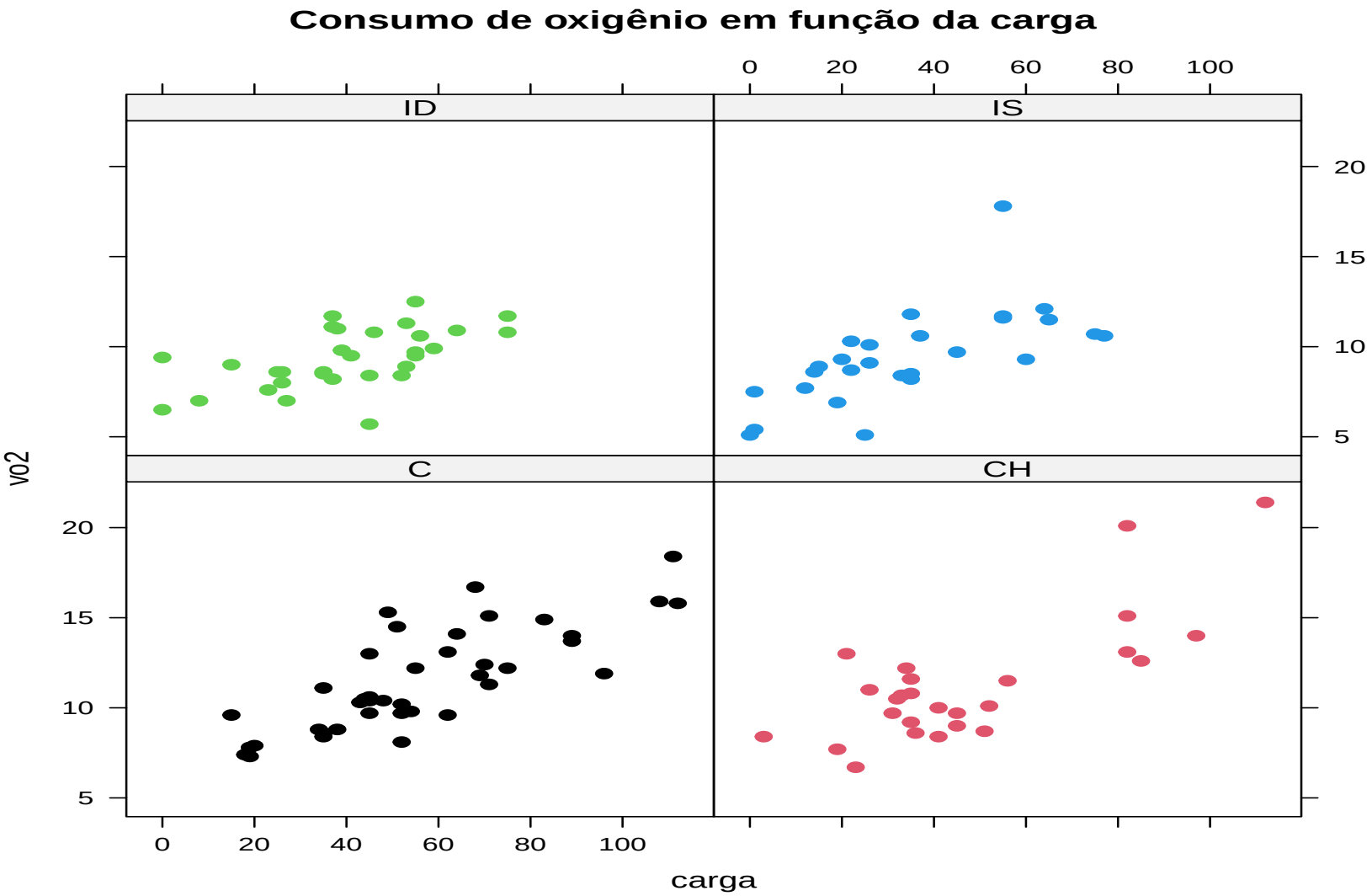
ID	Etiologia	Carga	VO2
1	CH	41	10,0
2	CH	56	11,5
3	ID	8	7,0
4	ID	53	8,9
⋮	⋮	⋮	
7	ID	0	6,5
⋮	⋮	⋮	
123	C	64	14,1
124	C	70	12,4


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
5
10
15
20
0
20
40
60
80
100
carga

Cont.

- 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

Cont.

- 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 5: desconsiderando as etiologias cardíacas

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

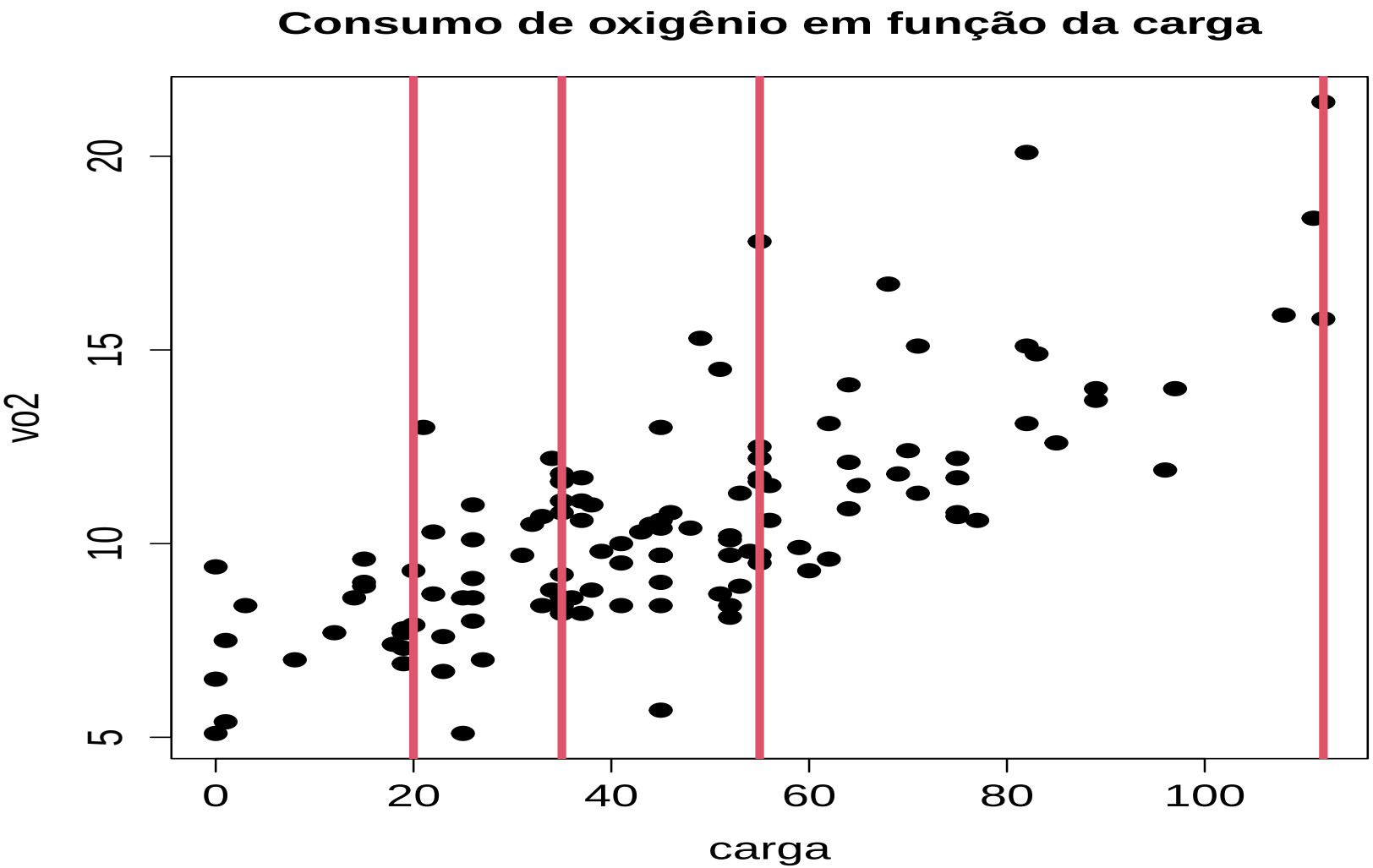
- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $(\beta_0, \beta_1, \sigma^2)'$: parâmetros desconhecidos.
- x_i: carga à que o paciente i foi submetido (conhecida e não aleatória).
- Parte sistemática: $\mathcal{E}(Y_i) = \beta_0 + \beta_1 x_i$.
- Parte aleatória: ξ_i.
- O modelo acima implica que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

Cont.

- 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
carga
5
10
15
20
0
20
40
60
80
100

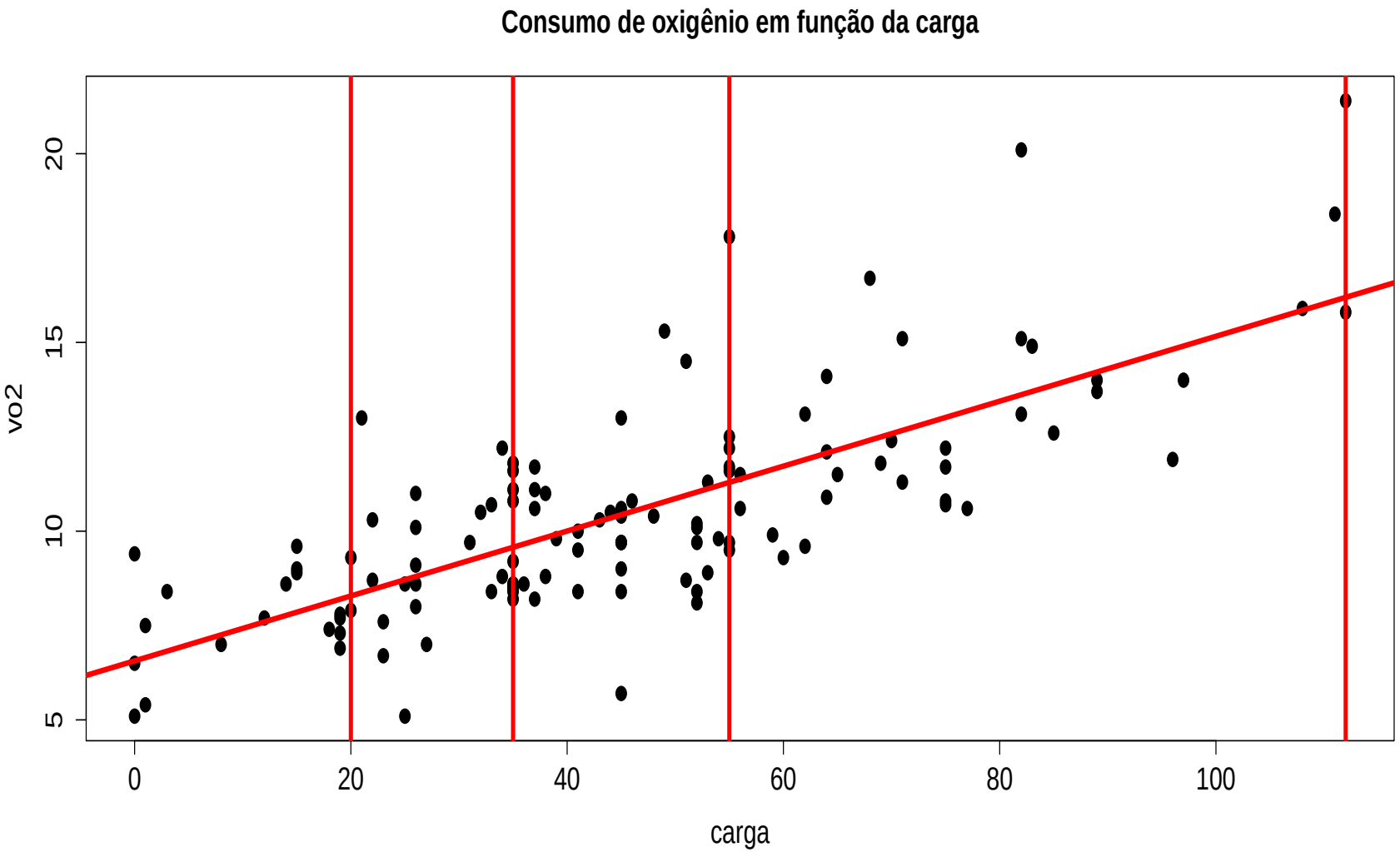
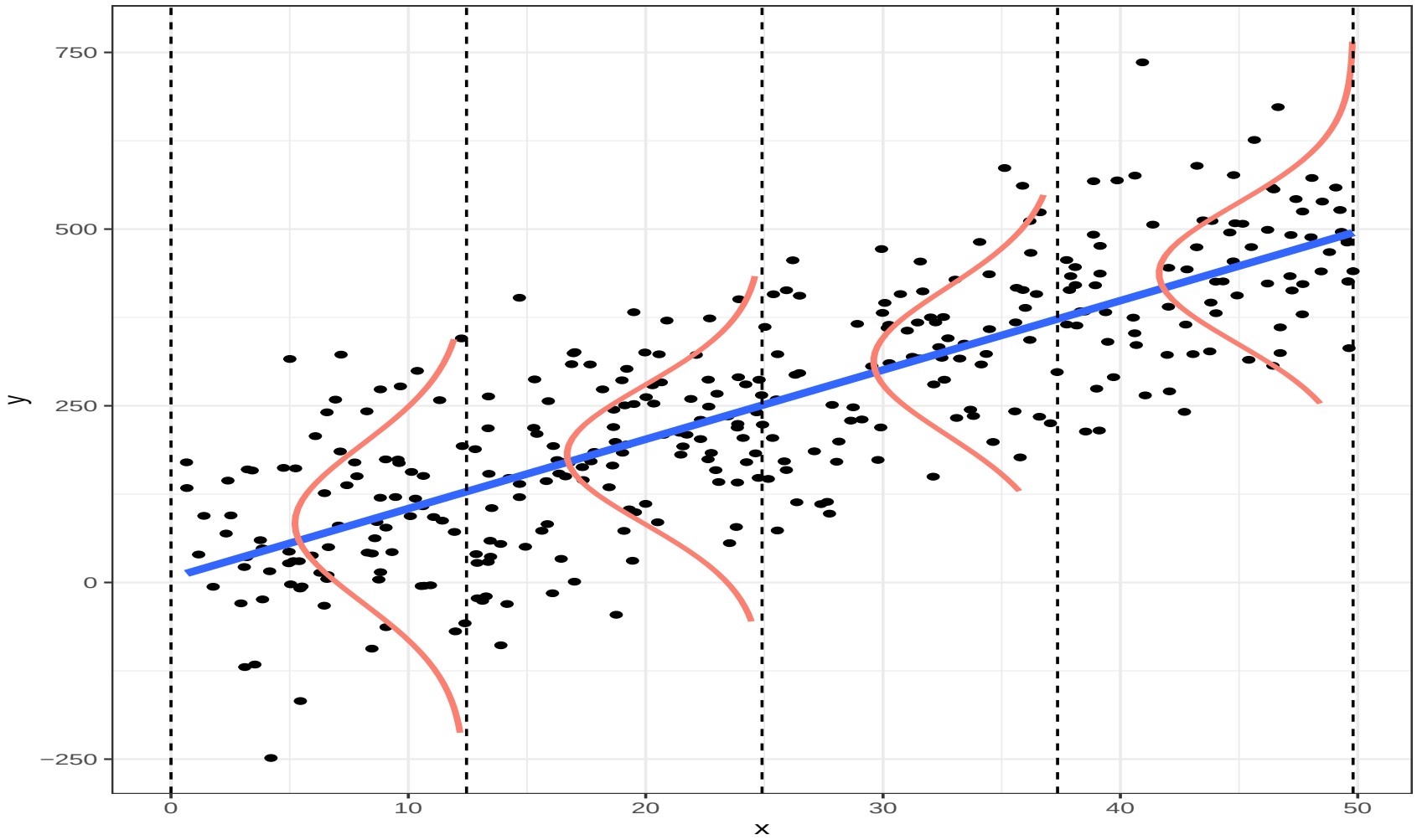

Consumo de oxigênio em função da carga
vo2
carga
5
10
15
20
0
20
40
60
80
100

Ilustração de um MRNLH



Notação matricial para o MNL

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

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

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

Continuação

- $\boldsymbol{X}$ é a matriz de plajenamento (ou delineamento) que define a parte sistemática do modelo (conhecida e não aleatória).
- $\boldsymbol{Y}$ é o vetor associado à variável resposta. Assim, temos que

$$\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n).$$

- Depois dos dados coletados teremos um conjunto $\boldsymbol{y} = (y_1, ..., y_n)'$ de observações.

Exemplo 5

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

Estimação dos parâmetros

- Estimador usual para $\boldsymbol{\beta}$: mínimos quadrados ordinários (MQO).
- Objetivo: obter $\boldsymbol{\beta}$ que minimiza $Q(\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})$. 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:

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

Cont.

- Por outro lado, temos que:

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

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

Cpont.

- 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 $\boldsymbol{X}'\boldsymbol{X}$ for inversível ($rank(\boldsymbol{X}) = p$), a solução é única.
  - Se $\boldsymbol{X}'\boldsymbol{X}$ for não inversível ($rank(\boldsymbol{X}) < p$), podemos considerar alguma inversa generalizada de $\boldsymbol{X}'\boldsymbol{X}$. Neste caso, o sistema pode apresentar infinitas soluções e as funções estimáveis passam a ter uma importância maior do que os parâmetros isoladamente.
  - Nesse último caso, uma solução é dada por $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-}\boldsymbol{X}'\boldsymbol{Y}$.
- Em geral, vamos trabalhar, no curso de MLG, com modelos em que a solução é única.

Propriedades do Estimador de MQO (exercício)

- Uma vez que $\widehat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$, $\boldsymbol{Y} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n)$ e pelas propriedades associados à vetores aleatórios e a distribuição normal multivariada, temos que:
  - $\mathcal{E}(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\mathcal{E}(\boldsymbol{Y}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta}$. (não viciado).
  - $Cov(\widehat{\boldsymbol{\beta}}) = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'Cov(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
  - $\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ (normalidade).
  - $\widehat{\beta}_j \sim N_1(\beta_j, \sigma^2\psi_j)$, ψ_j é o j-ésimo elemento da diagonal principal da matriz $(\boldsymbol{X}'\boldsymbol{X})^{-1}$.
  - 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

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

  o qual é viciado.

- Na prática considera-se o seguinte estimador:

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

  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).

Inferência

- Adicionalmente, temos que

$$\widehat{\beta}_j \sim N(\beta_j, \sigma^2\psi_j); \frac{(n-p)\widehat{\sigma}^2}{\sigma^2} \sim \chi^2_{(n-p)}; \widehat{\beta}_j \perp \frac{(n-p)\widehat{\sigma}^2}{\sigma^2}, j = 1, 2, ..., p.$$

- Logo,

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}^2\psi_j}} \sim t_{(n-p)}, j = 1, ..., p,$$

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

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

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.
- Estatística do teste

$$T_t = \frac{\widehat{\beta}_j - \beta_{j0}}{\sqrt{\widehat{\sigma}^2 \psi_j}},$$

em que $\widehat{\beta}_j$ é o estimador de MQO de β_j e

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

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 $P(X \geq t_c | H_0) = \alpha/2, X \sim t_{(n-p)}$, $\widetilde{\beta}_j$ é o j-ésimo elemento de $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{y}$ e $\widetilde{\sigma}^2 = \frac{1}{n-p}\left(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X}\widetilde{\boldsymbol{\beta}}\right)$
- De modo equivalente, rejeita-se H_0 se p-valor $\leq \alpha$, em que p-valor $= 2P(X \geq |t_t||H_0)$, $X \sim t_{(n-p)}$.
- Para o MRNLH, a grande maioria dos resultados inferenciais são exatos.