# Mais sobre diagnóstico de modelos de regressão normais lineares

Prof. Caio Azevedo

(grande parte do material apresentado foi extraído do livro Modelos de regressão com apoio computacional do Prof. Gilberto A. Paula)

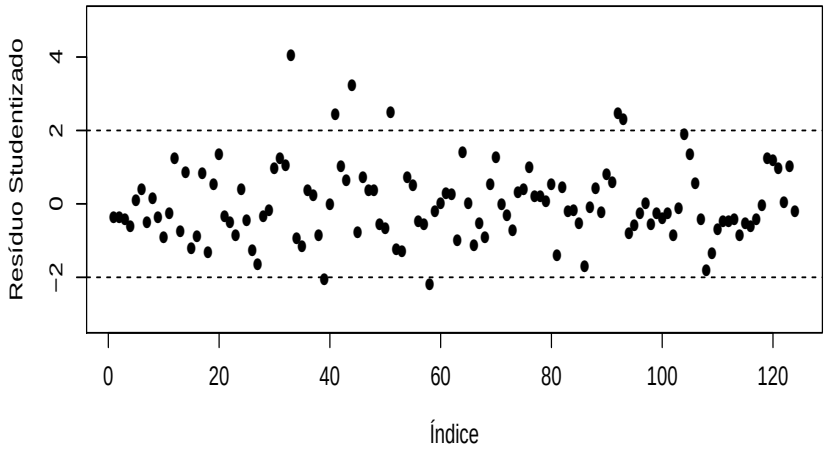
https://www.ime.usp.br/~giapaula/texto_2013.pdf

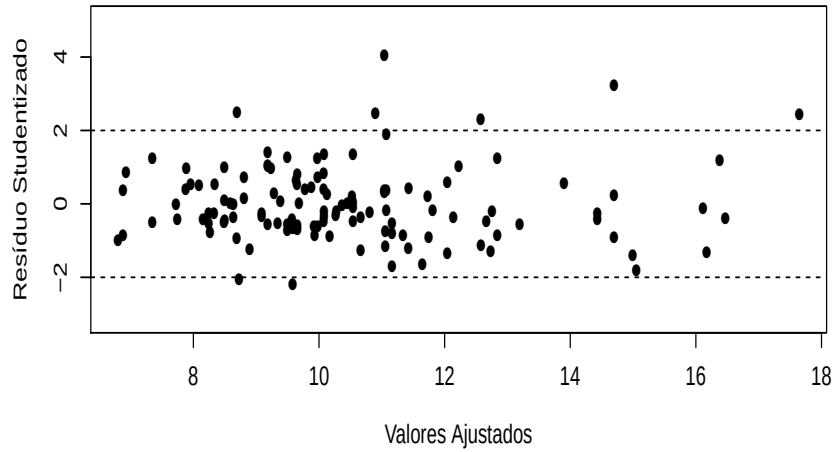
# Exemplo 1: considerando as etiologias cardíacas

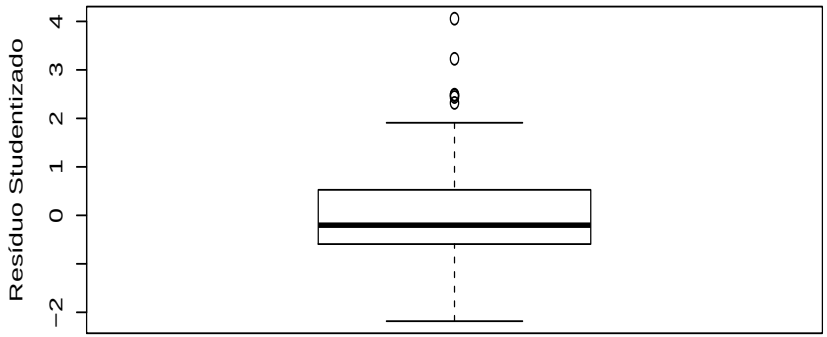
$$Y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \xi_{ij}, i = 1, ..., ; j = 1, ..., n_i$$

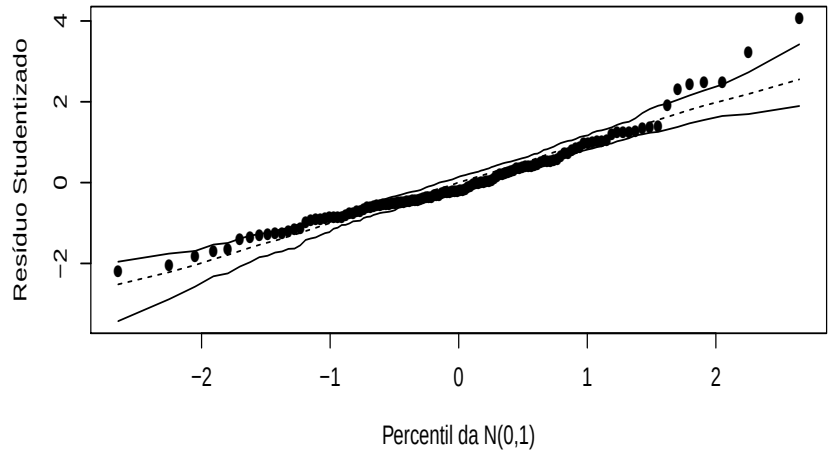
- Etiologias : CH ($i = 1$), ID ($i = 2$), IS ($i = 3$), C: ($i = 4$).
- $\xi_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $x_{ij}$: carga à que o paciente $j$ que apresenta a etiologia cardíaca $i$ foi submetido (conhecido e não aleatório).
- $\beta_{0i}$ : consumo esperado para pacientes da i-ésima etiologia submetidos à uma carga igual a 0.
- $\beta_{1i}$ : incremento (positivo ou negativo) no consumo esperado, de pacientes da i-ésima etiologia, para o aumento em uma unidade da carga.

# Gráficos de resíduos

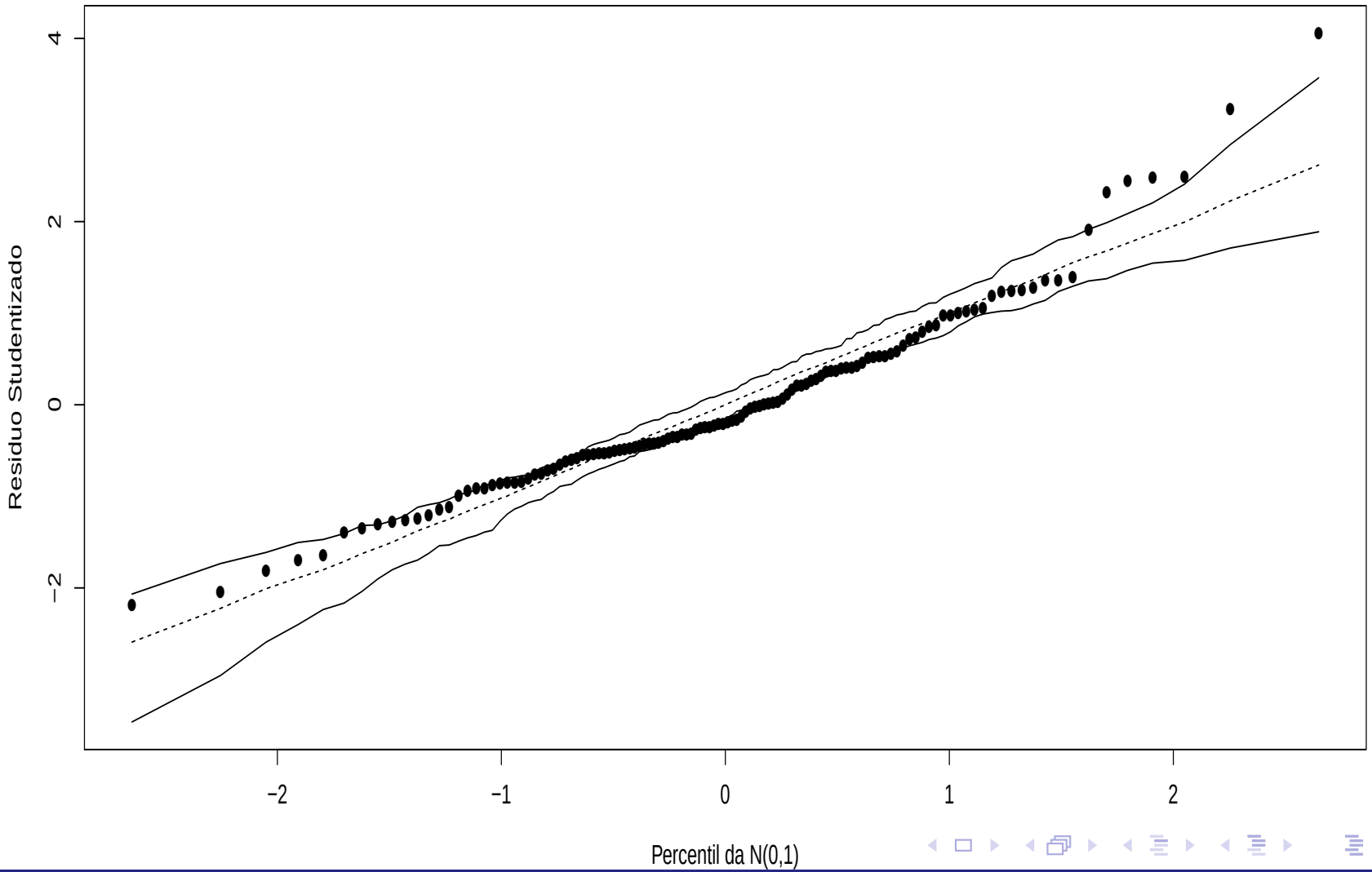


# Gráfico de envelopes para os resíduos



# Objetivos

- Vamos estudar algumas medidas que avaliam o quão cada observação influencia (pode influenciar) as estimativas dos parâmetros de interesse $(\boldsymbol{\beta}', \sigma^2)'$.
- Assumiremos que o modelo de regressão normal linear homocedástico se ajustou bem aos dados.
- Assim, as técnicas que veremos não se aplicam a avaliar se o modelo de ajustou bem ou não aos dados, mas somente em identificar as observações (possivelmente) influentes segundo algum critério.
- Estudaremos apenas duas medidas: de "alavancagem" e a distância de Cook.

# Medida de alavancagem (pontos alavancas) "'h'

- O resíduo oridinário $R_i = Y_i - \widehat{Y}_i$ mede a discrepância entre o valor observado e o ajustado (em que o sinal indica a direção dessa discrepância).
- Matricialmente (como já vimos) temos que : $\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}$.
- Defina $h_{ii} = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i$, em que $\boldsymbol{x}_i'$ é a i-esima linha da matriz $\boldsymbol{X}$.
- A matriz $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ é simétrica e idempotente tal que $tr(\boldsymbol{H}) = r(\boldsymbol{H}) = p = \sum_{i=1}^{n} h_{ii}$. Além disso, pode-se demonstrar que $\frac{1}{n} \leq h_{ii} \leq \frac{1}{c}$, em que $c$ é o número de linhas de $\boldsymbol{X}$ que são idênticas à $\boldsymbol{x}_i'$.

# Medida de alavancagem (pontos alavancas) "'h'

- Note que:

$$\boldsymbol{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \ddots & \vdots & \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1' & \boldsymbol{x}_2' & \ldots & \boldsymbol{x}_n' \end{bmatrix}$$

- Assim (próximo slide):

# Medida de alavancagem (pontos alavancas) "'h'

$$\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \begin{bmatrix} \boldsymbol{x}_1' \\ \boldsymbol{x}_2' \\ \dots \\ \boldsymbol{x}_n' \end{bmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1} \begin{bmatrix} \boldsymbol{x}_1 & \boldsymbol{x}_2 & \dots & \boldsymbol{x}_n \end{bmatrix}$$

$$= \begin{bmatrix} \boldsymbol{x}_1'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_1 & \boldsymbol{x}_1'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_1'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_n \\ \boldsymbol{x}_2'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_1 & \boldsymbol{x}_2'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_2'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_n \\ \dots & \dots & \ddots & \dots \\ \boldsymbol{x}_n'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_1 & \boldsymbol{x}_n'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_2 & \dots & \boldsymbol{x}_n'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_n \end{bmatrix}$$

$$= \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1n} \\ h_{21} & h_{22} & \dots & h_{2n} \\ \dots & \dots & \ddots & \dots \\ h_{n1} & h_{n2} & \dots & h_{nn} \end{bmatrix}$$

# Medida de alavancagem (pontos alavancas) "'h'

- Lembremos que $\widehat{\boldsymbol{Y}} = \boldsymbol{HY}$.
- Pode-se provar que o i-ésimo valor ajustado é dado por:

$$\widehat{Y}_i = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j \rightarrow \widetilde{y}_i = h_{ii} y_i + \sum_{j \neq i} h_{ij} y_j$$

  e, que por $\boldsymbol{H}$ ser idempotente, temos que $\sum_{j \neq i} h_{ij}^2 = h_{ii}(1 - h_{ii})$
- Assim, se $h_{ii} = 1$, então $\widehat{Y}_i = Y_i$ (embora a recíproca não seja verdadeira).
- Portanto, quanto maior for o valor de $h_{ii}$ mais influente é o valor da observação $y_i$ sobre o correspondente valor ajustado.

# Medida de alavancagem (pontos alavancas) “’h’

- Note ainda que $\frac{\partial \widetilde{y}_i}{\partial y_i} = h_{ii}$. Assim, $h_{ii}$ corresponde à variação em $\widetilde{y}_i$ quando $y_i$ é acrescido de um infinitésimo.
- Portanto, se todos os pontos exercerem a mesma influência sobre os valores ajustados, espera-se que que $h_{ii}$ esteja próximo de $\frac{tr(\boldsymbol{H})}{n} = \frac{p}{n}$ (média aritmética).
- Heuristicamente, se $h_{ii} \geq \frac{2p}{n}$, considera-se que a observação $i$ como ponto de alavancagem ou alavanca.
- Esses pontos também podem influenciar à estimativas de $\boldsymbol{\beta}$ e $\sigma^2$.

# Distância de Cook

- Logverossimilhança do modelo :

$$l_\delta(\boldsymbol{\beta}, \sigma^2; y) = \sum_{i=1}^{n} \delta_i l(\boldsymbol{\beta}, \sigma^2; y_i) \qquad (1)$$

em que $0 \leq \delta_i \leq 1, i = 1, 2, ..., n$ é um tipo de perturbação.

- Estimativa de MQ sob (1):

$$\widehat{\boldsymbol{\beta}}_{\boldsymbol{\Delta}} = (\boldsymbol{X}'\boldsymbol{\Delta}\boldsymbol{X})^{-1}\boldsymbol{X}\boldsymbol{\Delta}\boldsymbol{Y},$$

em que $\boldsymbol{\Delta} = \text{diag}(\delta_1, ..., \delta_n)$.

- Em geral para se verificar o quão influente é a i-ésima observação faz-se, para $i = 1, 2, ..., n$, $\delta_i = 0$ e $\delta_k = 1, \forall k \neq i$. Ou seja, verifica-se o impacto que a retirada da i-ésima observação provoca nas estimativas.
- Nesse caso, temos que

$$\widehat{\boldsymbol{\beta}}_{(i)} = \widehat{\boldsymbol{\beta}} - \frac{R_i}{1 - h_{ii}} (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{x}_i$$

- A medida de Cook (relativa à retirada da i-ésima observação), é dada por

$$D_i = \frac{\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}'_{(i)}\right) (\boldsymbol{X}'\boldsymbol{X})^{-1} \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right)}{p\widehat{\sigma}^2}$$

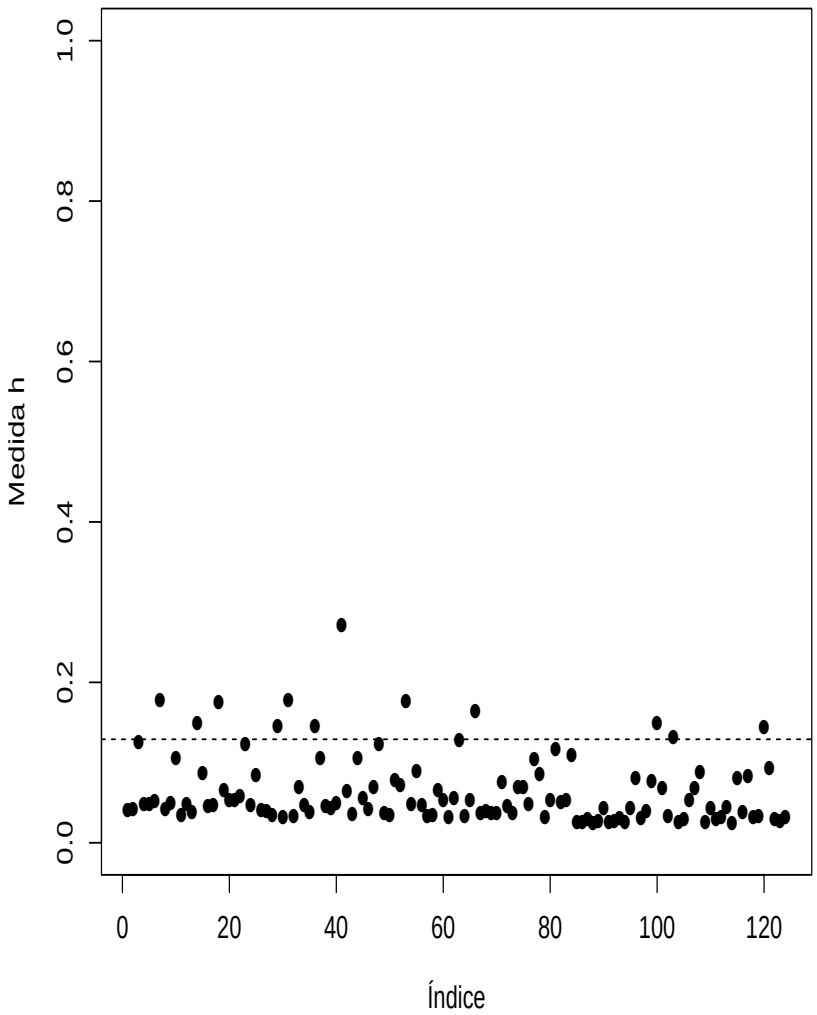
em que $\widehat{\sigma}^2 = \frac{1}{n-p} \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)' \left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$

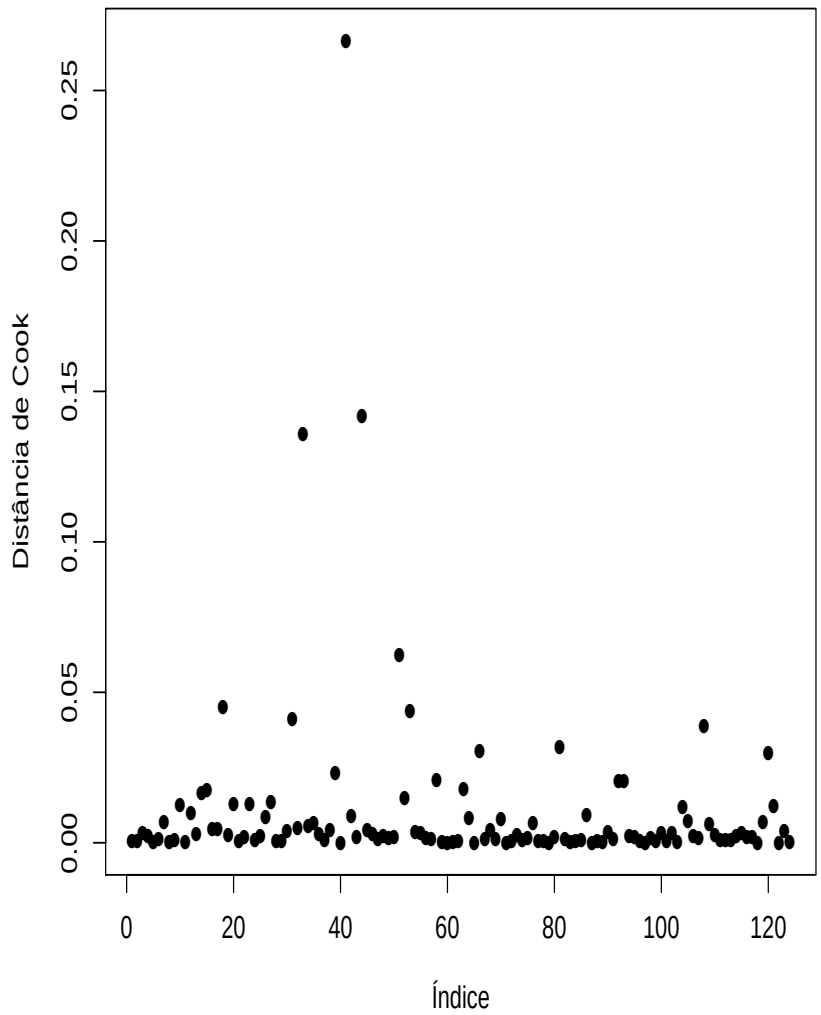
# Exemplo 1: considerando as etiologias cardíacas

$$Y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \xi_{ij}, i = 1, ..., ; j = 1, ..., n_i$$

- Etiologias : CH ($i = 1$), ID ($i = 2$), IS ($i = 3$), C: ($i = 4$).
- $\xi_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $x_{ij}$: carga à que o paciente $j$ que apresenta a etiologia cardíaca $i$ foi submetido (conhecido e não aleatório).
- $\beta_{0i}$ : consumo esperado para pacientes da i-ésima etiologia submetidos à uma carga igual a 0.
- $\beta_{1i}$ : incremento (positivo ou negativo) no consumo esperado, de pacientes da i-ésima etiologia, para o aumento em uma unidade da carga.

# Pontos alavanca e distância de Cook



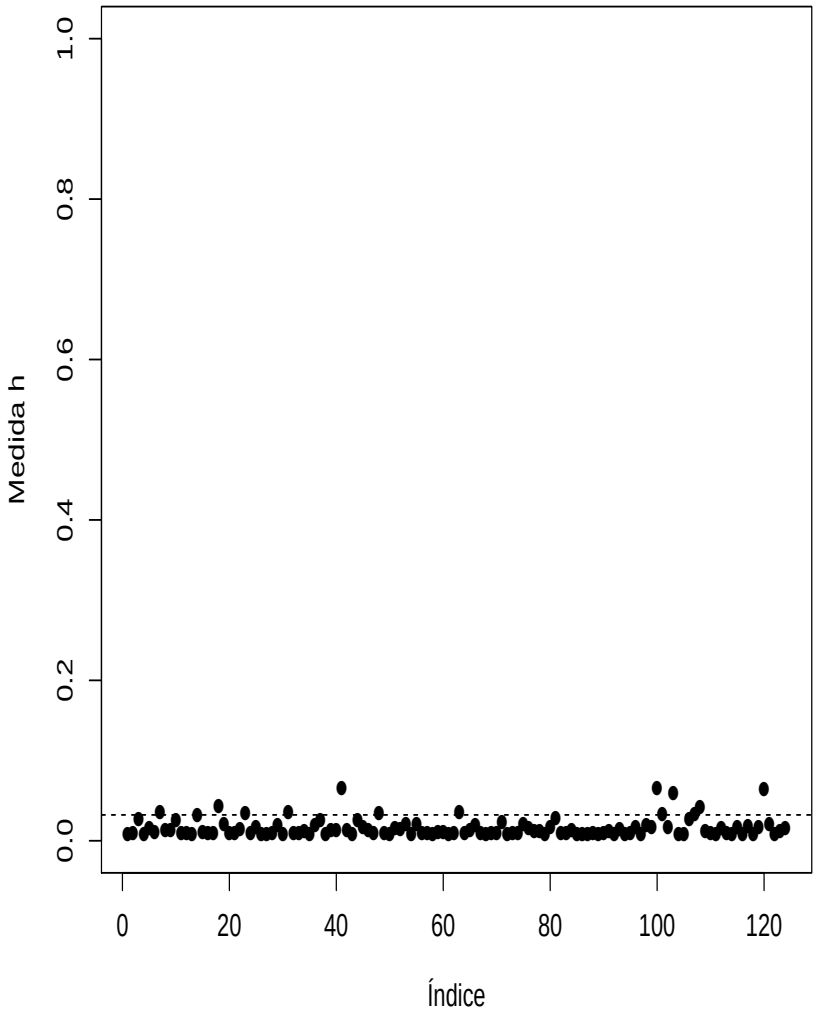



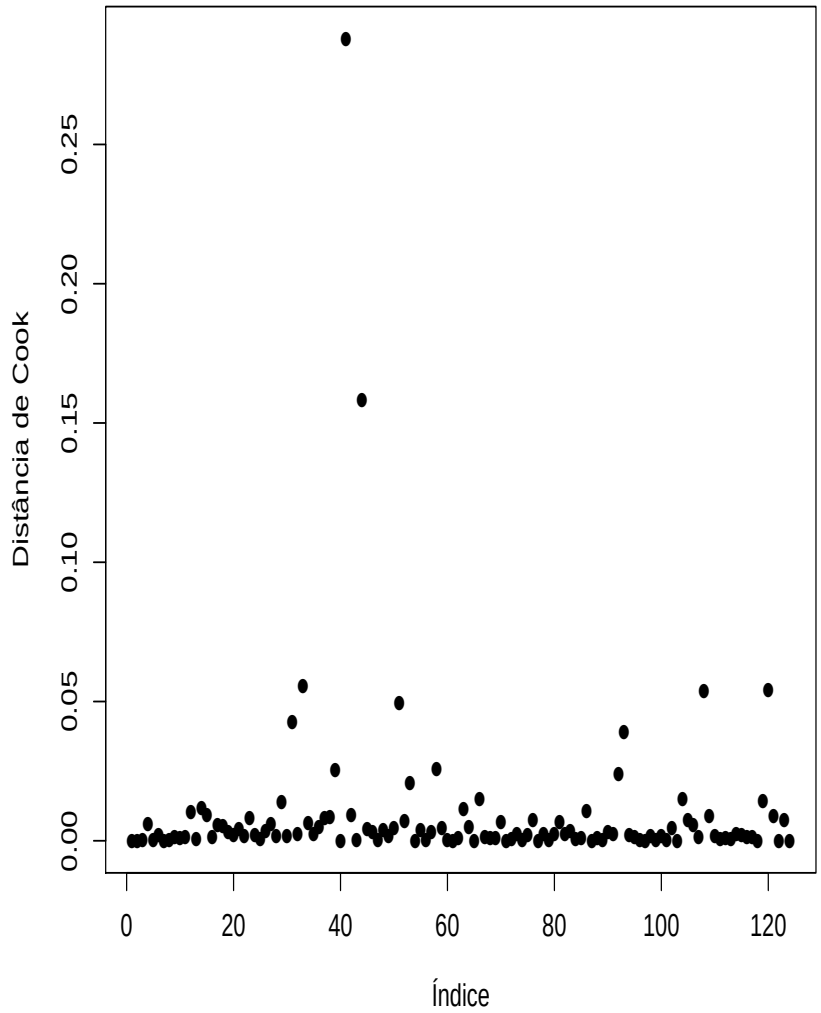
# Exemplo 1: modelo reduzido (desconsiderando as etiologias cardíacas)

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

- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $(\beta_0, \beta_1, \sigma^2)'$ : 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: $\xi_i$.
- 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 $i$.

# Pontos alavanca e distância de Cook



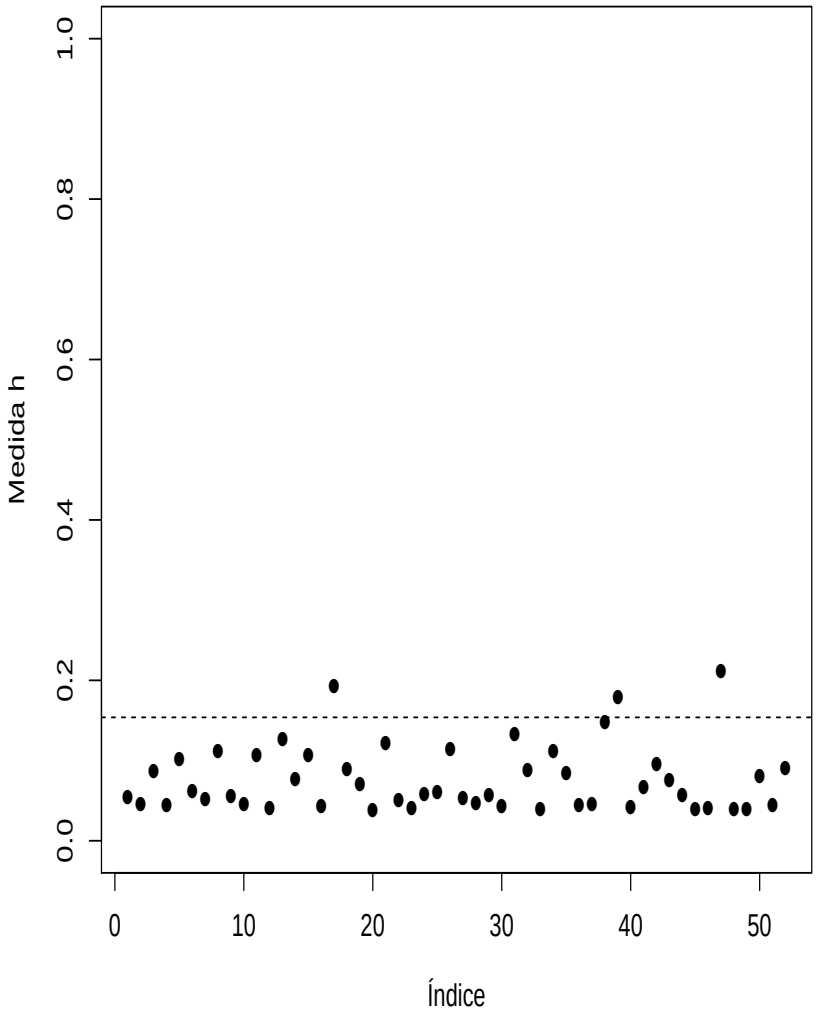
# Exemplo 2: desconsiderando o sexo

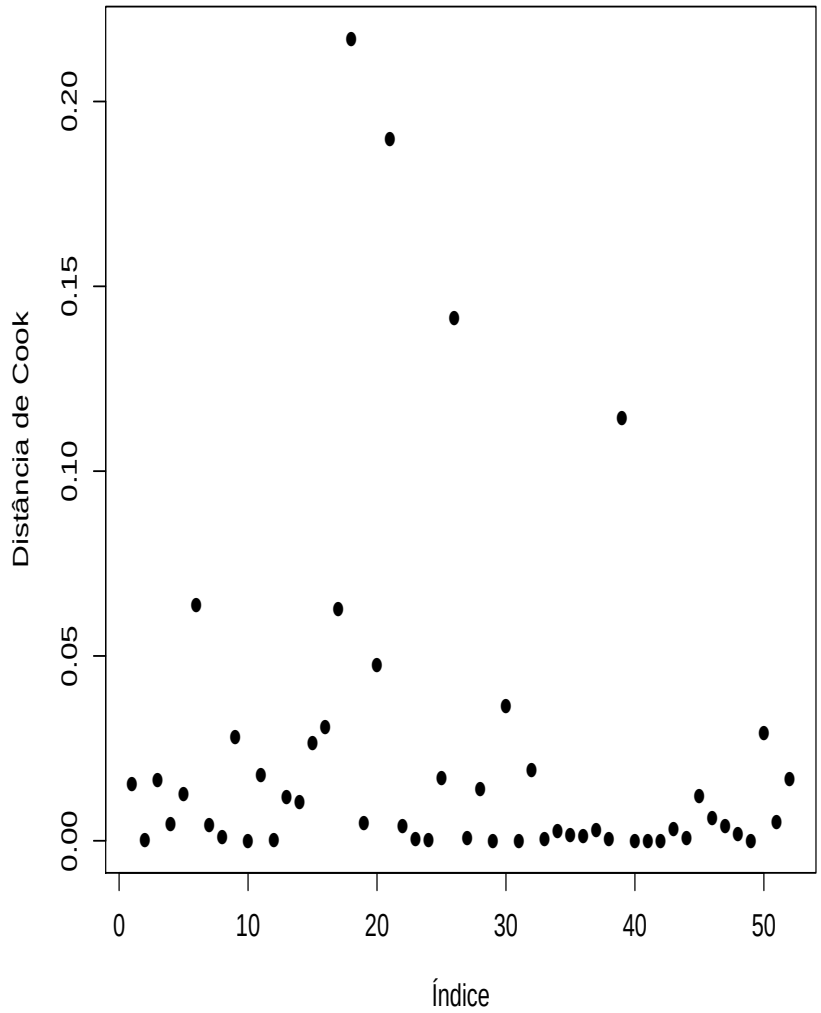
$$Y_{ij} = \beta_{0i} + \beta_{1i} x_{ij} + \xi_{ij}, i = 1(\text{convencional}), 2(\text{hugger})(\text{tipo de escova}),$$

$$j = 1, ..., 26(\text{criança}).$$

- $\xi_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $x_{ij}$: IPB pré-escovação da criança $j$ utilizando a escova do tipo $i$.
- $Y_{ij}$: IPB pós-escovação da criança $j$ utilizando a escova do tipo $i$.
- $\beta_{0i}$ : IPB pós-escovação esperado quando se utiliza a escova do tipo $i$ para um IPB pré-escovação igual a 0.
- $\beta_{1i}$ : incremento (positivo ou negativo) no IPB pós-escovação esperado quando se utiliza a escova do tipo $i$, para o aumento em uma unidade no IPB pré-escovação.

# Pontos alavanca e distância de Cook



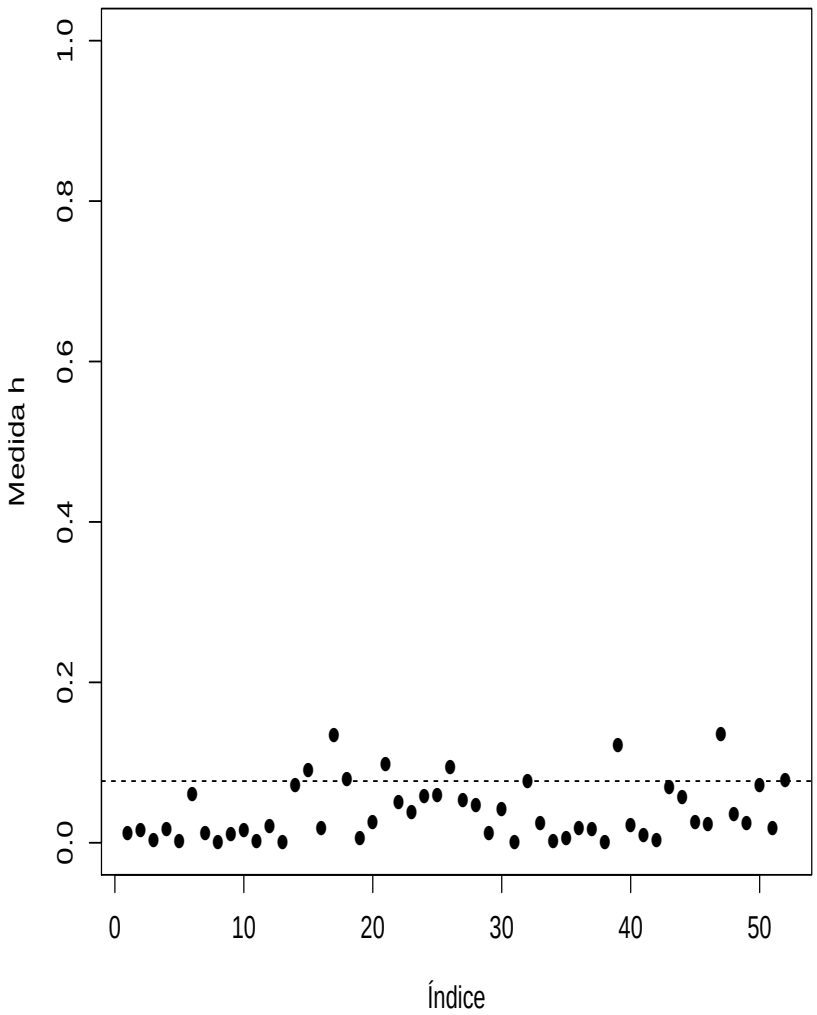
# Exemplo 2 (modelo reduzido): desconsiderando o sexo

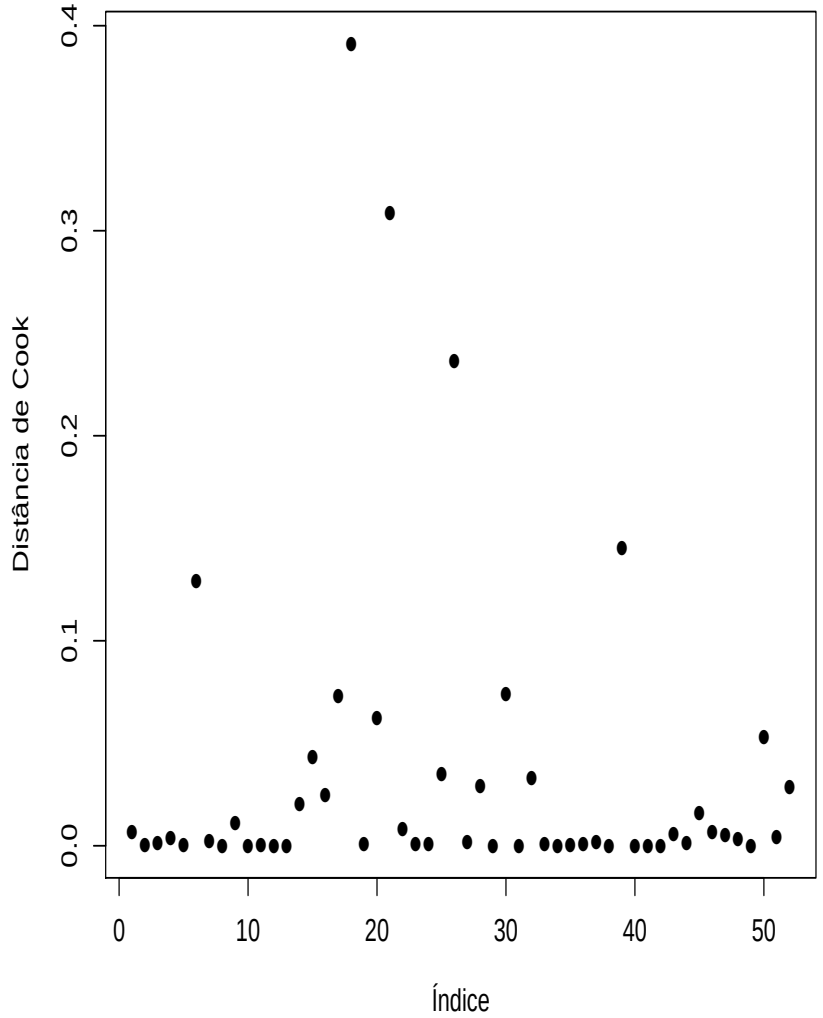
$$Y_{ij} = \beta_{1i} x_{ij} + \xi_{ij}, i = 1(\text{convencional}), 2(\text{hugger})(\text{tipo de escova});$$

$$j = 1, ..., 26(\text{criança}).$$

- $\xi_{ij} \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $x_{ij}$: IPB pré-escovação da criança $j$ utilizando a escova do tipo $i$.
- $Y_{ij}$: IPB pós-escovação da criança $j$ utilizando a escova do tipo $i$.
- $\beta_{1i}$ : diminuição (se $\beta_{i1} \in (0, 1)$) ou aumento (se $\beta_{i1} > 1$), no IPB quando se usa a escova do tipo $i$.

# Pontos alavanca e distância de Cook



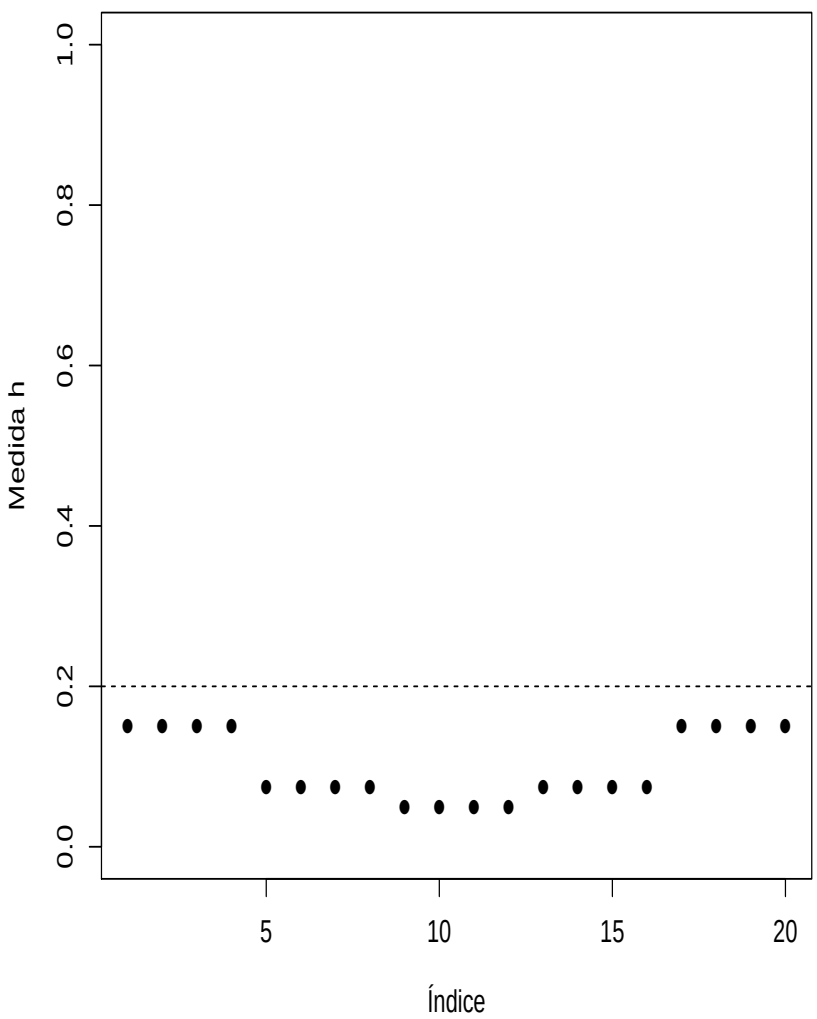
# Exemplo 3 (efeito do fósforo na produção de milho)

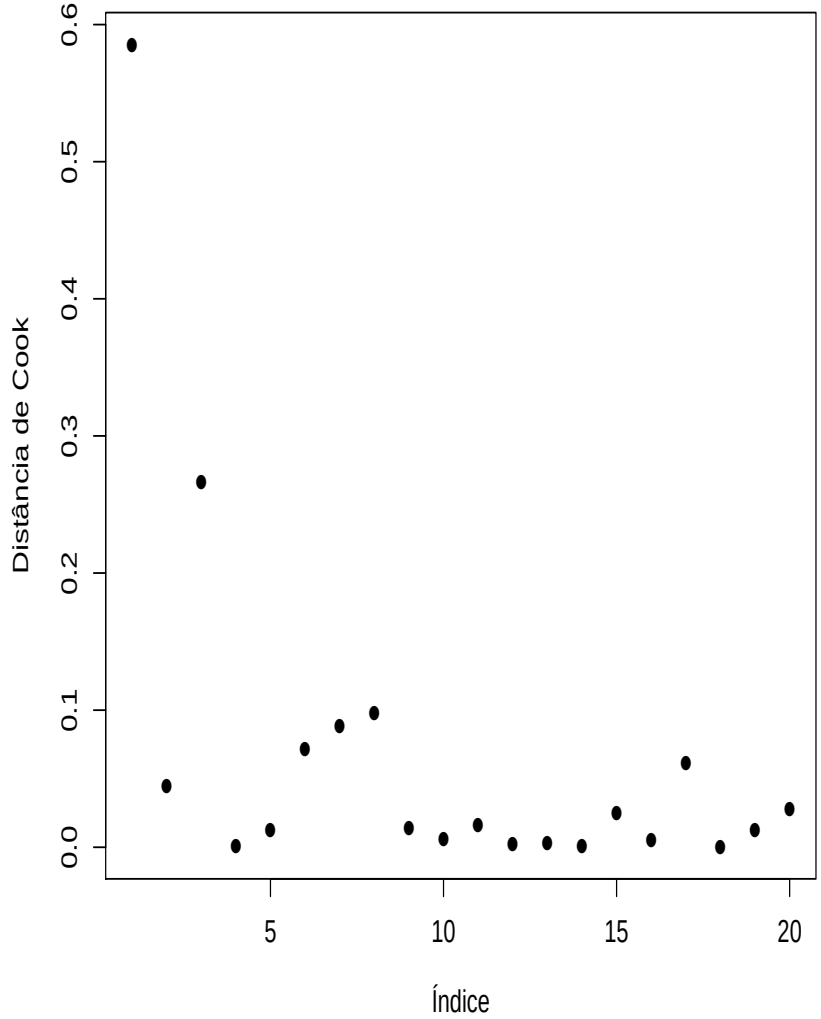
## Modelo linear 1: reta

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

- $x_i$ : quantidade de fósforo ministrada a i-ésima parcela.
- $\beta_0$ : valor esperado (média) da produção de milho quando a quantidade de fósforo aplicada é igual à 0.
- $\beta_1$ : incremento no valor esperado da produção de milho quando a quantidade de fósforo aplicada aumenta em uma unidade.
- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

# Pontos alavanca e distância de Cook



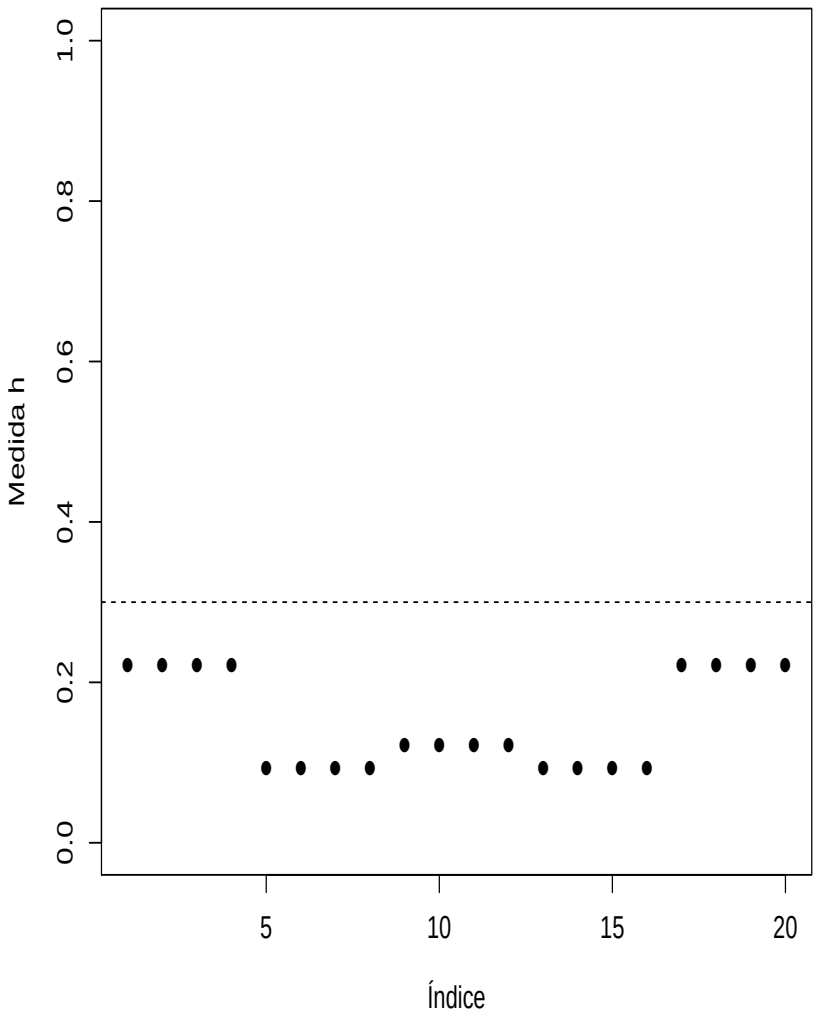



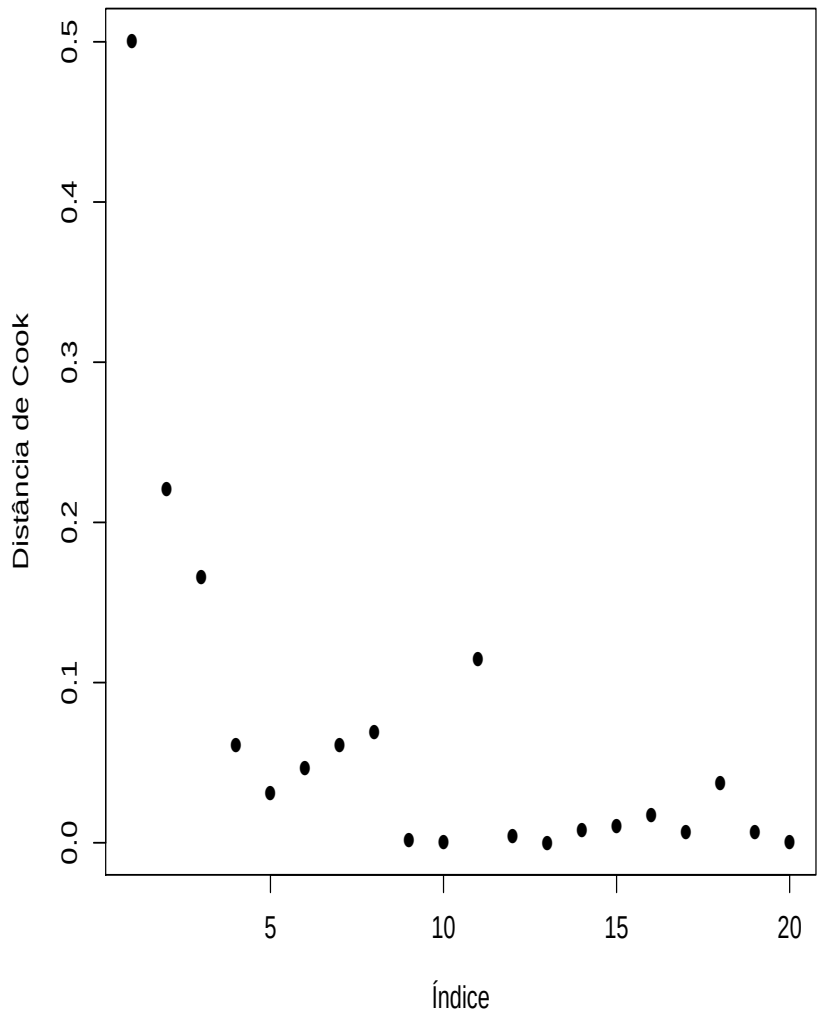
# Modelo linear 2: parábola

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

- $x_i$ : quantidade de fósforo ministrada a i-ésima parcela.
- $\beta_0$ : valor esperado (média) da produção de milho quando a quantidade de fósforo aplicada é igual à 0.
- A interpretação isolada dos parâmetros $\beta_1$ e $\beta_2$ é complicada mas, podemos dizer que $\frac{-\beta_1}{2\beta_2}$ é o máximo (ou mínimo) do valor esperado da produção de milho.
- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.

# Pontos alavanca e distância de Cook



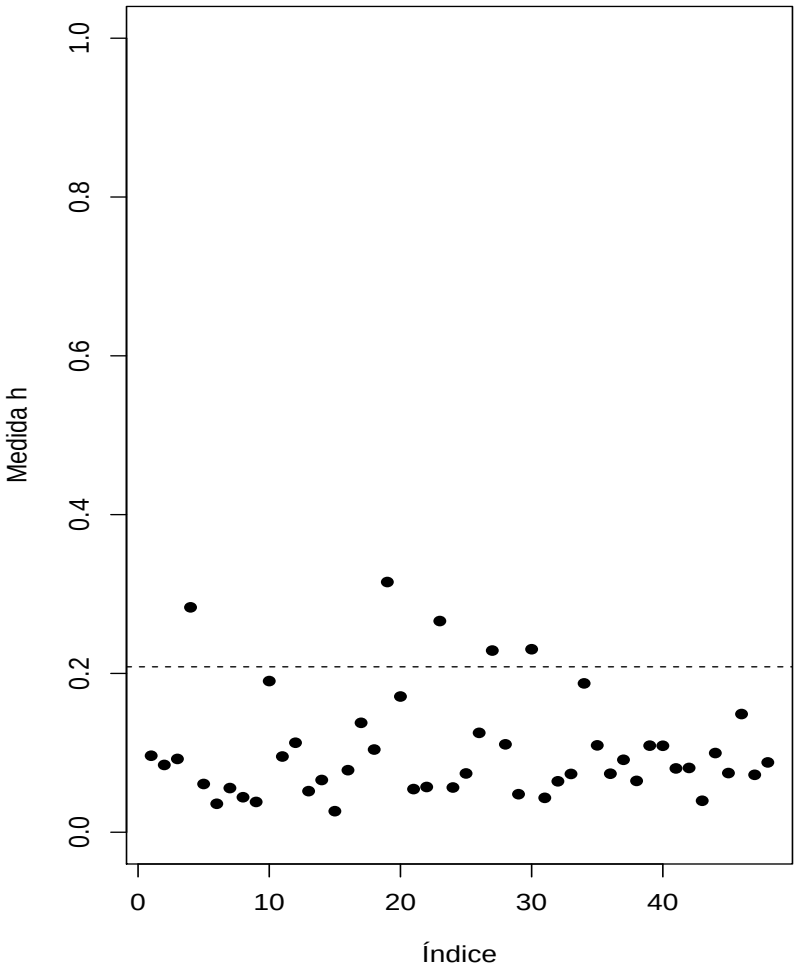
# Exemplo 4: consumo de combustível

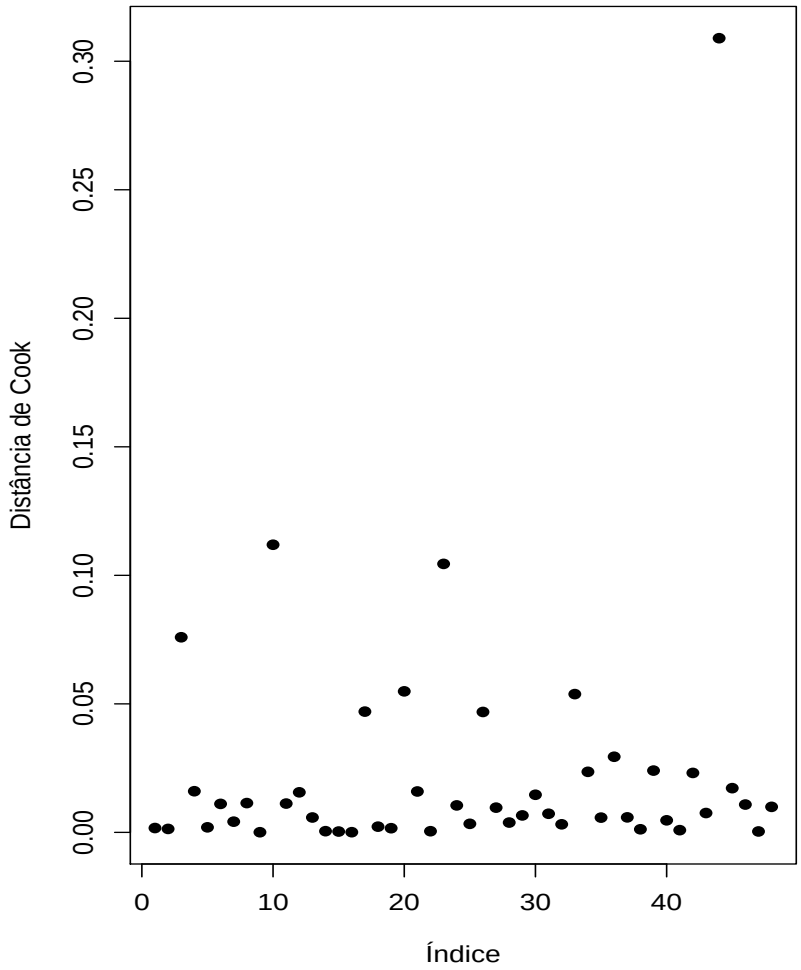
$$Y_i = \beta_0 + \beta_1 \frac{x_{1i} - \overline{x}_1}{s_1} + \beta_2 \frac{x_{2i} - \overline{x}_2}{s_1} + \beta_3 \frac{x_{3i} - \overline{x}_3}{s_3} + \beta_4 \frac{x_{4i} - \overline{x}_4}{s_4} + \xi_i, \quad (2)$$

$i = 1, ..., 48, \overline{x}_j = \frac{1}{48} \sum_{i=1}^{48} x_{ij}, j = 1, 2, 3, 4$

- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $\beta_0$ : consumo esperado para estados com valor de cada covariável igual à sua respectiva média.
- $\beta_j / s_j$ : incremento (positivo ou negativo) no consumo esperado para o aumento em uma unidade da variável $j, j = 1, 2, 3, 4$, mantendo-se as outras fixas.

# Pontos alavanca e distância de Cook



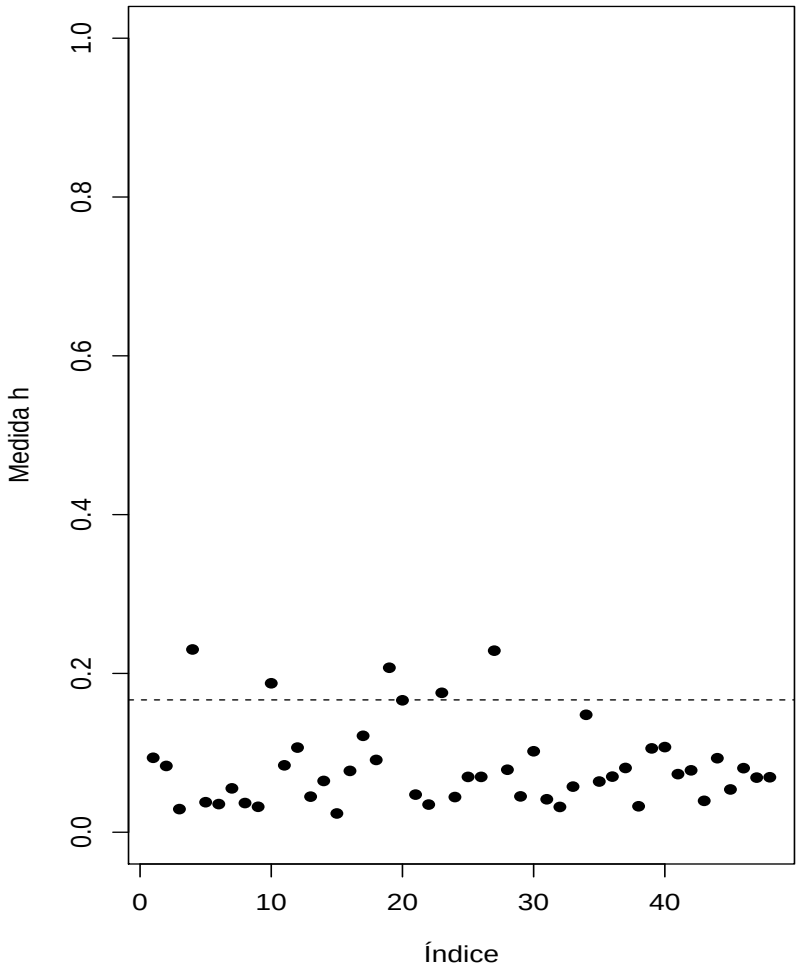
# Exemplo 4: consumo de combustível (modelo reduzido)

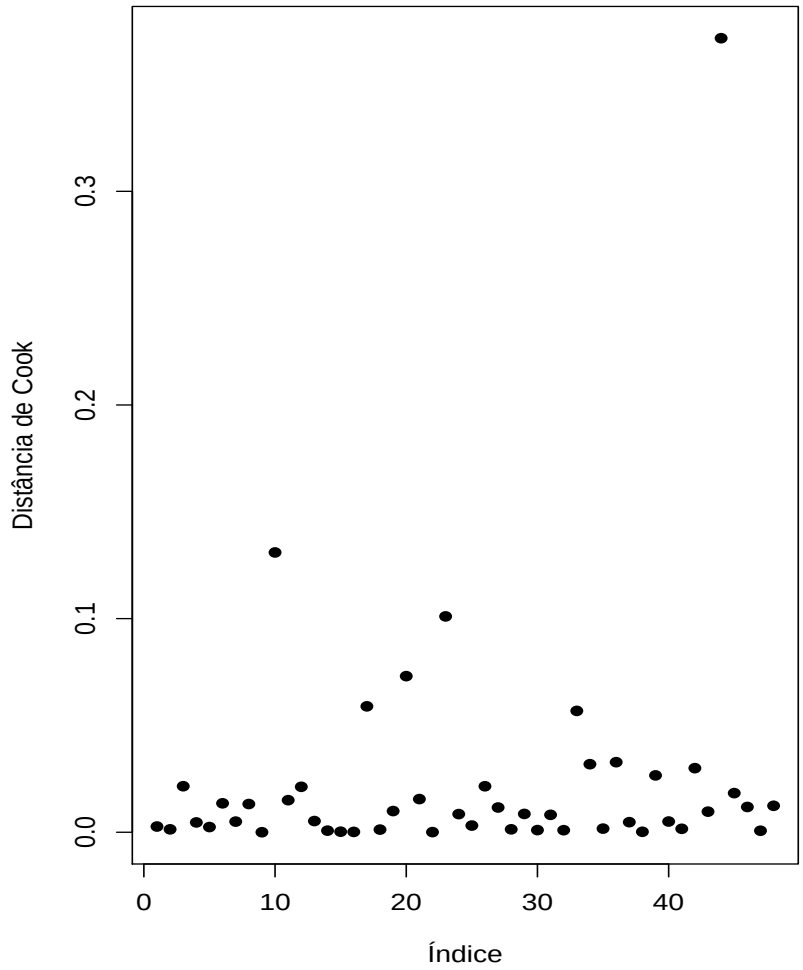
$$Y_i = \beta_0 + \beta_1 \frac{x_{1i} - \overline{x}_1}{s_1} + \beta_2 \frac{x_{2i} - \overline{x}_2}{s_2} + \beta_3 \frac{x_{3i} - \overline{x}_3}{s_3} + \xi_i, \qquad (3)$$

$i = 1, ..., 48, \overline{x}_j = \frac{1}{48} \sum_{i=1}^{48} x_{ij}, j = 1, 2, 3, 4$

- $\xi_i \overset{i.i.d.}{\sim} N(0, \sigma^2)$.
- $\beta_0$ : consumo esperado para estados com valor de cada covariável igual à sua respectiva média.
- $\beta_j / s_j$ : incremento (positivo ou negativo) no consumo esperado para o aumento em uma unidade da variável $j, j = 1, 2, 3, 4$, mantendo-se as outras fixas.

# Pontos alavanca e distância de Cook



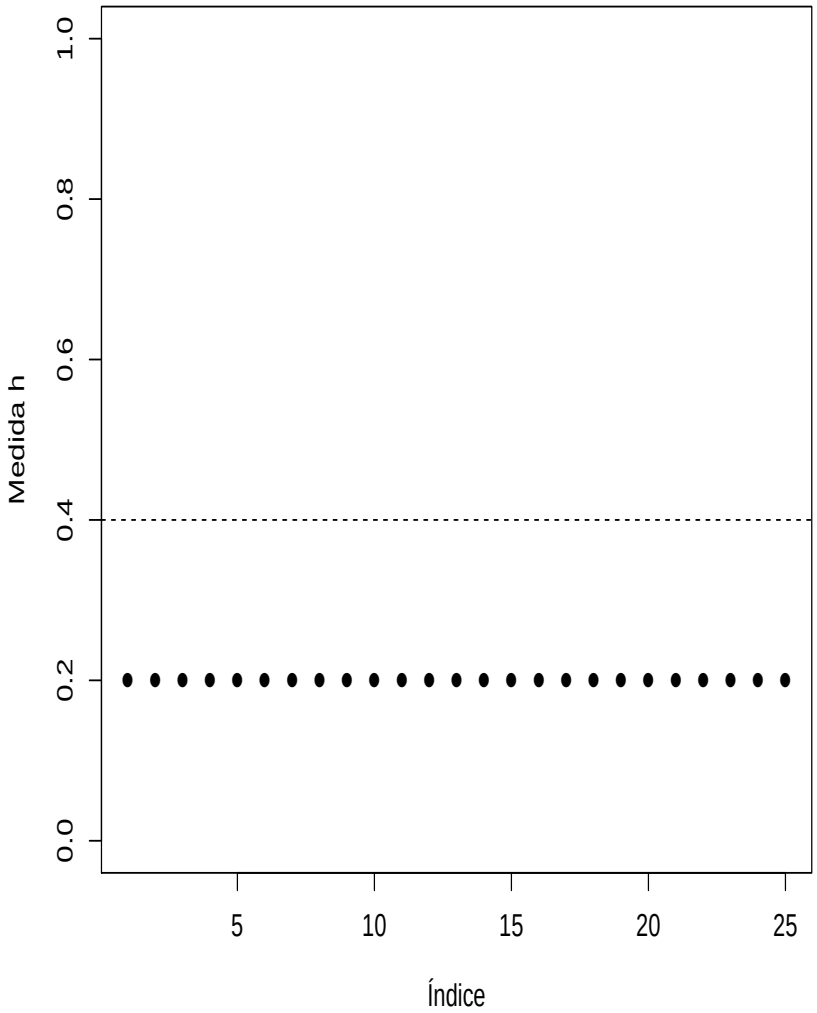
# Exemplo 5: Modelo (casela de referência)

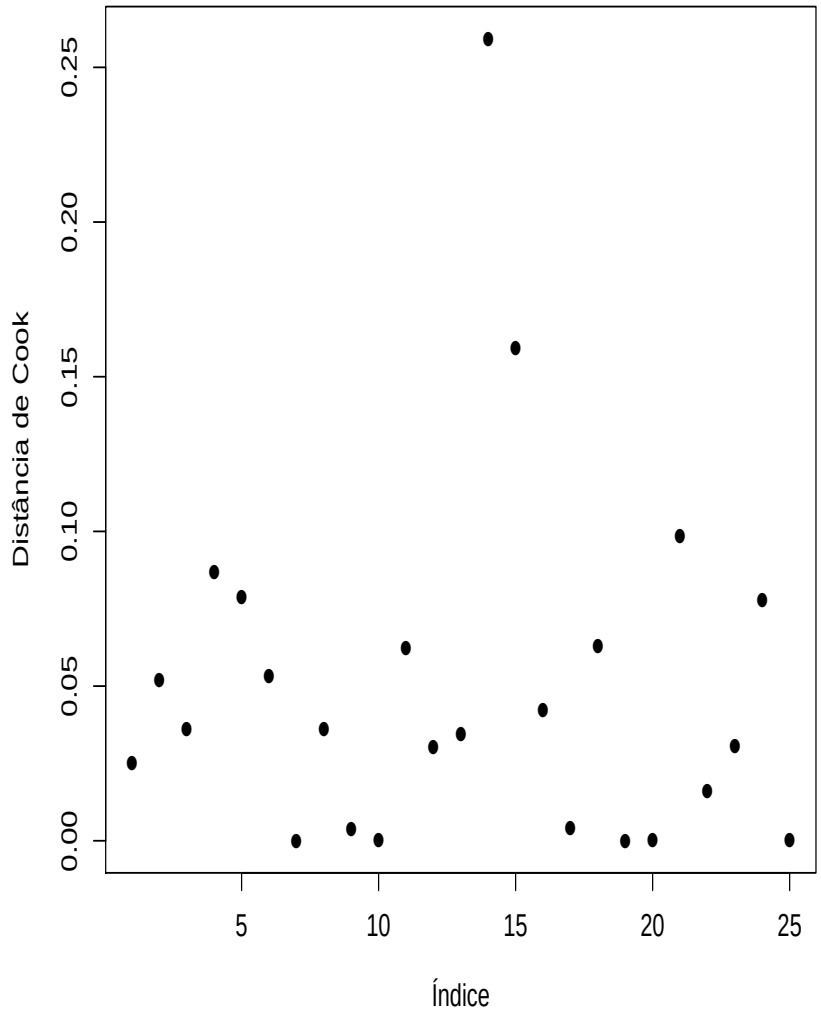
$$Y_{ij} \quad = \quad \mu + \alpha_i + \xi_{ij},$$

$i = 1, 2, .., 5$ (grupos); $j = 1, ..., 5$(unidades experimentais)

- $\xi_{ij} \overset{ind.}{\sim} N(0, \sigma^2)$. Parte sistemática: $\mu_i = \mu + \alpha_i$, é a média populacional relacionada ao i-ésimo fator, $\alpha_1 = 0$ (restrição de identificabilidade) .
- $\mu$ : é a média populacional do grupo de referência, $\mu_1 = \mu$.
- $\alpha_i = \mu_i - \mu_1, i = 2, ..., 5$, é o incremento (positivo ou negativo) entre a média do grupo $i$ e a média do grupo de referência.
- Grupos : grupo 1(E50), grupo 2(E70), grupo 3(EAW), grupo 4(M1M), grupo 5(MAW).

# Pontos alavanca e distância de Cook



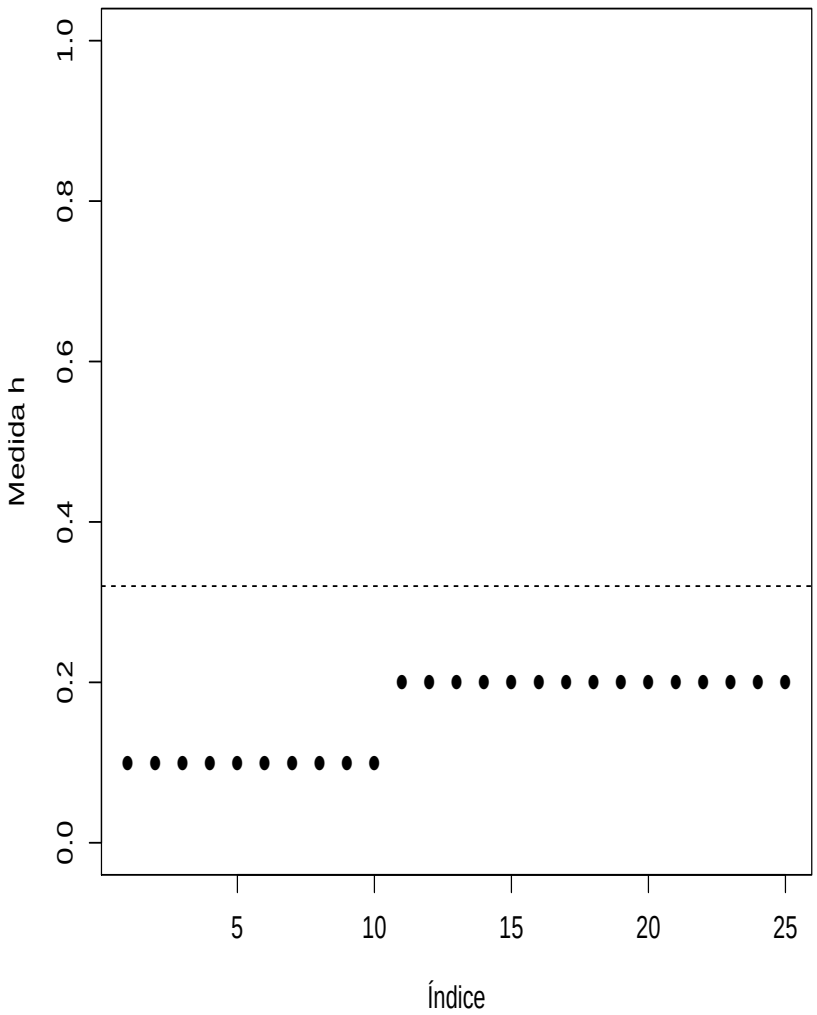
# Modelo reduzido (casela de referência)

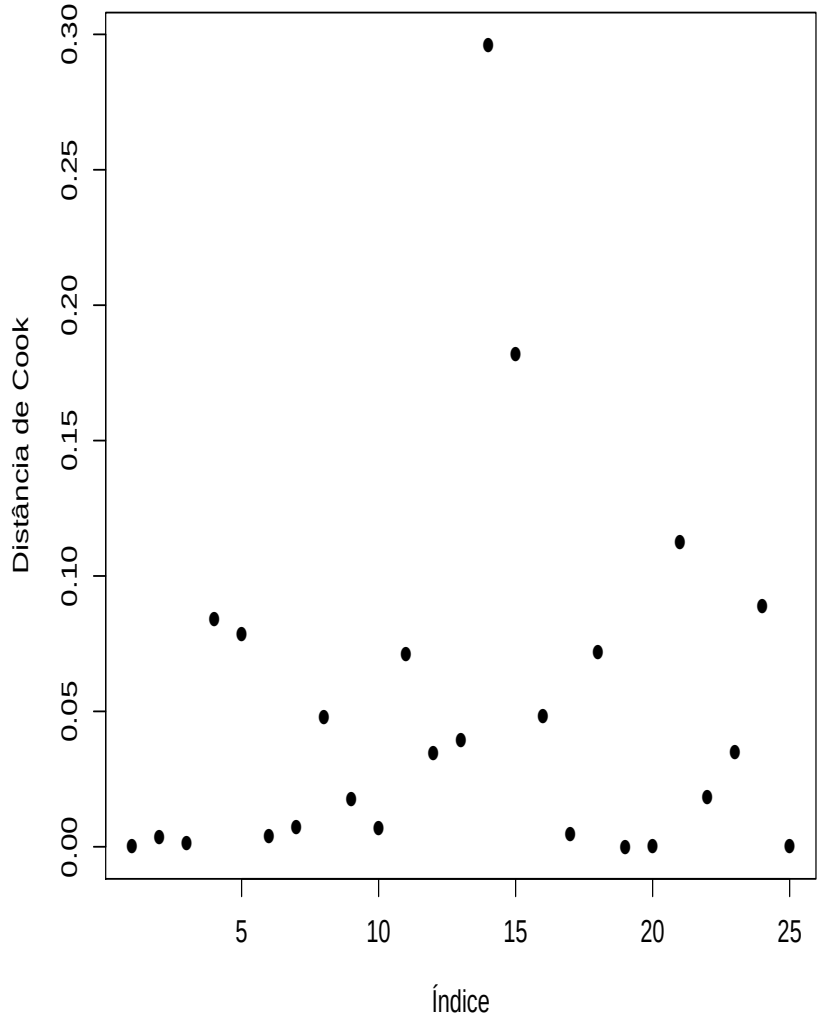
$$Y_{ij} = \mu + \alpha_i + \xi_{ij},$$

$i = 1, 2, 3, 4$(grupos); $j = 1, ..., 4$(unidades experimentais)

- Parte sistemática: $\mu_i = \mu + \alpha_i$, é a média populacional relacionada ao i-ésimo fator, $\alpha_1 = 0$ (restrição de identificabilidade).
- $\mu$ : é a média populacional do grupo de referência, $\mu_1 = \mu$.
- $\alpha_i = \mu_i - \mu_1, i = 2, 3, 4$, é o incremento (positivo ou negativo) entre a média do grupo $i$ e a média do grupo de referência.
- Grupos : grupo 1(E50/EAW), grupo 2(E70), grupo 3(M1M), grupo 4(MAW).

# Pontos alavanca e distância de Cook



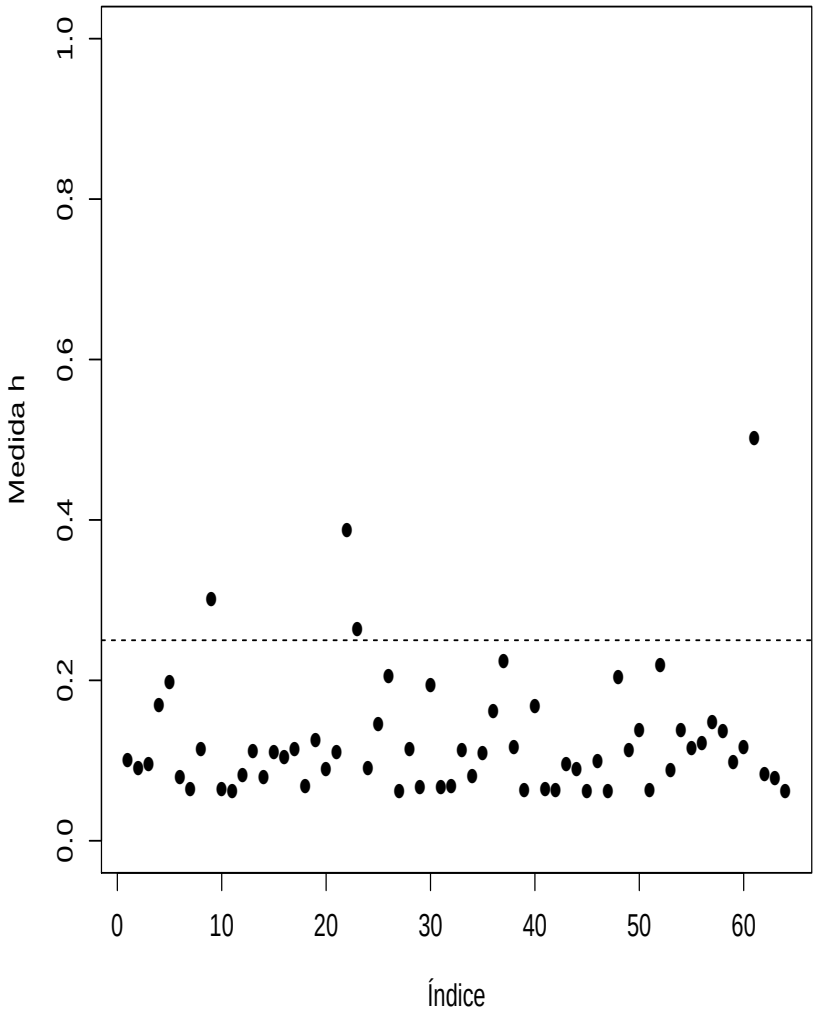
# Exemplo 11: Estudo (longitudinal) da eficácia de escovas de dentes

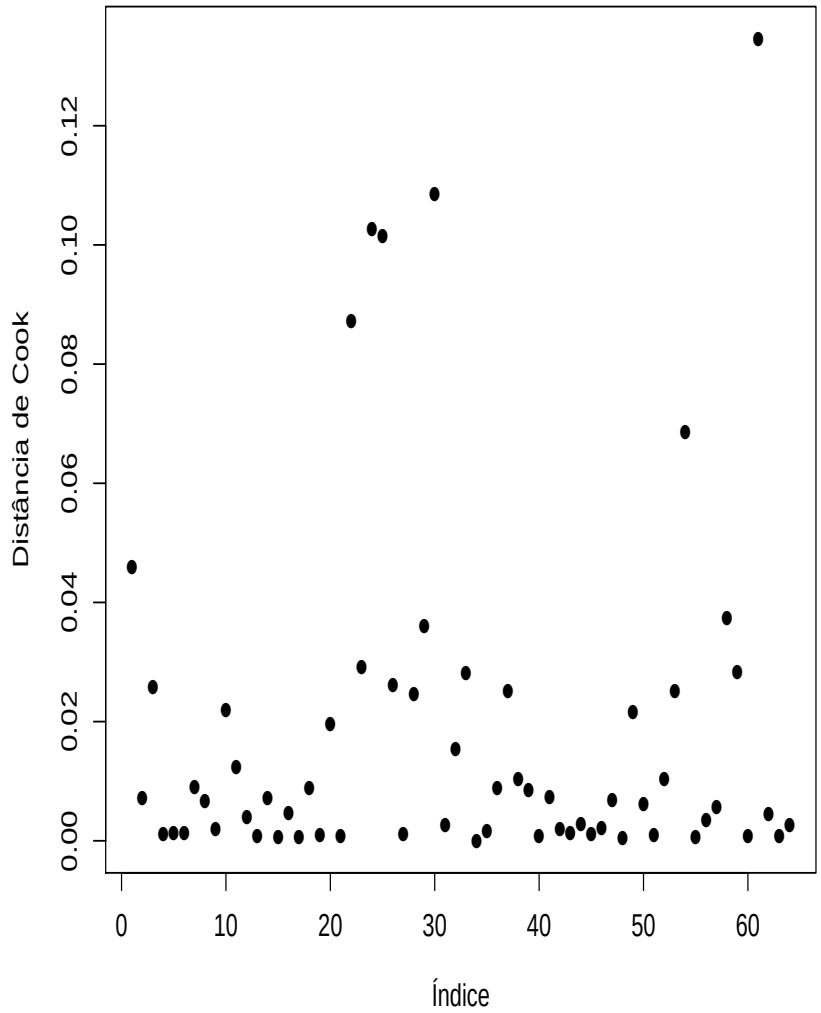
$$Y_{ijk} = \alpha_{ij} + \beta_{ij} x_{ijk} + \xi_{ijk},$$

$i = 1$(convencional (CT)), $2$(monobloco (MT)) (tipo de escova), $j = 1, 2$ (sessão), $k = 1, 2, ..., 16$ (indivíduo)

- $x_{ijk}$ : é o IPB da criança $k$, submetida ao tipo de escova $i$, na sessão $j$, antes da escovação.
- $Y_{ijk}$ : é o IPB da criança $k$, submetida ao tipo de escova $i$, na sessão $j$, depois da escovação.
- $\alpha_{ij}$ : é o IPB pós escovação esperado quando se utiliza a escova do tipo $i$ na j-ésima sessão. $\beta_{ij}$ : é o incremento (positivo ou negativo) no IPB esperado pós escovação para crianças submetidas ao tipo de escova $i$ na

# Pontos alavanca e distância de Cook



# Exercício

Conduzir uma análise de influência para os outros exemplos vistos em sala, bem como nas listas de exercícios.