ME720 - Modelos Lineares Generalizados

Parte 6 - Métodos de diagnóstico em modelos normais lineares

Profa. **Larissa Avila Matos**

Forma matricial para do MNL

Considere o modelo,

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

com

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

- Suposição: $\boldsymbol{\xi} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$ (vetor $n \times 1$ de erros).
- $\boldsymbol{Y}_{n\times 1}$: vetor das variáveis resposta.
- $\boldsymbol{X}_{n\times p}$: matriz de plajenamento (ou delineamento) que define a parte sistemática do modelo.

Suposições

As principais suposições do MNL são:

- Homocedasticidade.
- Independência dos erros.
- Normalidade dos erros.

Como verificar as suposições do modelo?

Como proceder se uma ou mais de uma suposição não for (satisfatoriamente) válida?

Resíduos

Como os erros ($\boldsymbol{\xi}$) não são observados, precisamos de algum preditor apropriado para avaliar as suposições feitas sobre eles.

Já definimos os resíduos: $\widehat{\xi}_i = R_i = Y_i - \widehat{Y}_i = Y_i - \boldsymbol{x}_i'\widehat{\boldsymbol{\beta}}$. Definimos também $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ (matriz de projeção), então matricialmente,

$$\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\xi}.$$

Em termos dos elementos da matriz $\boldsymbol{H}$, temos que

$$R_i = \xi_i - \sum_{j=1}^{n} h_{ij}\xi_j, \quad i = 1, \cdots, n.$$

$\Rightarrow$ Se os h_{ij}'s são pequenos (em valor absoluto), $\boldsymbol{R}$ é próximo de $\boldsymbol{\xi}$.

Propriedades dos Resíduos

- $\mathbb{E}(\boldsymbol{R}) = \mathbb{E}((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\xi}) = \mathbf{0}$.
- $Cov(\boldsymbol{R}) = Cov((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\xi}) = (\boldsymbol{I} - \boldsymbol{H})\sigma^2 = (\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}')\sigma^2$.

  Então, sob as suposições do modelo,

  $$\boldsymbol{R} \sim N_n(\mathbf{0}, \sigma^2(\boldsymbol{I} - \boldsymbol{H})),$$

  são correlacionados.

- $Cov(\boldsymbol{R}, \boldsymbol{Y}) = Cov((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \boldsymbol{Y}) = (\boldsymbol{I} - \boldsymbol{H})\sigma^2$.
- $Cov(\boldsymbol{R}, \hat{\boldsymbol{Y}}) = Cov((\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}, \boldsymbol{H}\boldsymbol{Y}) = \mathbf{0}$.

- $\sum_{i=1}^{n} R_i = 0 \Rightarrow \bar{R} = 0.$
- $\boldsymbol{R'Y} = SQR = \boldsymbol{Y'}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{Y}.$
- $\boldsymbol{R'}\hat{\boldsymbol{Y}} = \boldsymbol{Y'}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{H}\boldsymbol{Y} = 0.$
- $\boldsymbol{R'X} = \boldsymbol{Y'}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X} = \boldsymbol{0}.$

Potanto, através das propeiedades dos resíduos podemos verificar se as suposições estão corretas.

Propriedades da Matriz $\boldsymbol{H}$

Se $\boldsymbol{X}$ é uma matriz $n \times p$, de posto $p < n$, e se a primeira coluna de $\boldsymbol{X}$ é $\mathbf{1}_n$. Então, os elementos da matriz $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ tem as seguintes propriedades:

1. $\frac{1}{n} \leq h_{ii} \leq 1$, para $i = \cdots, n$.
2. $-\frac{1}{2} \leq h_{ij} \leq \frac{1}{2}$, $\forall\ j \neq i$.
3. $tr(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = p$.

Prova: ??

Resíduos Padronizados / Studentizados

Como vimos, as variâncias dos resíduos não são constantes, então desejamos padronizar os resíduos para que eles tenham mesma variância. Sabemos que $Var(R_i) = (1 - h_{ii})\sigma^2$.

Resíduos Padronizados:

$$Z_i = \frac{R_i}{\sqrt{\sigma^2(1 - h_{ii})}},$$

onde Z_i tem média 0 e variância 1, e h_{ii} é o i-ésimo elemento da diagonal principal de $\boldsymbol{H}$.

Podemos substituir σ^2 por $\hat{\sigma}^2 = S^2 = \frac{SQR}{n-p}$.

Resíduos Studentizados: Defina,

$$T_i^* = \frac{R_i}{\sqrt{S^2(1-h_{ii})}},$$

em que

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

A divisão por $(1-h_{ii})$ atenua a correlação entre os resíduos.

Contudo, R_i e S^2 não são independentes.

Podemos padronizar os resíduos considerando um estimador para σ^2 que exclua a i-ésima observação.

Então, temos que $S^2_{(i)}$ e R_i são independentes (em que $S^2_{(i)}$ é S^2 correspondente ao modelo sem a i-ésima observação).

Pode-se provar, além disso, que

$$S^2_{(i)} = S^2 \left(\frac{n - p - T_i^2}{n - p - 1} \right),$$

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

Tem-se, então, que

$$T_i = \frac{R_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}} \sim t_{(n-p-1)},$$

sob a validade das hipóteses do modelo.

O que e como observar os resíduos?

Gráfico de dispersão dos resíduos versus seu índice: Ausência de tendência (autocorrelações, por exemplo).

Gráfico de dispersão dos resíduos versus valores ajustados: Variâncias parecidas para diferentes "grupos" de resíduos.

Boxplot e/ou gráfico de quantis-quantis: Simetria, ausência de "outliers" e multimodalidade.

Problema no gráfico de quantis-quantis: Visualmente, muitas vezes, é complicado avaliar a proximidade dos quantis.

Solução: criar bandas de confiança.

Gráfico de quantis-quantis (Q-Q plot)

Usado para checar a suposição de normalidade dos erros (resíduos).

Graficamos as estatísticas de odem dos resíduos (T_i) versus os quantis da $N(0,1)$ (s_i).

Temos que, $T_{(1)} \leq T_{(2)} \leq \cdots \leq T_{(n)}$ (estatísticas de ordem, quantis amostrais).

Também que, $s_1 \leq s_2 \leq \cdots \leq s_n$ os scores da normal, onde são definidos como

$$\mathbb{P}(N(0,1) \leq s_i) = \frac{i - 1/2}{n} \quad \text{(quantis teóricos)}.$$

Gráfico de envelopes

1. Ajusta-se o modelo e obtem-se

$$t_i = \frac{r_i}{\sqrt{s^2_{(i)}(1 - h_{ii})}}.$$

2. Gera-se n observações N(0,1) as quais são armazenadas em $\boldsymbol{y} = (y_1, ..., y_n)$.

3. Ajusta-se o modelo considerando-se $\boldsymbol{y}$ e obtem-se $r_i = y_i - \widehat{y}_i$ (relativo ao modelo ajustado nesta etapa).

4. Obtem-se agora,

$$t_i = \frac{r_i}{\sqrt{s^2_{(i)}(1 - h_{ii})}}, i = 1, ..., n.$$

Gráfico de envelopes

5. Repete-se os passos (2)-(4), m vezes. Logo, teremos $t_{ij}, i = 1, ..., n$ e $j = 1, ..., m$.

6. Colocamos cada grupo de n resíduos em ordem crescente, obtendo-se $t_{(i)j}$.

7. Obtemos os limites

$$t_{(i)I} = min_j t_{(i)j} \quad \text{e} \quad t_{(i)S} = max_j t_{(i)j}.$$

Assim, os limites correspondentes ao i-ésimo resíduo serão dados por $t_{(i)I}$ e $t_{(i)S}$.

Outliers

Para identificar possíveis "outliers" podemos fazer o boxplot e/ou gráfico de quantis-quantis. Além disso, podemos fazer o gráfico de dispersão dos resíduos versus valores ajustados.

Mas, podemos também examinar os *"resíduos deletados"*. O i-ésimo resíduo, deletando a i-ésima observação, pode ser computado usando

$$R_{(i)} = \boldsymbol{Y}_i - \hat{\boldsymbol{Y}}_{(i)} = \boldsymbol{Y}_i - \boldsymbol{x}_i' \widehat{\boldsymbol{\beta}}_{(i)} = \frac{R_i}{1 - h_{ii}}.$$

Graficamos R_i versus $R_{(i)}$, se o ajuste não muda substancialmente no cálculo de $\widehat{\boldsymbol{\beta}}$ quando a i-ésima observação é deletada, os pontos no gráfico devem estar em uma reta com inclinação 1.

Observações influentes e alavanca

Alavanca:

Para investigar a influência de cada observação, vamos começar analisando

$$\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y} \Rightarrow \hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j = h_{ii} Y_i + \sum_{j \neq i} h_{ij} Y_j.$$

Sabemos das propriedades da matriz $\boldsymbol{H}$ que

$$1 = h_{ii} + \frac{\sum_{j \neq i} h_{ij}}{h_{ii}}.$$

Então, temos que, se h_{ii} é grande (perto de 1), os h_{ij}'s $i \neq j$ são pequenos e Y_i contribui muito mais que os outros Y's para $\hat{Y}_i$.

h_{ii} é chamado de ponto de alavanca da observação i e pode ser interpretado como a alavanca exercida pela i-ésima observação (Y_i) na resposta predita $\hat{Y}_i$.

Como $tr(\boldsymbol{H}) = \sum_{i=1}^{n} h_{ii} = p \Rightarrow \sum_{i=1}^{n} h_{ii}/n = p/n$ (média dos h_{ii}'s), é recomendado destacar as obserções onde $h_{ii} \geq 2p/n$.

Observação: Nem toda observação com h_{ii} alto é influente na análise de regressão. Observe que, a diagonal da matriz $\boldsymbol{H}$ avalia a localização da observação no espaço de variáveis explicativas. Como consequência, podemos ter observações com ponto de alavanca alto, mas não influenciarem nas estimativas dos parâmetros.

Distância de Cook:

Para formalizar a influência de um ponto, vamos considerar o efeito dessa observação em $\widehat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

Já vimos que $\widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}'_{(i)}\boldsymbol{Y}_{(i)}$, então podemos comparar $\widehat{\boldsymbol{\beta}}$ com $\widehat{\boldsymbol{\beta}}_{(i)}$ pela *distância de Cook*, definida como

$$D_i = \frac{(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})'\boldsymbol{X}'\boldsymbol{X}(\widehat{\boldsymbol{\beta}}_{(i)} - \widehat{\boldsymbol{\beta}})}{pS^2} = \frac{(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})'(\widehat{\boldsymbol{Y}}_{(i)} - \widehat{\boldsymbol{Y}})}{pS^2},$$

no qual D_i é proporcional a distância euclediana entre $\widehat{\boldsymbol{Y}}_{(i)}$ e $\widehat{\boldsymbol{Y}}$. Então, se D_i é grande a i-ésima observação tem influência em $\widehat{\boldsymbol{\beta}}$ e $\hat{\boldsymbol{Y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$.

A *distância de Cook* pode ser escrita como

$$D_i = \frac{T_i^{*2}}{p} \frac{h_{ii}}{(1 - h_{ii})}.$$

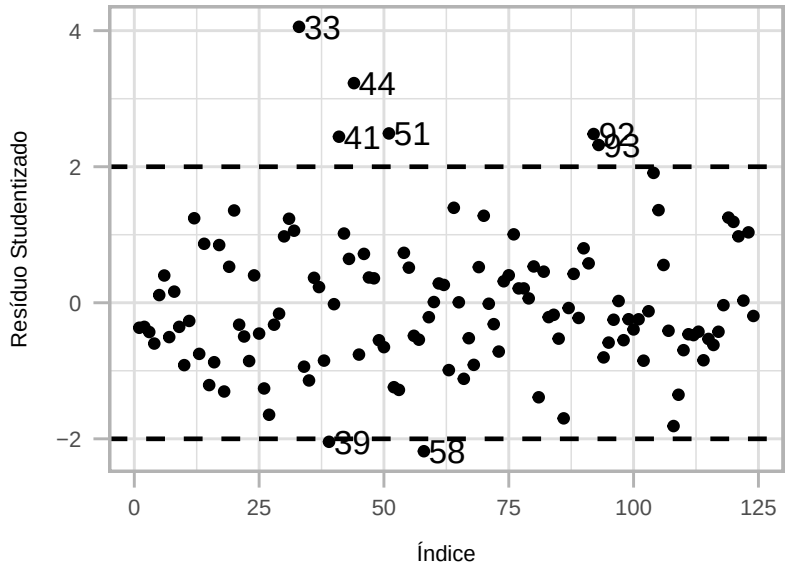
Destacamos as observações quando $D_i > 1$.

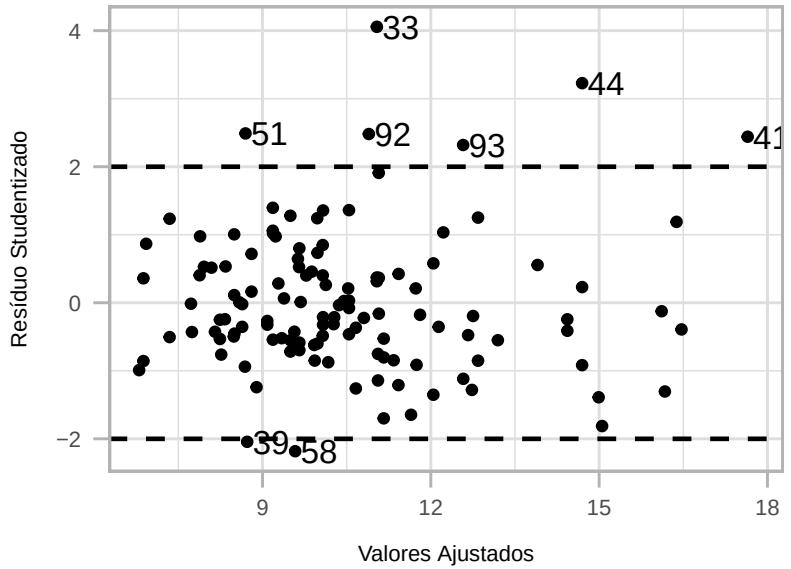
Exemplo 1

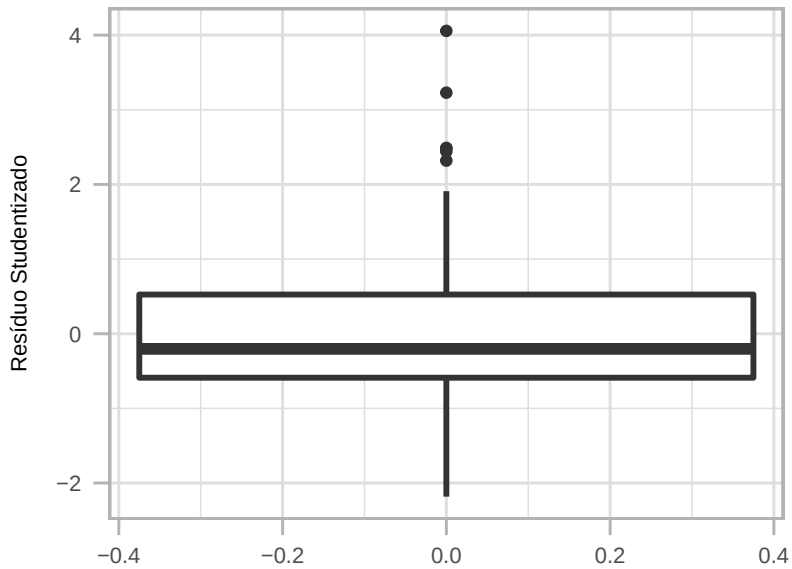
Voltemos ao exemplo 1.

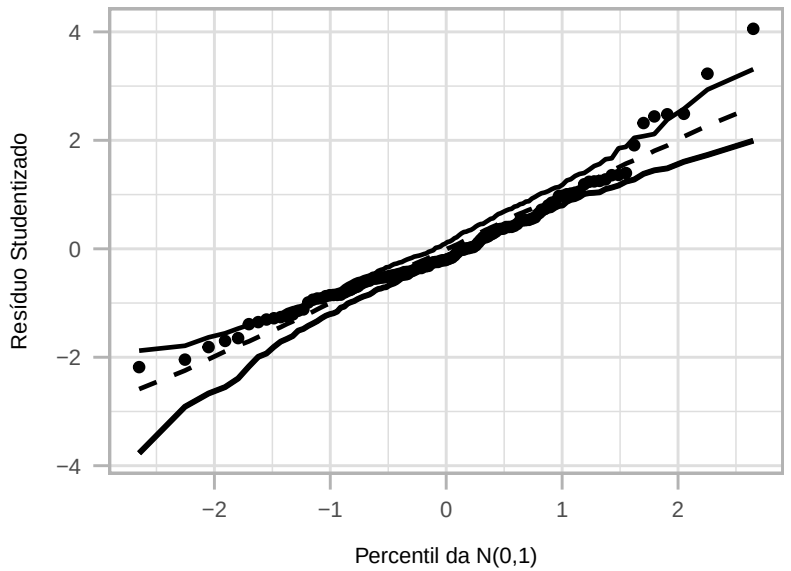
Considerando primeiro o modelo que contempla os grupos e depois o modelo reduzido.

Considerando as etiologias cardíacas

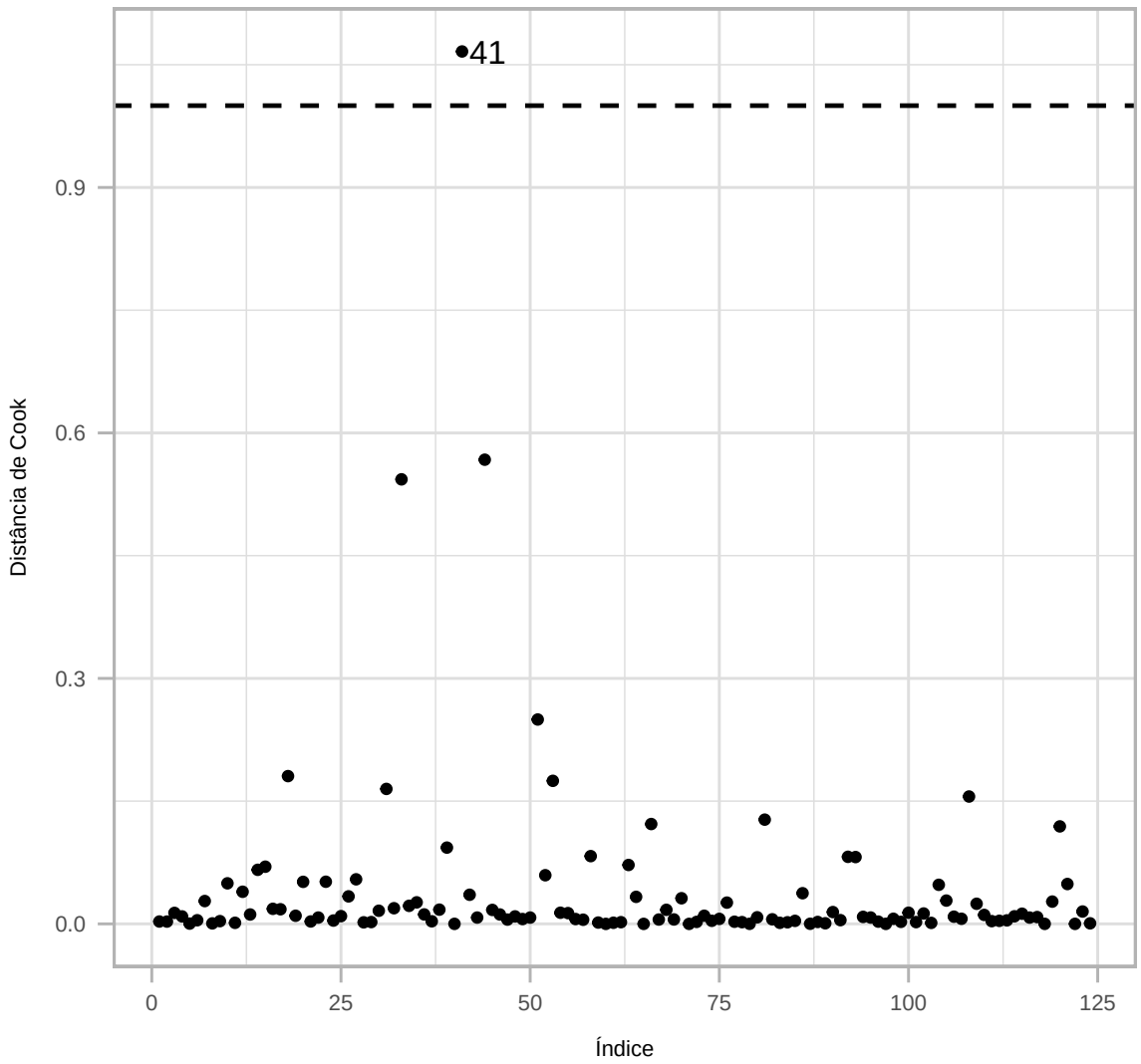


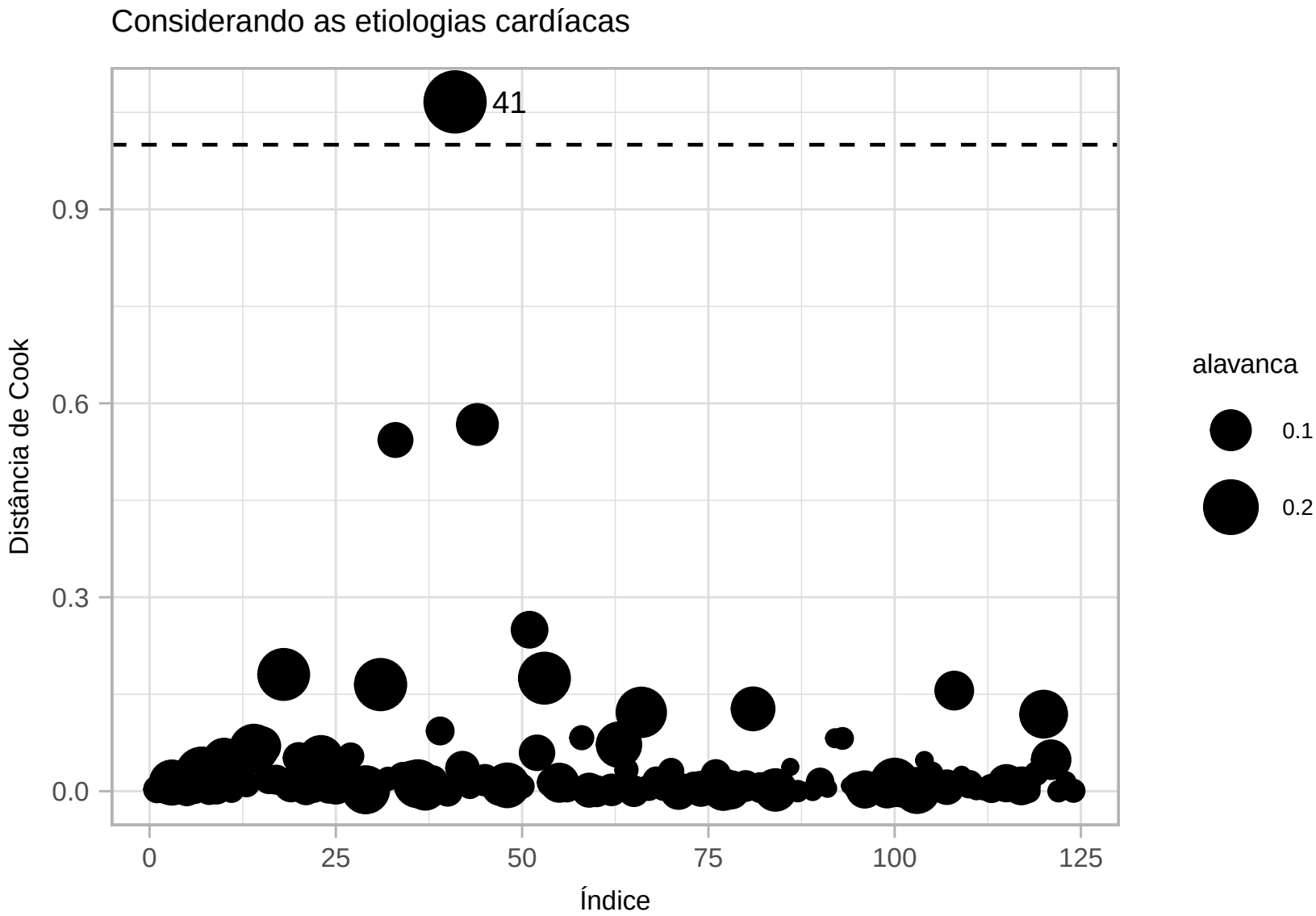




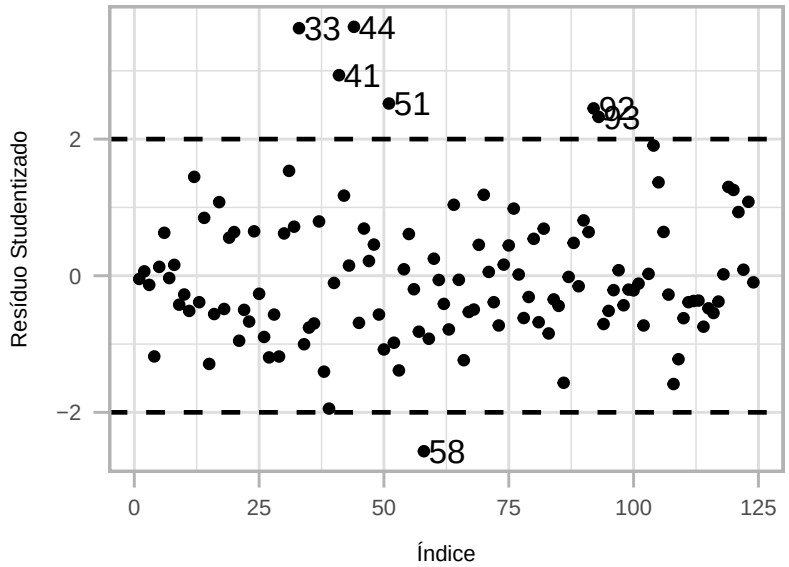


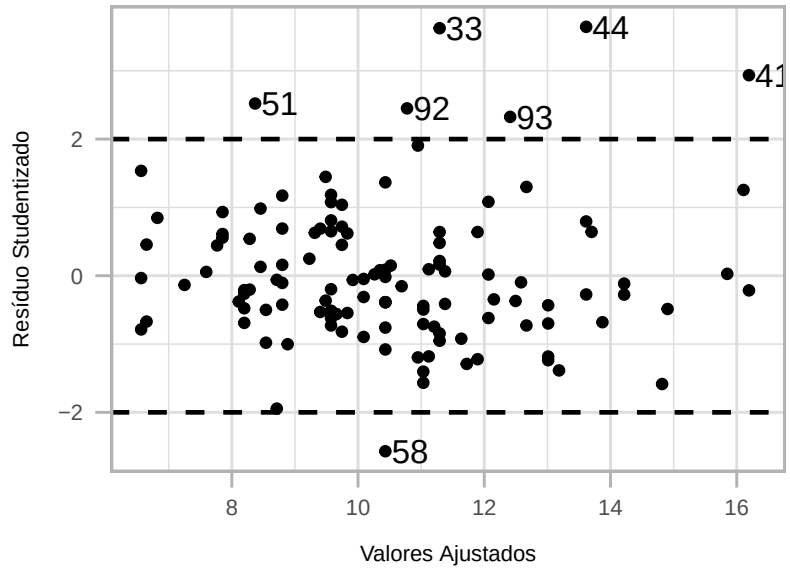
Considerando as etiologias cardíacas

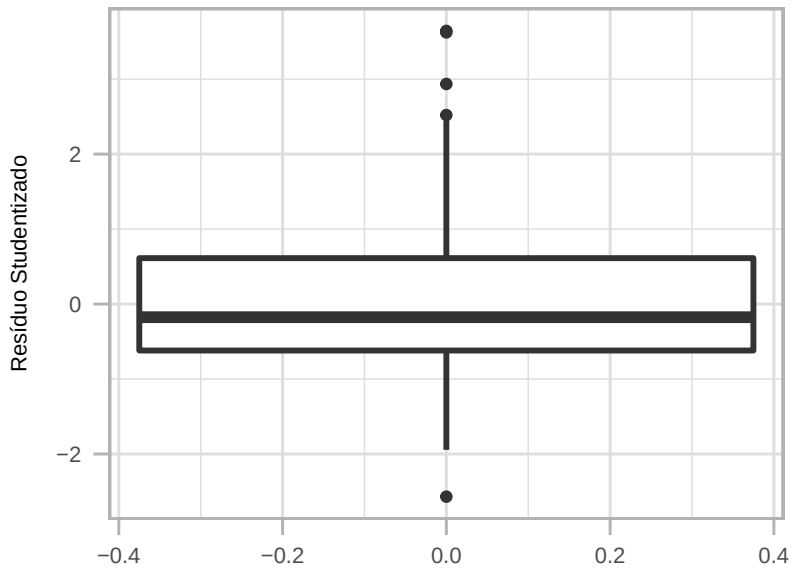


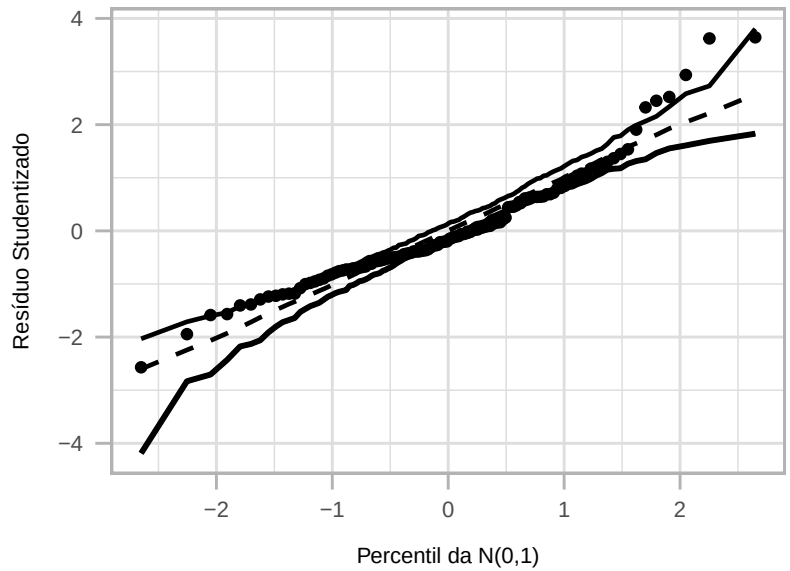

Considerando as etiologias cardíacas
41
0.9
0.6
0.3
0.0
Distância de Cook
0
25
50
75
100
125
Índice
alavanca
0.1
0.2

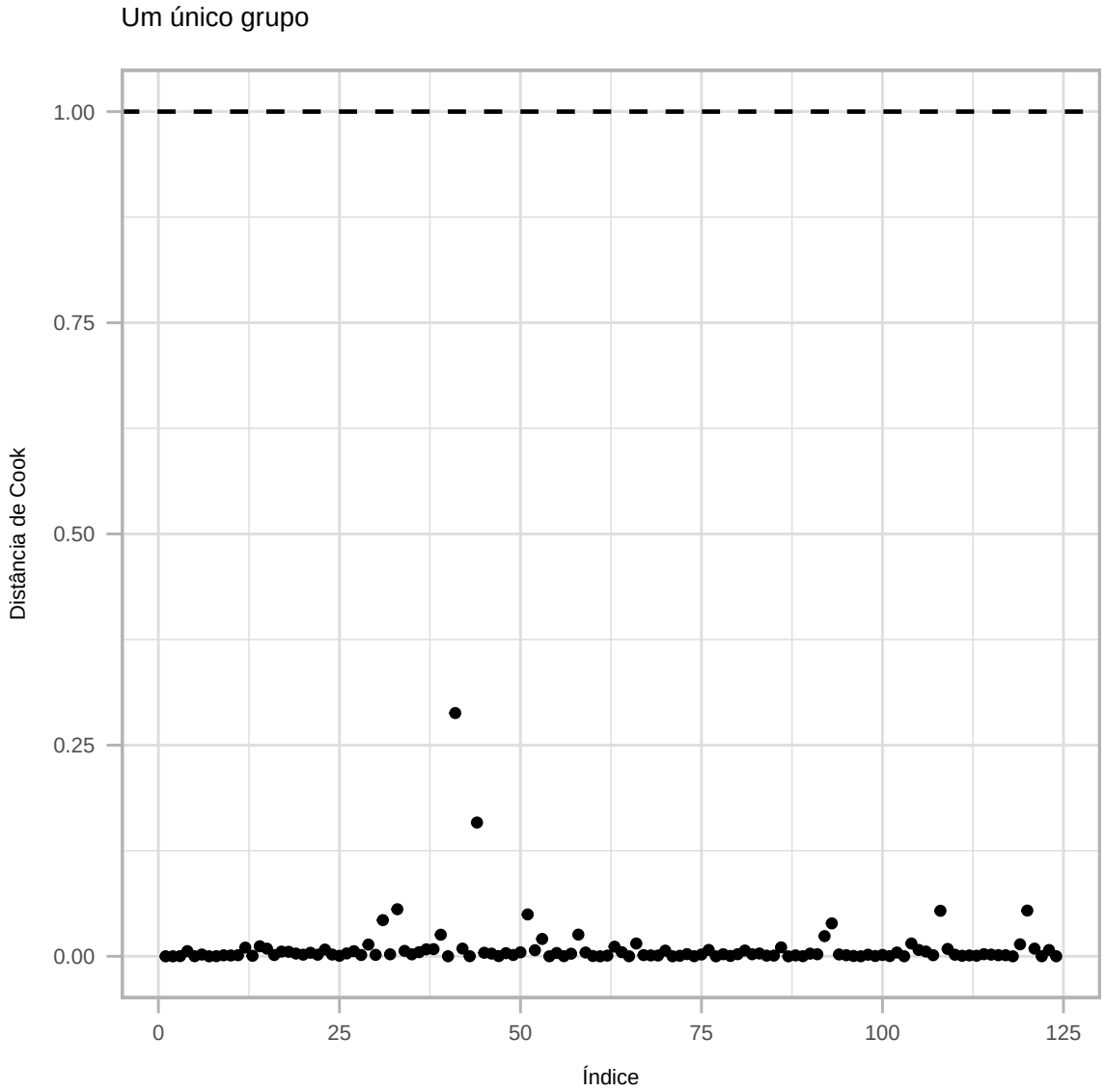
Um único grupo

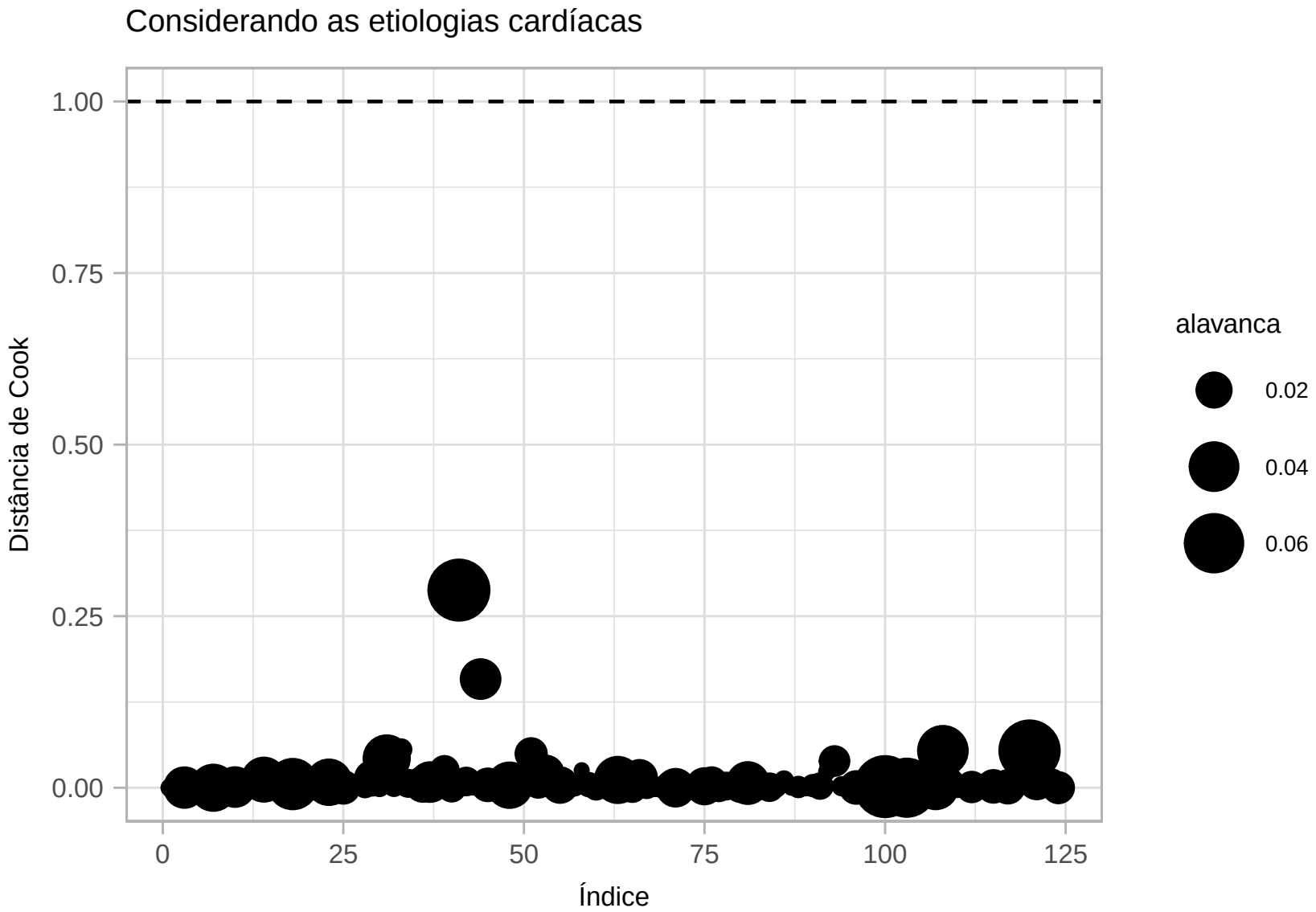









Um único grupo
1.00
0.75
0.50
0.25
0.00
Distância de Cook
0
25
50
75
100
125
Índice

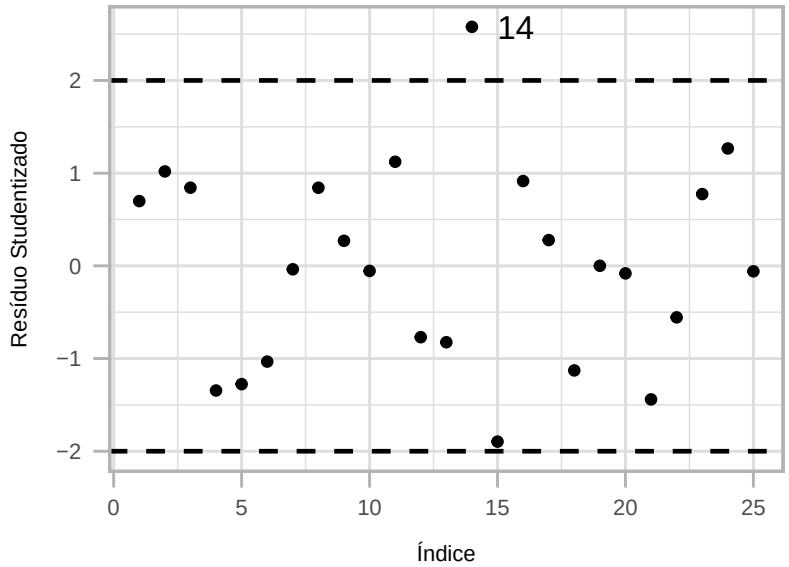

Considerando as etiologias cardíacas
1.00
0.75
0.50
0.25
0.00
Distância de Cook
0
25
50
75
100
125
Índice
alavanca
0.02
0.04
0.06

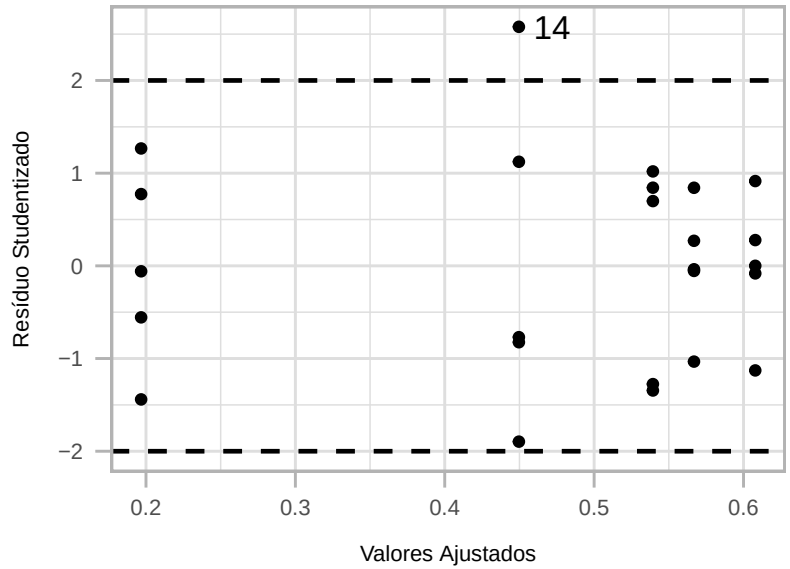
Exemplo 2: considerando todos os tipos de solvente

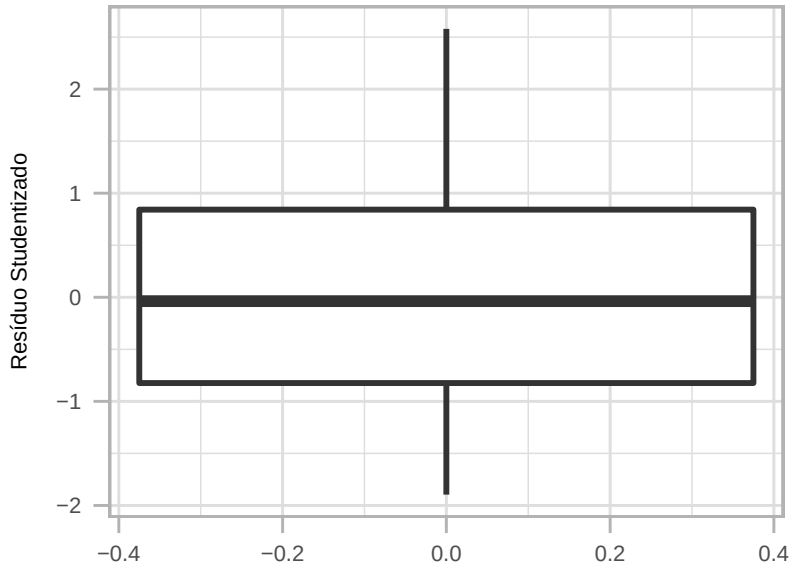
Voltemos ao exemplo 2.

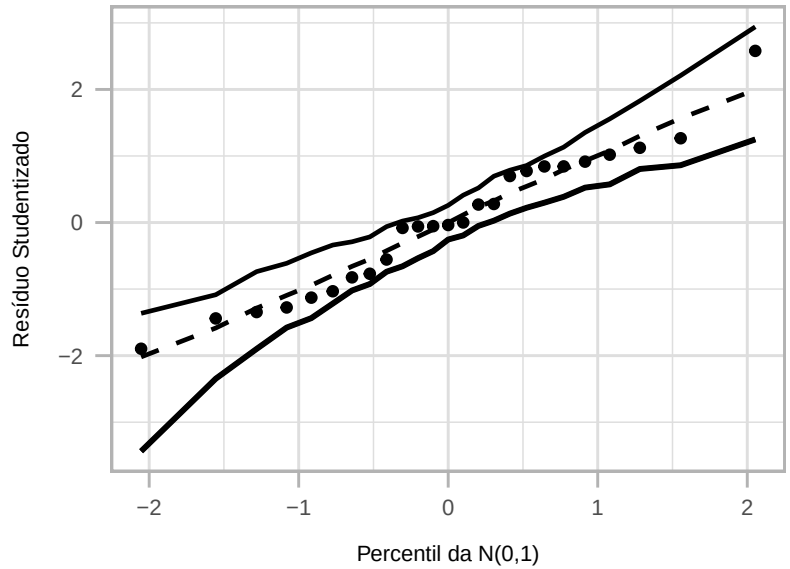
Considerando primeiro o modelo que contempla todos os tipos de solvente e depois o modelo reduzido.

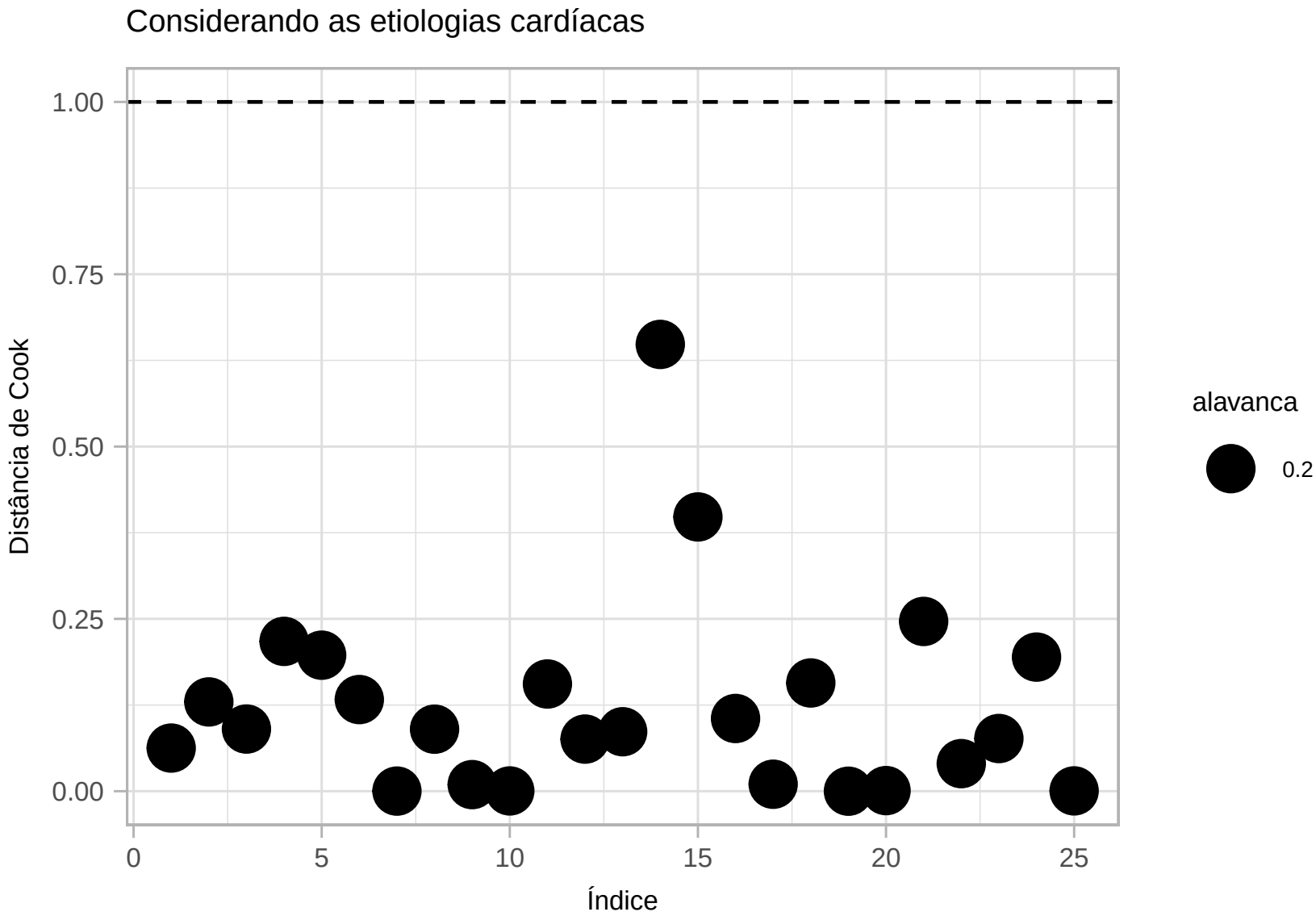
Considerando todos os tipos de solvente



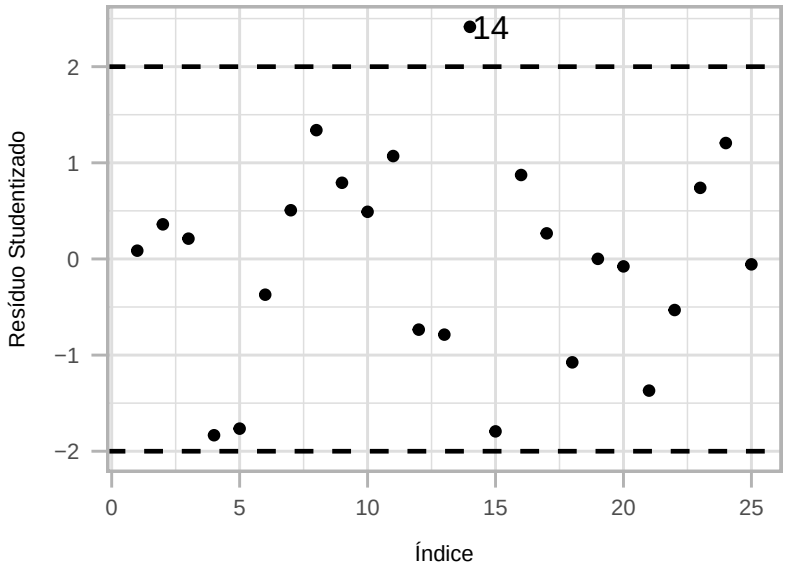


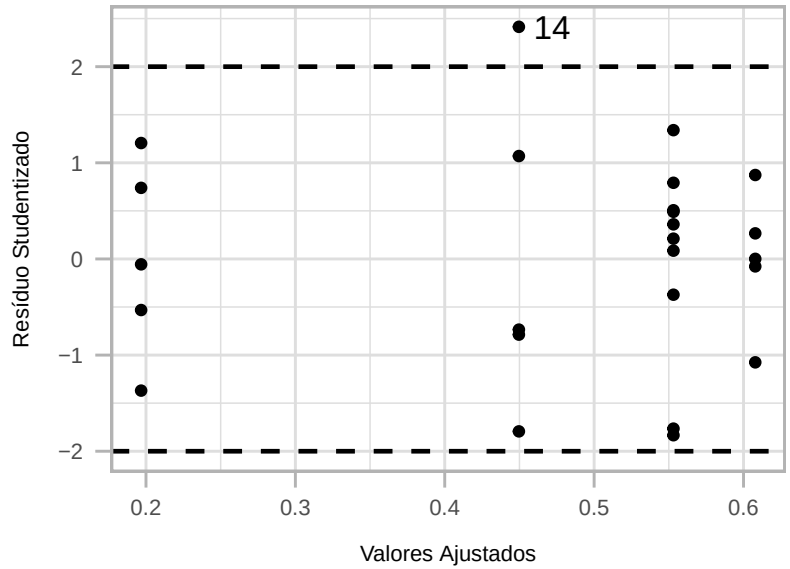


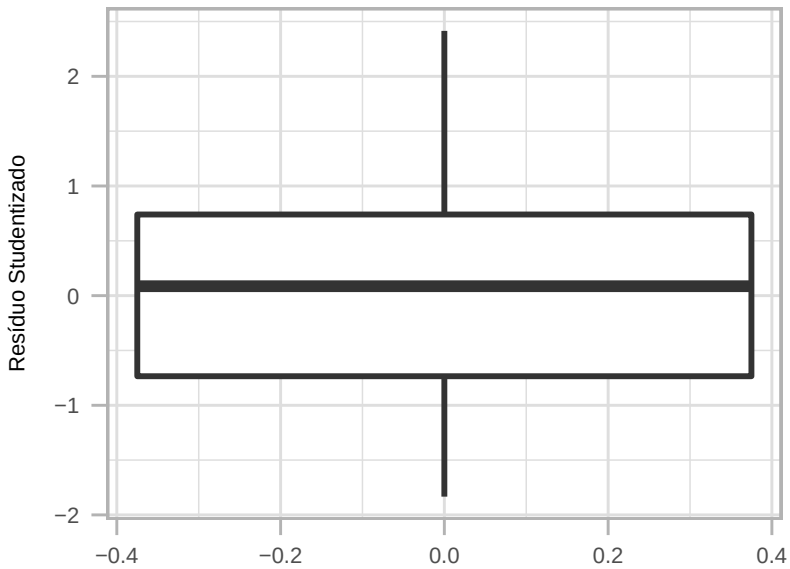


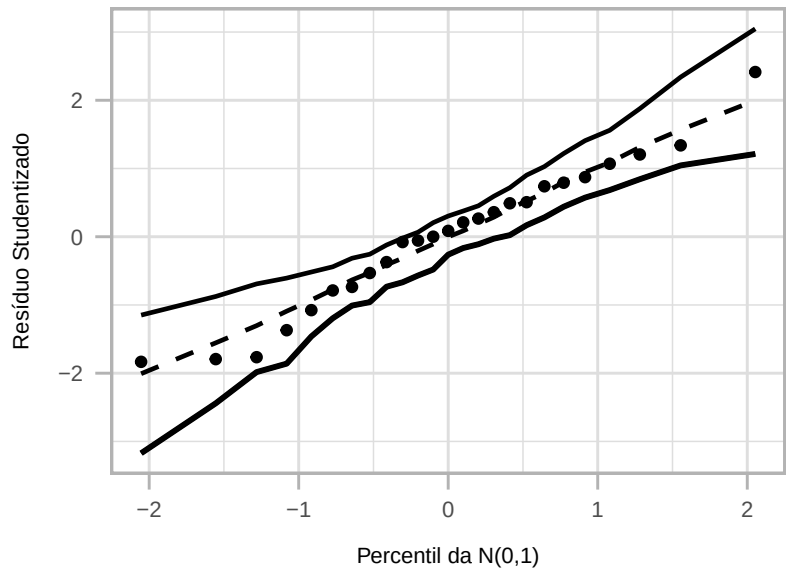

Considerando as etiologias cardíacas
Distância de Cook
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
Índice
alavanca
0.2

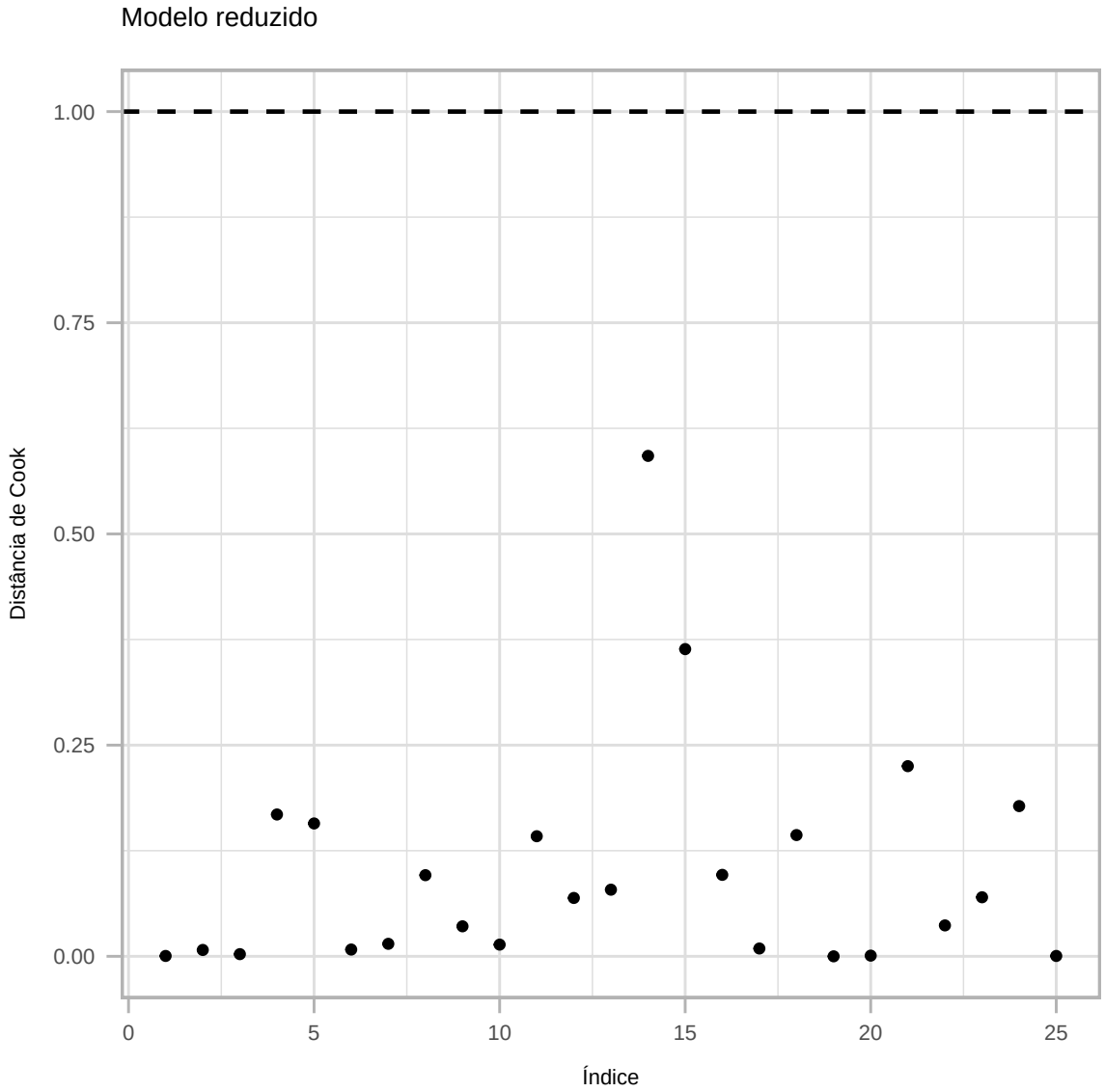
Modelo reduzido

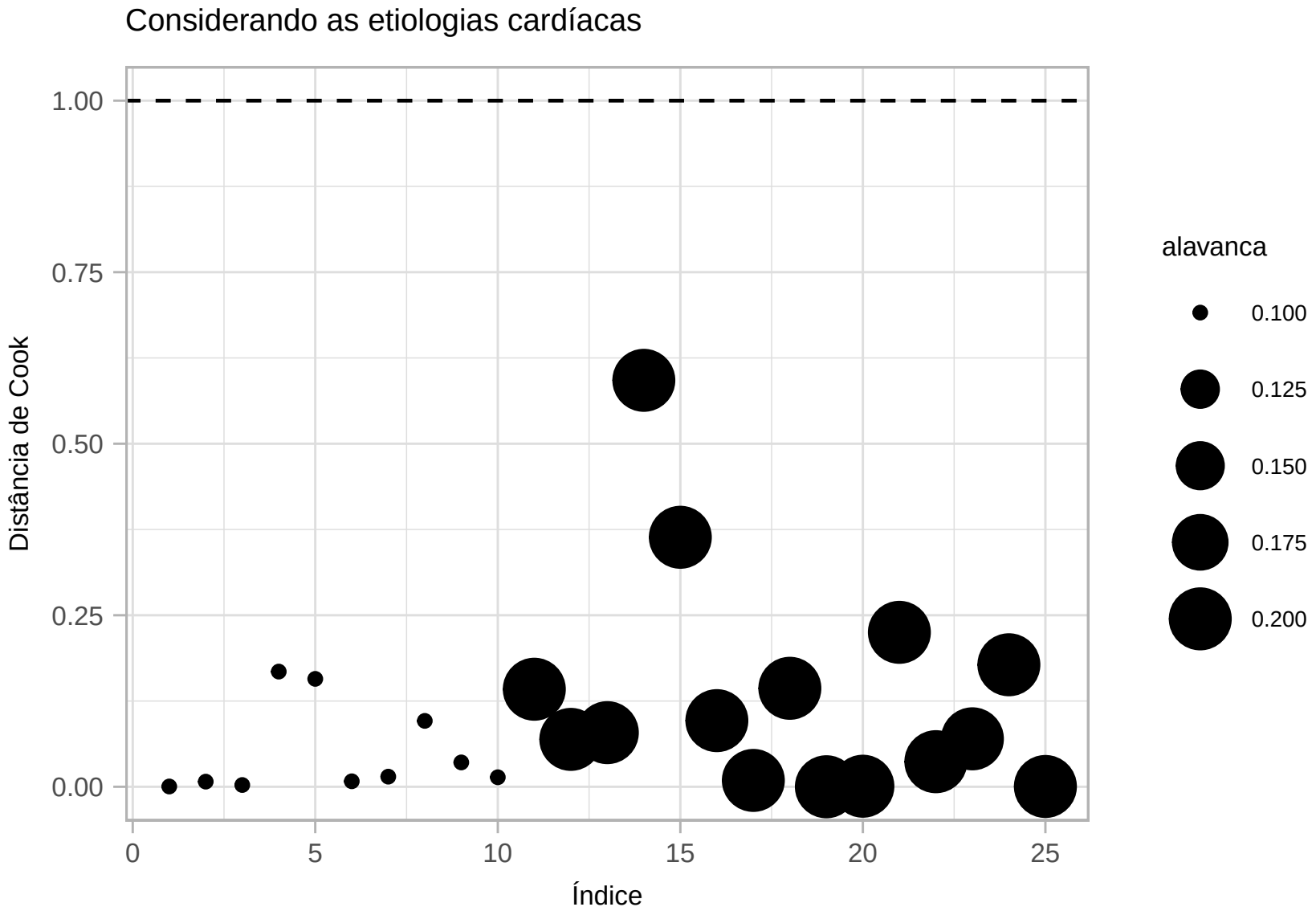









Modelo reduzido
1.00
0.75
0.50
0.25
0.00
Distância de Cook
0
5
10
15
20
25
Índice


Considerando as etiologias cardíacas
Distância de Cook
1.00
0.75
0.50
0.25
0.00
0
5
10
15
20
25
Índice
alavanca
0.100
0.125
0.150
0.175
0.200

Referência

- Notas de aula do Prof. Caio Azevedo.
- Agresti, A. (2015). *Foundations of Linear and Generalized Linear Models*. Wiley series in probability and statistics.