Análise de resíduos no MRNLH (modelo de regressão normal linear homocedástico)

Prof. Caio Azevedo

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

Forma matricial do MRNLH

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

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

Forma matricial do MRNLH (Cont.)

- Suposição: $\boldsymbol{\xi} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$.
- $\boldsymbol{Y}$ é o vetor das variáveis resposta.
- O índice n da variável resposta é geral e pode representar combinações de índices.
- $\boldsymbol{X}$ é a matriz de planejamento (ou delineamento) que define a parte sistemática do modelo.

Suposições

- As principais suposições do MNL são:
  - Homocedasticidade (dos erros).
  - Independência (correlação nula) dos erros.
  - Normalidade dos erros.
- Como verificar as suposições do modelo?
- Como proceder se uma ou mais suposições não forem (satisfatoriamente) válida(s)?

Resíduos

- Como os erros ($\boldsymbol{\xi}$) não são observados (observáveis), precisamos de algum preditor apropriado para avaliar as suposições feitas sobre eles.
- Lembre-se de que $\boldsymbol{\xi} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$ (não correlacionados).
- Já definimos os resíduos ordinários: $\widehat{\xi}_i = R_i = Y_i - \widehat{Y}_i = Y_i - \boldsymbol{X}_i' \widehat{\boldsymbol{\beta}}$.
- Matricialmente

$$\boldsymbol{R} \quad = \quad \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y},$$

em que $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Cont. (livro do Prof. Gilberto, págs. 46 a 48)

- Assim, temos que, sob as suposições do modelo,

  $\boldsymbol{R} \sim N_n(\boldsymbol{0}, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}))$ (são correlacionados). Mais especificamente, $R_i \sim N(0, \sigma^2(1 - h_{ii}))$ e $Cov(R_i, R_j) = -\sigma^2 h_{ij}$, em que h_{ij} é o elemento da i-ésima linha e j-ésima coluna da matriz $\boldsymbol{H}$.

- Defina

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

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

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

Cont. (livro do Prof. Gilberto, págs. 46 a 48)

- Contudo, R_i e S^2 não são independentes (exercício).
- Porém, $S^2_{(i)}$ e R_i o são (em que $S^2_{(i)}$ corresponde à S^2 obtido no modelo sem a i-ésima observação e (i) indica que a i-ésima observação foi excluída) (exercício), em que:

$$S^2_{(i)} = \frac{1}{n-p-1}(\boldsymbol{Y}_{(i)} - \boldsymbol{X}_{(i)}\widehat{\boldsymbol{\beta}}_{(i)})'(\boldsymbol{Y}_{(i)} - \boldsymbol{X}_{(i)}\widehat{\boldsymbol{\beta}}_{(i)}),$$

em que $\widehat{\boldsymbol{\beta}}_{(i)} = (\boldsymbol{X}'_{(i)}\boldsymbol{X}_{(i)})^{-1}\boldsymbol{X}'_{(i)}\boldsymbol{Y}_{(i)}$.

Cont. (livro do Prof. Gilberto, pags. 46 a 48)

- Pode-se provar, além disso, que $S^2_{(i)} = S^2 \left(\dfrac{n - p - V_i^2}{n - p - 1} \right)$ (facilita seu cálculo, dispensando o ajuste do modelo "n" vezes).
- Tem-se, então, que $T_i = \dfrac{R_i}{\sqrt{S^2_{(i)}(1 - h_{ii})}} \sim t_{(n-p-1)}$, sob a validade das hipóteses do modelo (exercício). Lembre-se de que, se $\nu \geq 30$, então $t_{(\nu)} \approx N(0, 1)$.
- Os erros ($\boldsymbol{\xi}$) são variáveis latentes (quantidades não observáveis). Os resíduos ($\boldsymbol{R}$) são valores preditos para os erros.

O que e como observar nos resíduos?

- Gráfico de dispersão dos resíduos versus o índice da observação: identificação de dependência/tendência/correlação.
- Gráfico de dispersão dos resíduos versus os valores ajustados: homocedasticidade.
- Boxplot e/ou gráfico de quantis-quantis: simetria, identificação de "out-liers" e multimodalidade.
- Problema no gráfico de quantis-quantis: Visualmente, muitas vezes, é complicado avaliar a proximidade dos quantis.
- Solucão: criar bandas de confiança (gráficos de envelope).

link

Procedimento para se gerar o gráfico de envelopes

1) Gera-se n observações N(0,1) as quais são armazenadas em $\boldsymbol{z} = (z_1, ..., z_n)'$.
2) Calcula-se $\boldsymbol{r}^* = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{z}$ e depois $t_i^* = \dfrac{r_i^*}{\sqrt{1 - h_{ii}}}$.
3) Repete-se os passos (1)-(2), m vezes. Logo, teremos $t_{ij}^*, i = 1, ..., n$ e $j = 1, ..., m$.

Procedimento para se gerar o gráfico de envelopes

4) Ao final teremos uma matriz com os resíduos, ou seja t_{ij}^*, i=1,...,n, (tamanho da amostra) j=1,...,m (réplica).

$$\boldsymbol{T}_1 = \begin{bmatrix} t_{11}^* & t_{12}^* & \cdots & t_{1m}^* \\ t_{21}^* & t_{22}^* & \cdots & t_{2m}^* \\ \vdots & \vdots & \ddots & \vdots \\ t_{n1}^* & t_{n2}^* & \cdots & t_{nm}^* \end{bmatrix}$$

Procedimento para se gerar o gráfico de envelopes

5) Dentro de cada amostra, ordena-se, de modo crescente, os resíduos, obtendo-se $t^*_{(i)j}$ (estatísticas de ordem):

$$\boldsymbol{T}_2 = \begin{bmatrix} t^*_{(1)1} & t^*_{(1)2} & \cdots & t^*_{(1)m} \\ t^*_{(2)1} & t^*_{(2)2} & \cdots & t^*_{(2)m} \\ \vdots & \vdots & \ddots & \vdots \\ t^*_{(n)1} & t^*_{(n)2} & \cdots & t^*_{(n)m} \end{bmatrix}$$

6) Obtem-se os limites $t^*_{(i)I} = \min\limits_{1\leq j\leq m} t^*_{(i)j}$ e $t^*_{(i)S} = \max\limits_{1\leq j\leq m} t^*_{(i)j}$, $j = 1, 2, ..., m$.

Procedimento para se gerar o gráfico de envelopes

7) Na prática considera-se $t^*_{(i)I} = \dfrac{t^*_{(i)(2)} + t^*_{(i)(3)}}{2}$ e $t^*_{(i)S} = \dfrac{t^*_{(i)(m-2)} + t^*_{i(m-1)}}{2}$ (refinamento das estimativas do mínimo e máximo), em que $t^*_{(i)(r)}$ é a r-ésima estatística de ordem dentro de cada linha, $i = 1, 2, ..., n$.

- Além disso, consideramos como linha de referência $t^*_{(i)} = \dfrac{1}{m}\sum_{j=1}^{m} t^*_{(i)j}, i = 1, 2, ..., n.$
- Obs: Como veremos adiante, os Passos 1) e 2) podem ser substituídos pelo cálculo dos resíduos com base em valores simulados e ajustados a partir do modelo de regressão em uso.

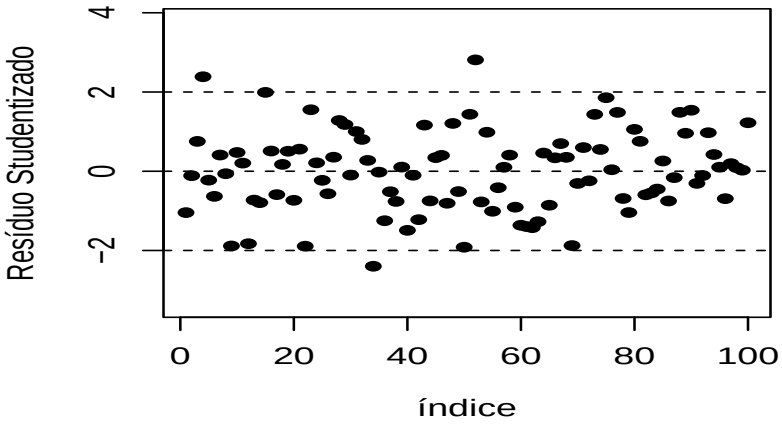
Estudo de simulação

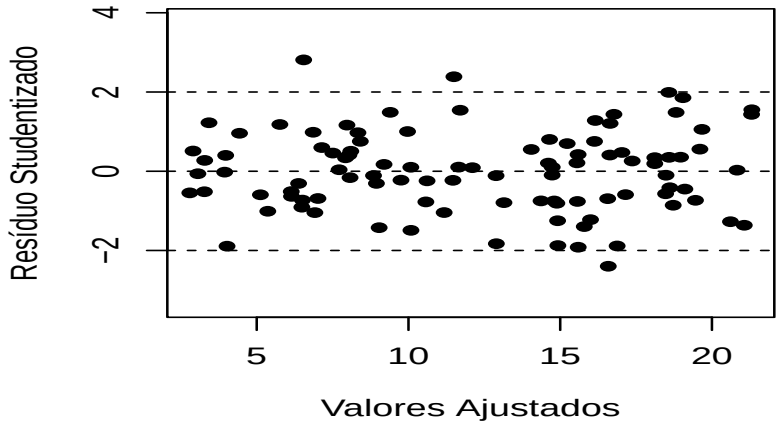
- Vamos avaliar o comportamento dos resíduos sob:
  - Heterocedasticidade.
  - Correlação entre as observações.
  - Ausência de normalidade.

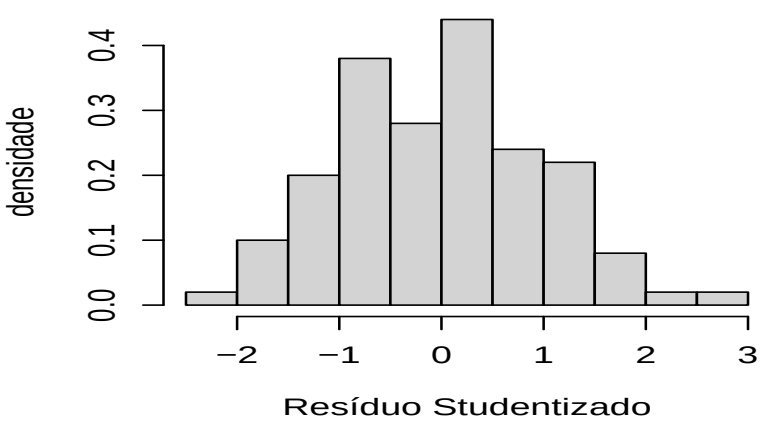
Heterocedasticidade

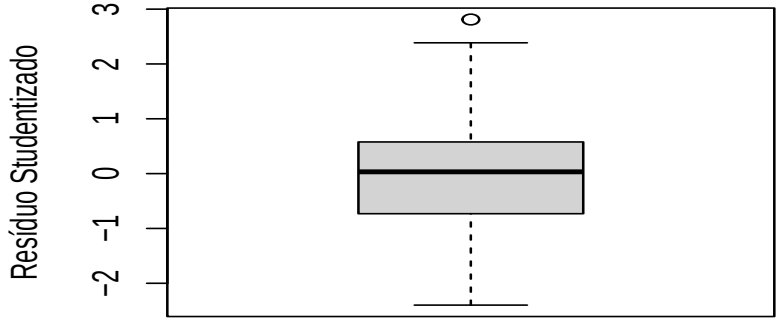
- Modelo 1 (M1): $Y_i = 1 + 2x_i + \xi_i, i = 1, 2, ..., 100$, $x_i \overset{i.i.d.}{\sim} U(1, 10)$ e $\xi \overset{i.i.d.}{\sim} N(0, 4)$.
- Modelo 2 (M2): M1 com $\xi_i \overset{ind.}{\sim} N(0, 4x_i)$.
- Modelo 3 (M3): M1 com $\xi_i \overset{ind.}{\sim} N(0, 4x_i^{-1})$.
- Modelo 4 (M4): M1 com $\xi_i \overset{ind.}{\sim} N(0, 4x_i), i = 1, 2, ..., 50$ e $\xi_i \overset{ind.}{\sim} N(0, 4x_i^{-1}), i = 51, 2, ..., 100$.

Modelo 1: diagnóstico

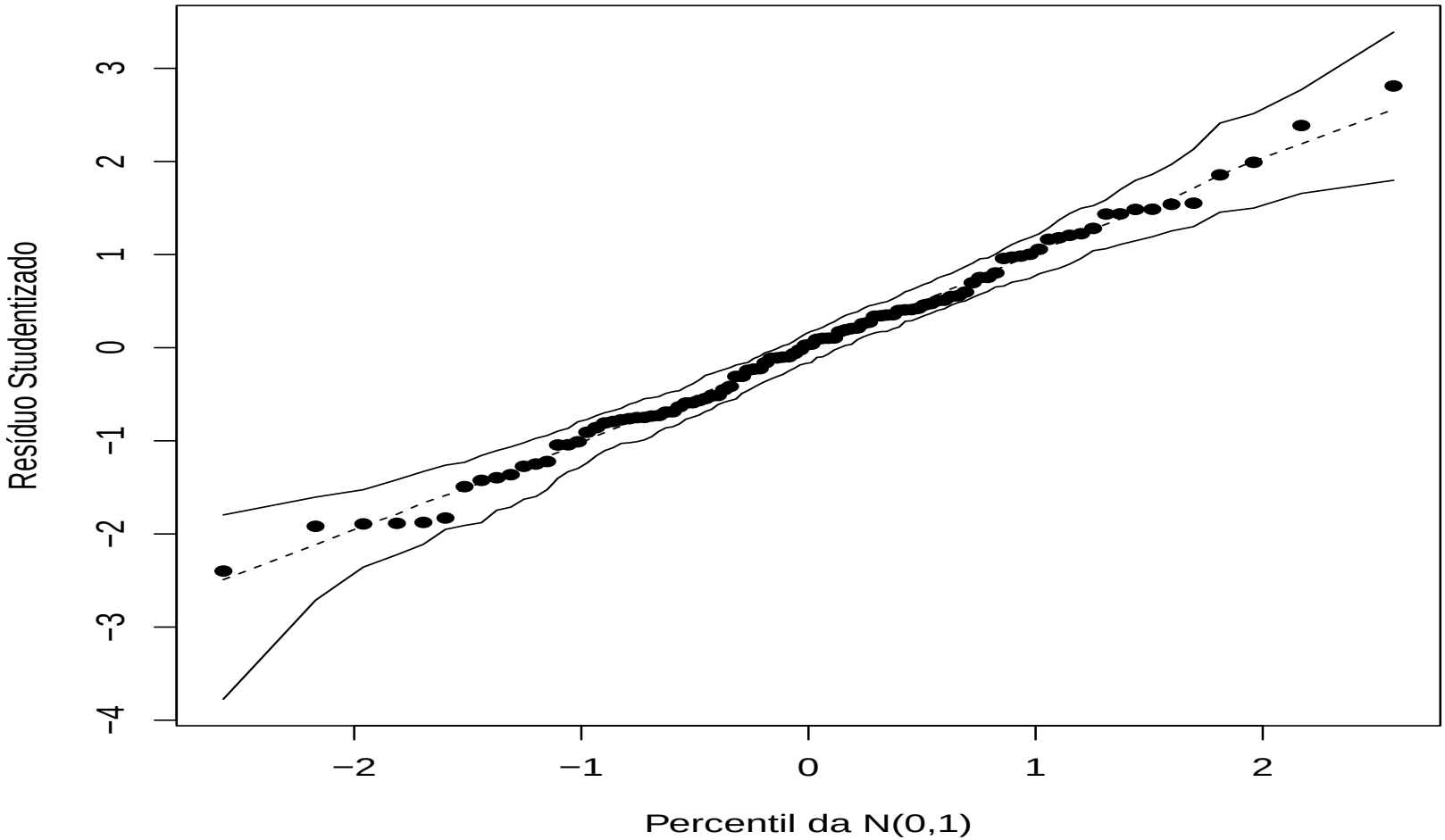




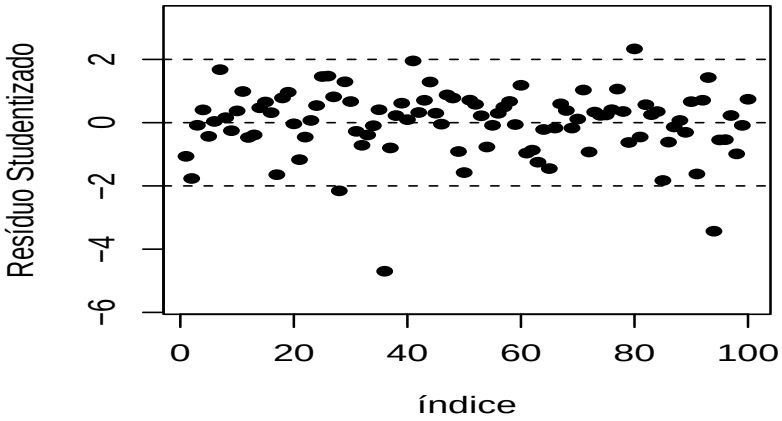


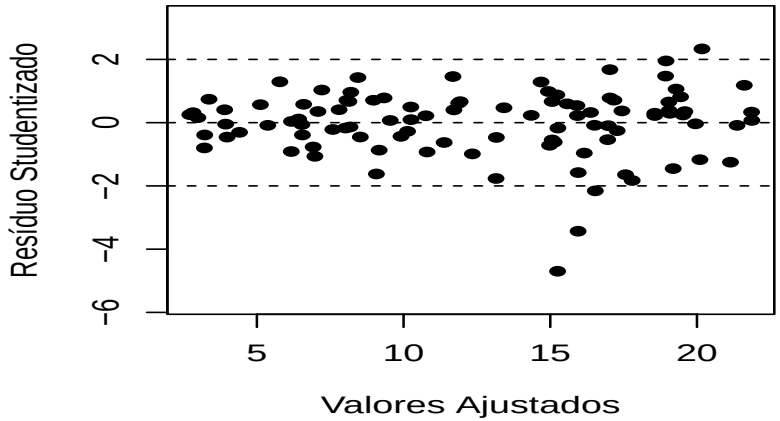


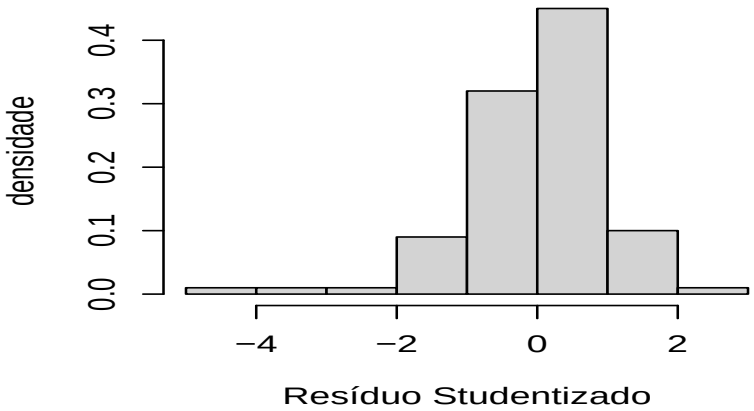
Modelo 1: envelope

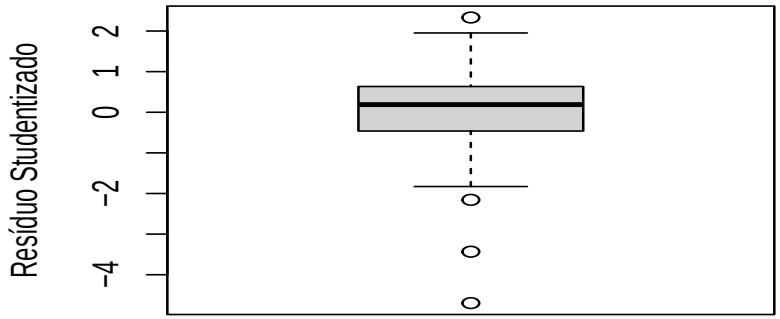


Modelo 2: diagnóstico

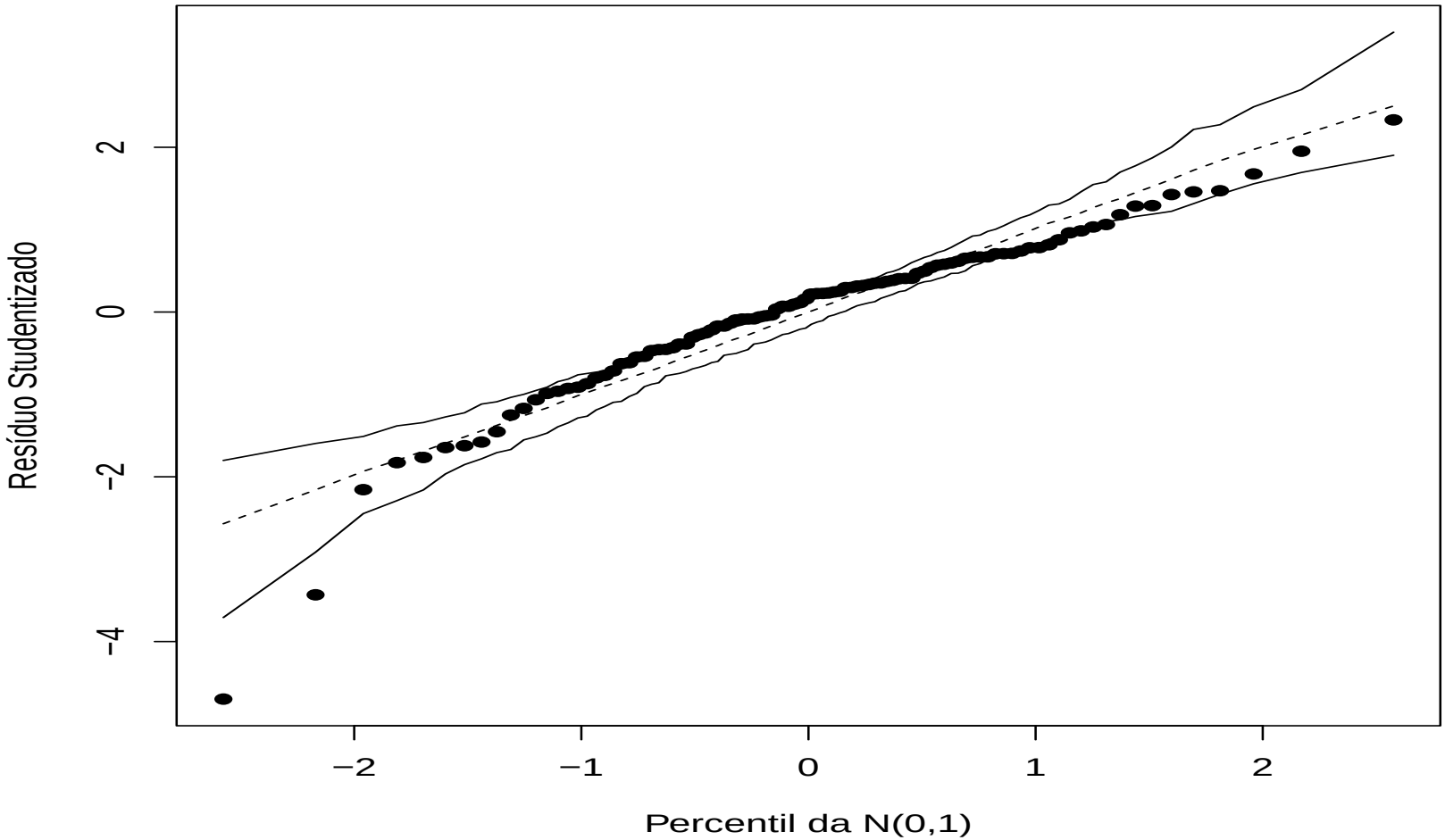




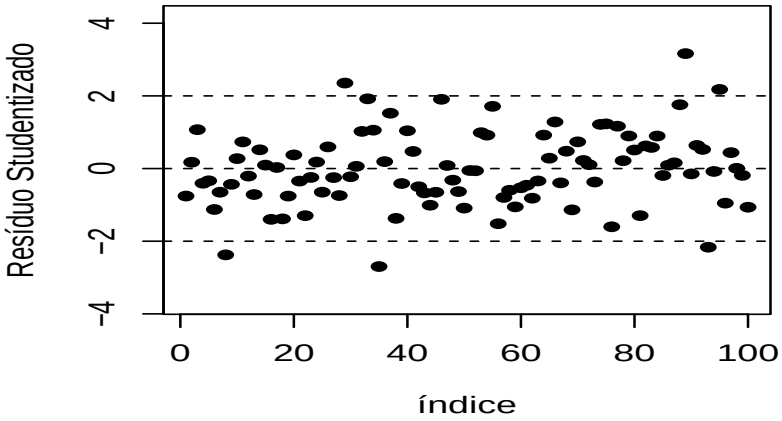


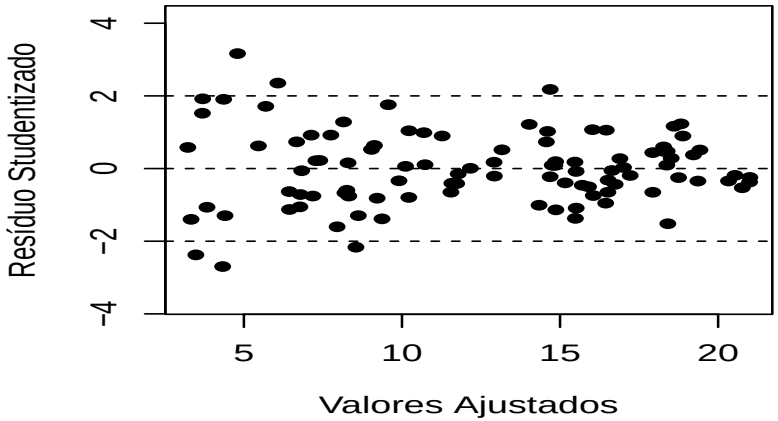


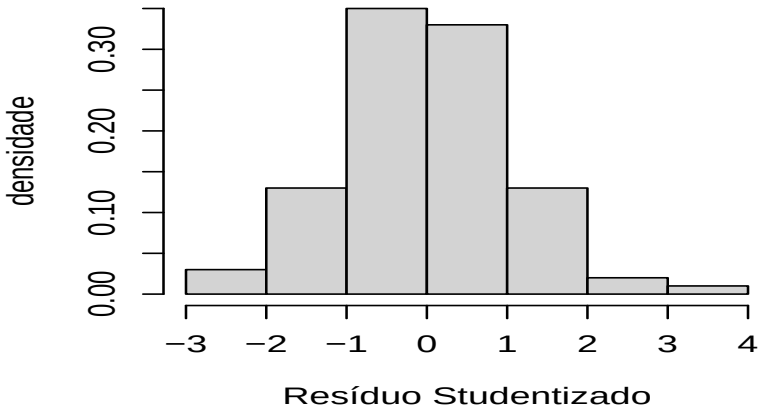
Modelo 2: envelope

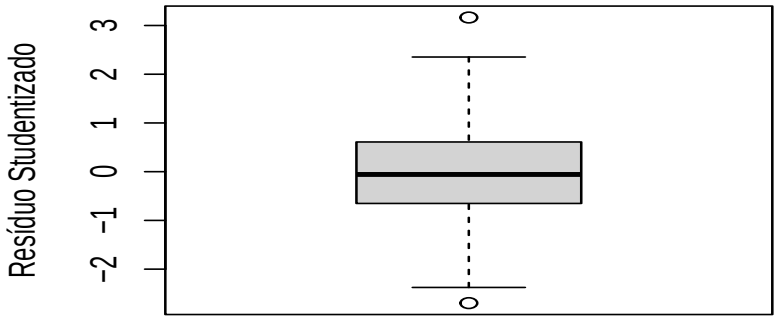


Modelo 3: diagnóstico

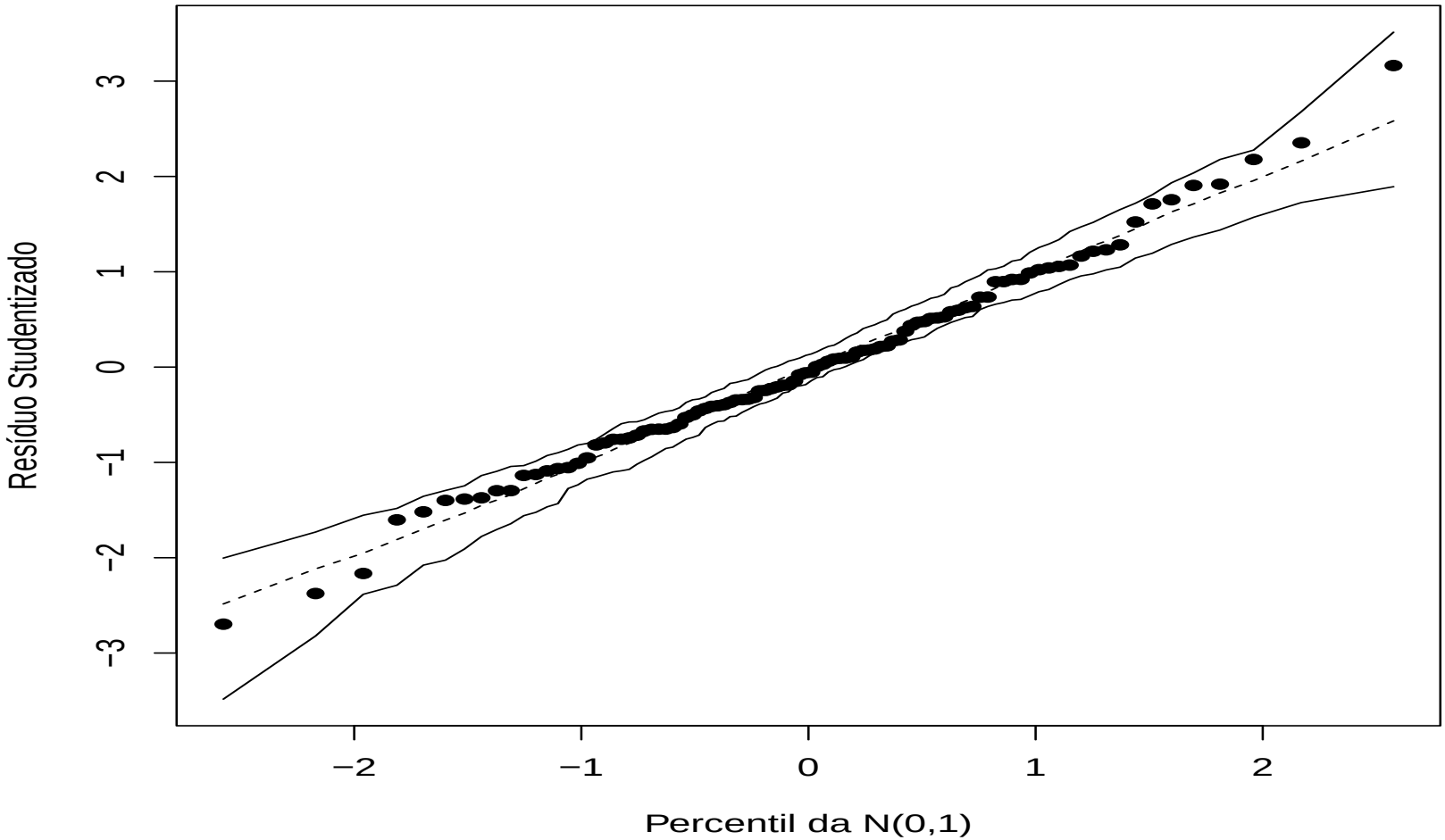




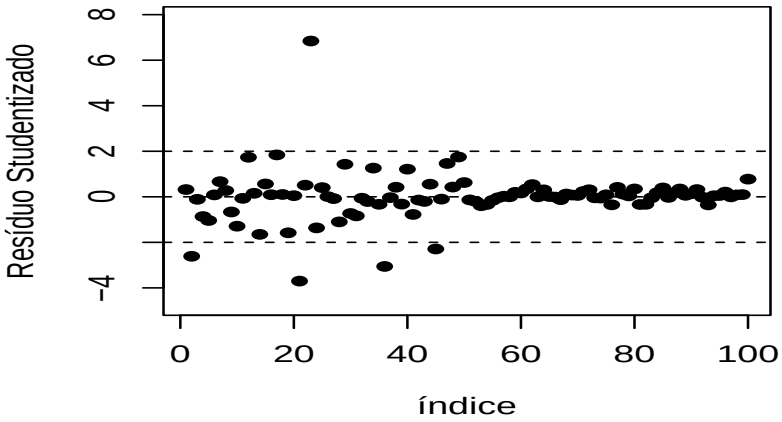


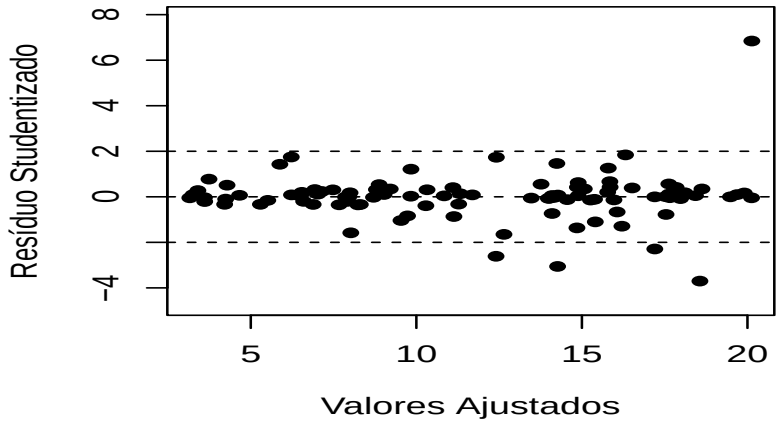


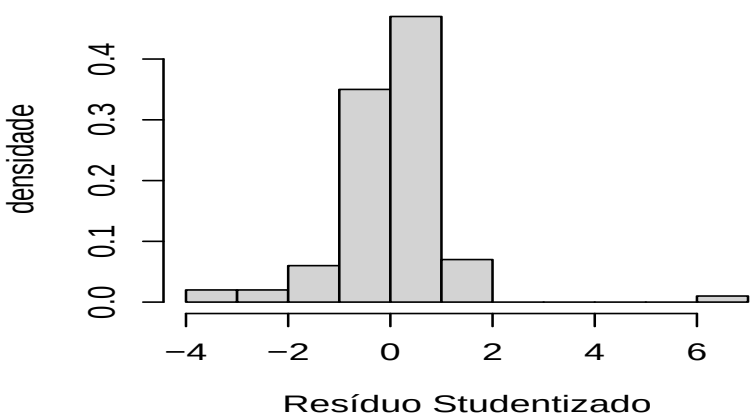
Modelo 3: envelope

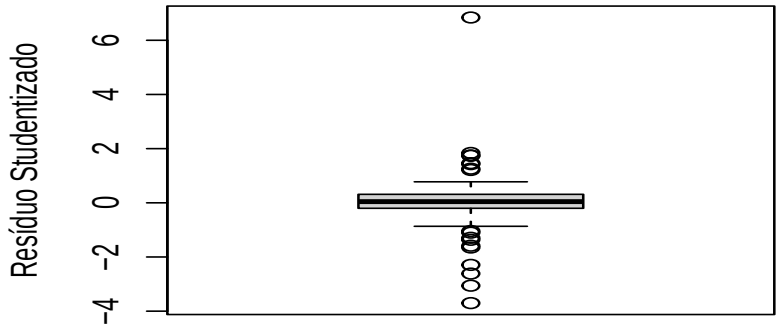


Modelo 4: diagnóstico

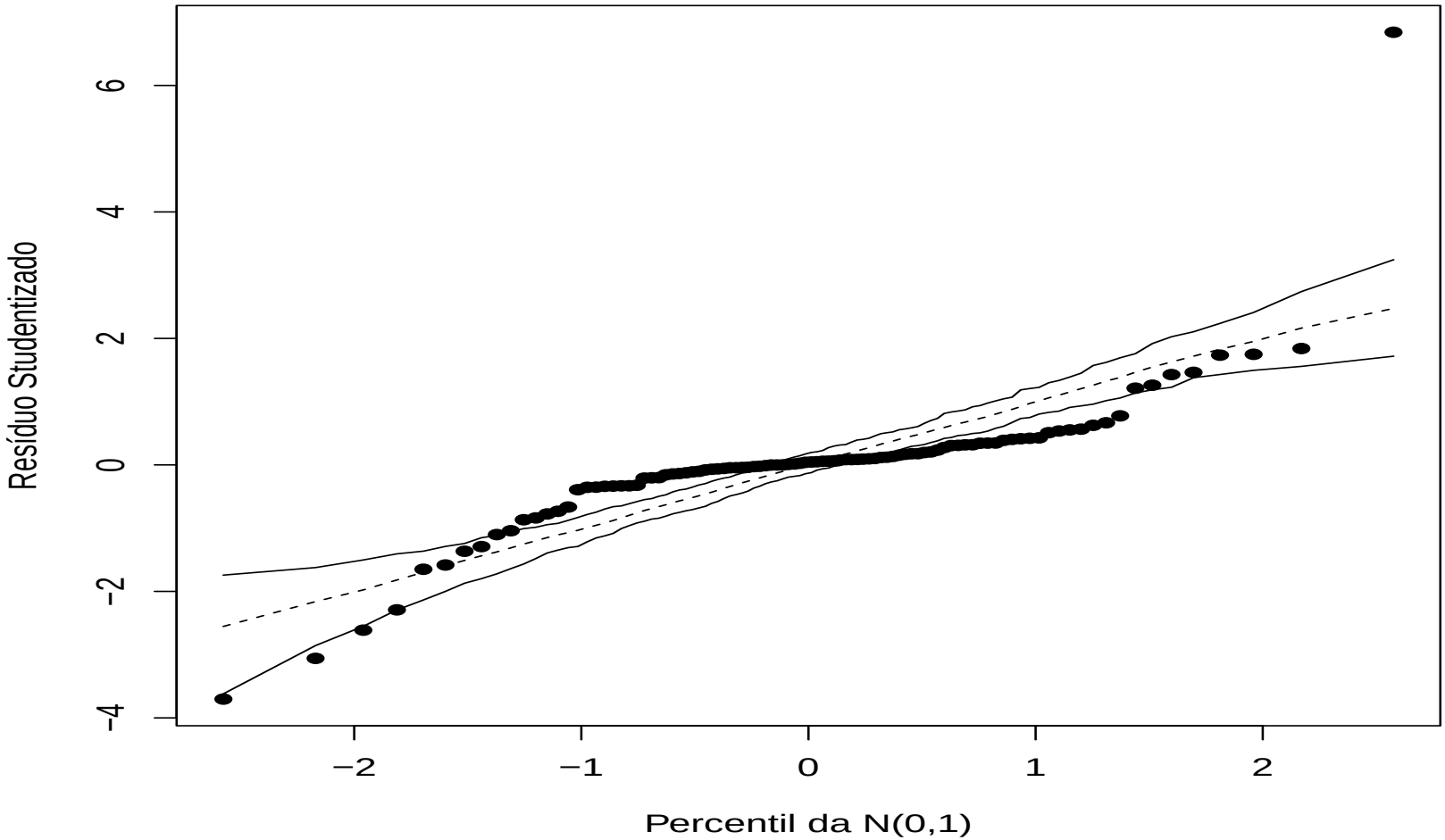








Modelo 4: envelope



Dependência

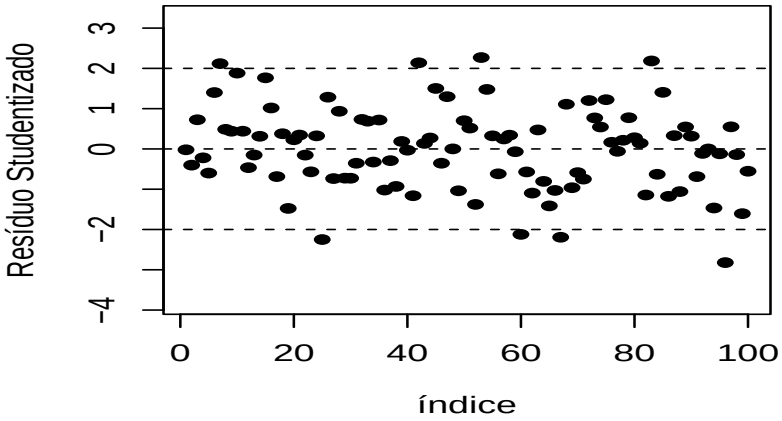
- Modelo 1 (M1): $Y_i = 1 + 2x_i + \xi_i, i = 1, 2, ..., 100$, $x_i \overset{i.i.d.}{\sim} U(1, 10)$ e $\xi \overset{i.i.d.}{\sim} N(0, 4)$.
- Modelo 2 (M2): $Y_i, i = 1, ..., 100$ segue um processo AR(1) com $\rho = 0,8$, ie, $Y_i = 1 + 2x_i + \xi_i$, $\xi_i = \rho\xi_{i-1} + \epsilon_i$, $i = 2, 3, ..., 100$, $\xi_1 = \epsilon_1, \epsilon_1 \sim N(0, 1)$ (correlação entre as observações).
- Modelo 3 (M3): M1 com $(\xi_i, \xi_{i+1})' \sim N_2\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 3,2 \\ 3,2 & 4 \end{bmatrix}\right)$, i=1,3,5,...,99.

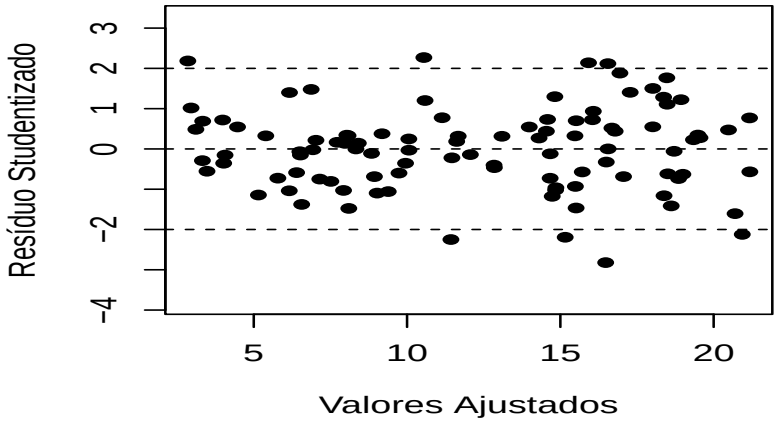
Dependência

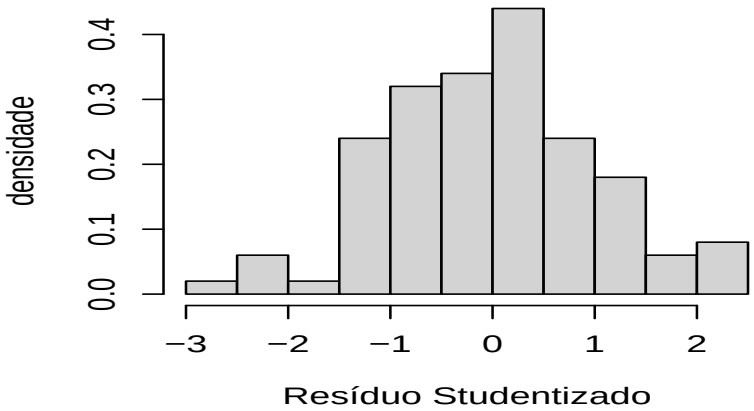
- Modelo 4 (M4): M1 com $\boldsymbol{\xi}_1 \sim N_{50}(\mathbf{0}, \boldsymbol{\Sigma}_1)$ e $\boldsymbol{\xi}_2 \sim N_{50}(\mathbf{0}, \boldsymbol{\Sigma}_2)$, em que $\boldsymbol{\xi}_1 = (\xi_1, ..., \xi_{50})'$ e $\boldsymbol{\xi}_2 = (\xi_{51}, ..., \xi_{100})'$.

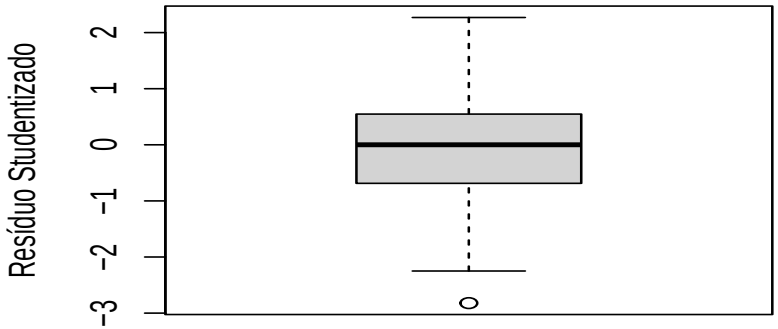
$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 4 & 3.2 & \dots & 3,2 \\ 3,2 & 4 & \dots & 3,2 \\ \vdots & \vdots & \ddots & \vdots \\ 3,2 & 3,2 & \dots & 4 \end{bmatrix}; \boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 3,6 & \dots & 3,6 \\ 3,6 & 4 & \dots & 3,6 \\ \vdots & \vdots & \ddots & \vdots \\ 3,6 & 3,6 & \dots & 4 \end{bmatrix}$$

Modelo 1: diagnóstico

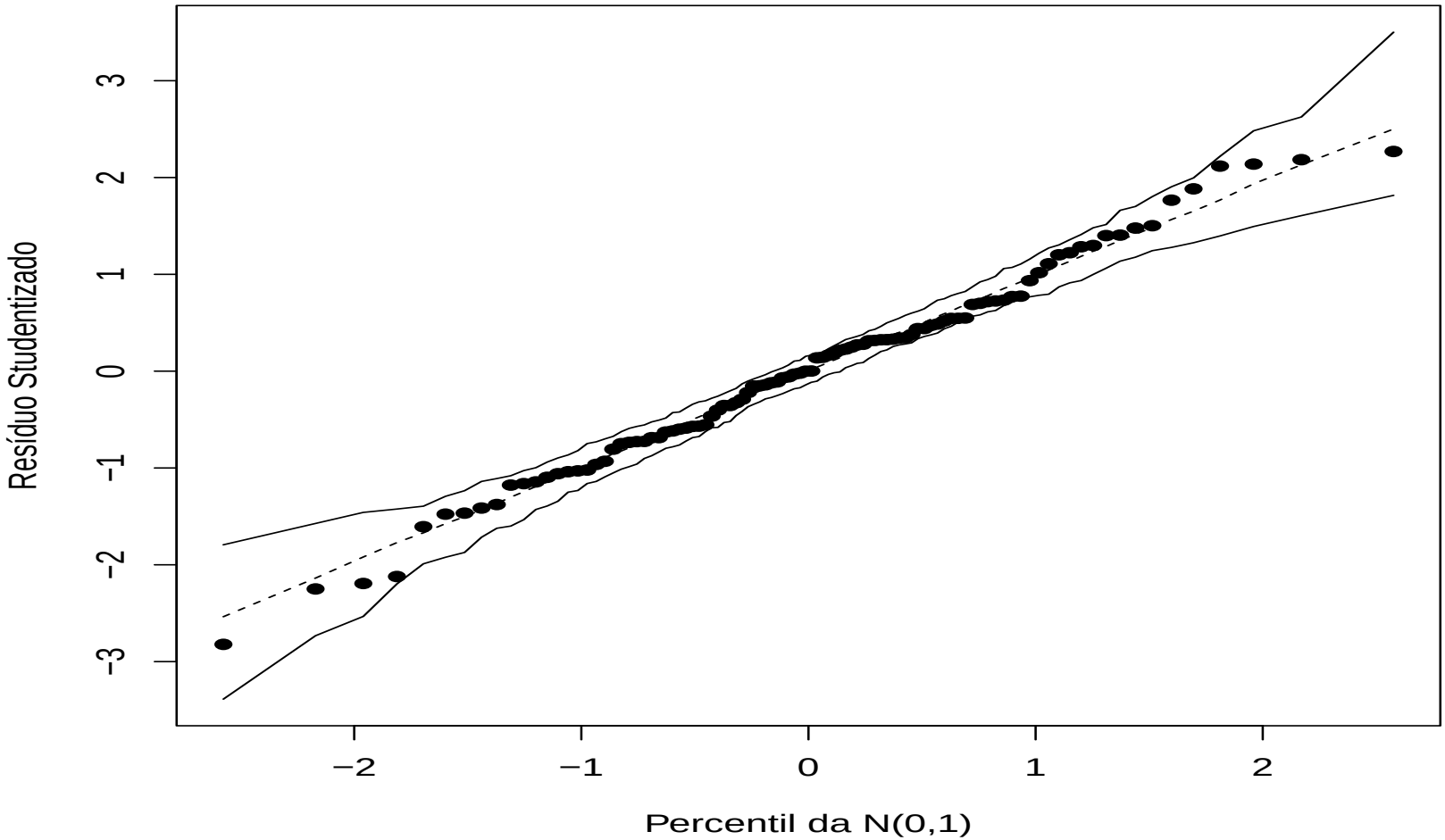




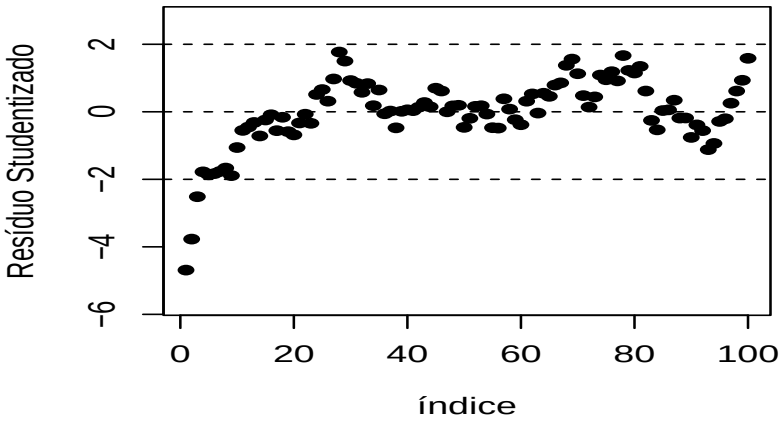


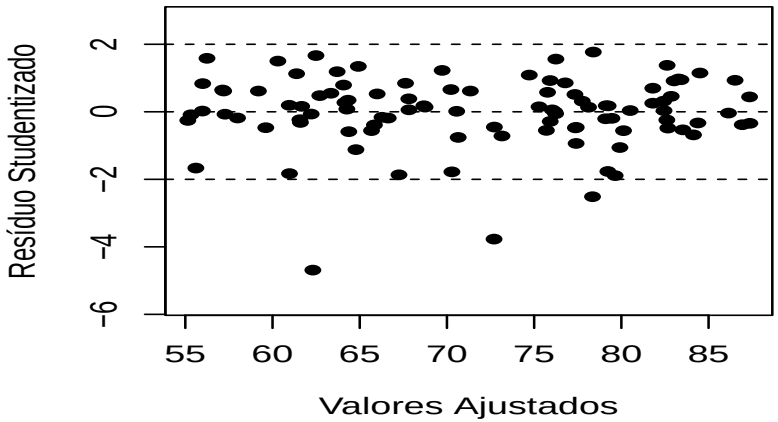


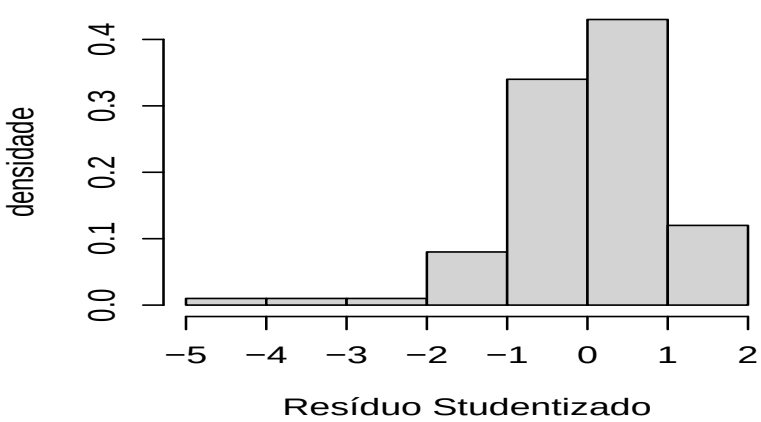
Modelo 1: envelope

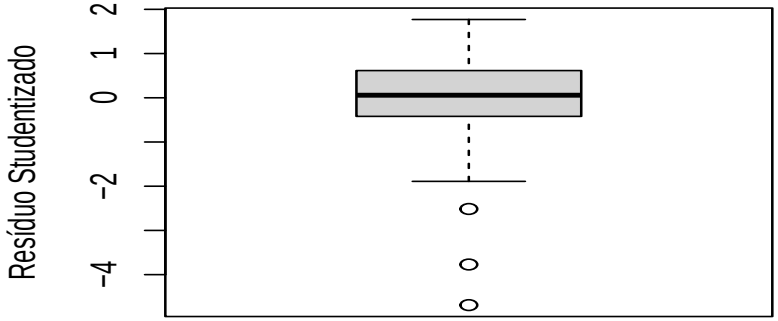


Modelo 2: diagnóstico

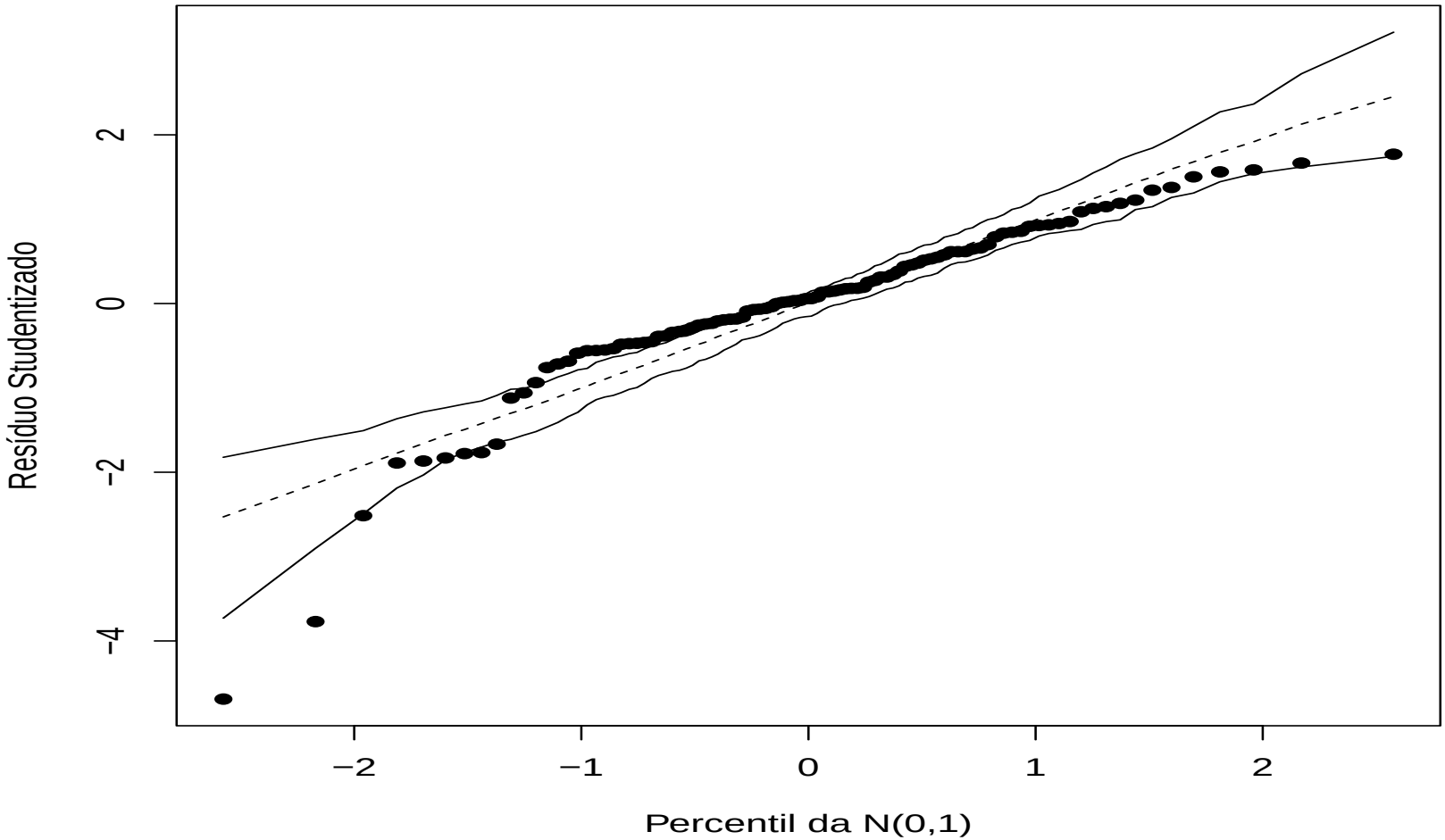




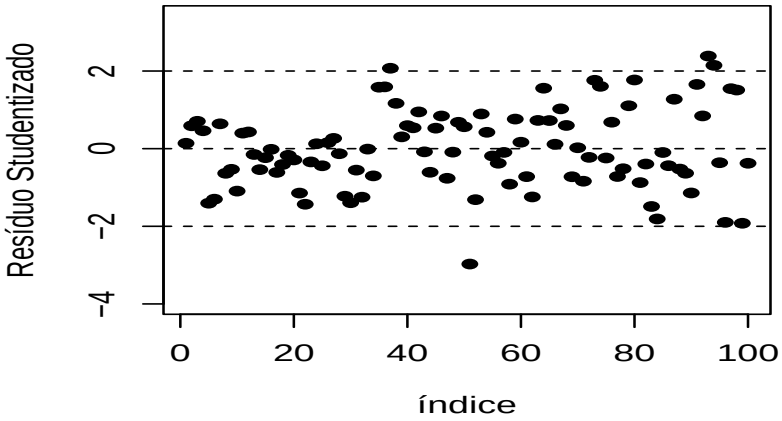


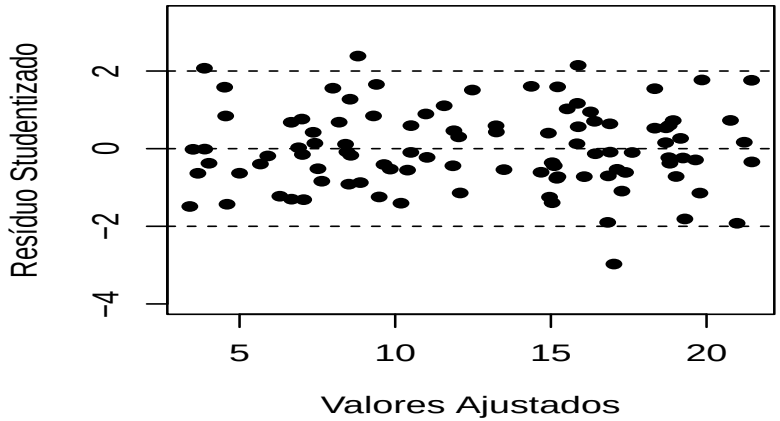


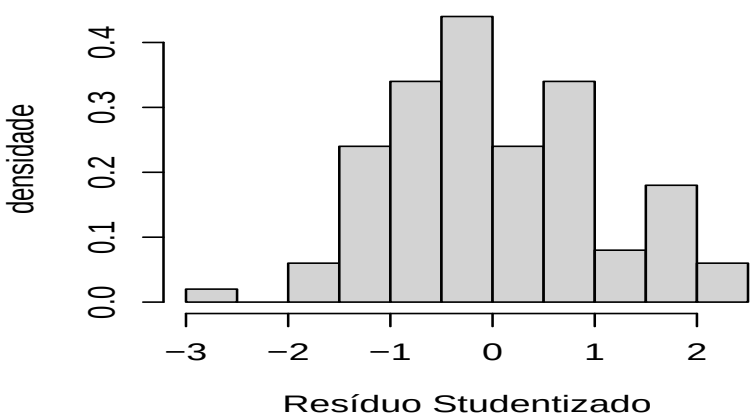
Modelo 2: envelope

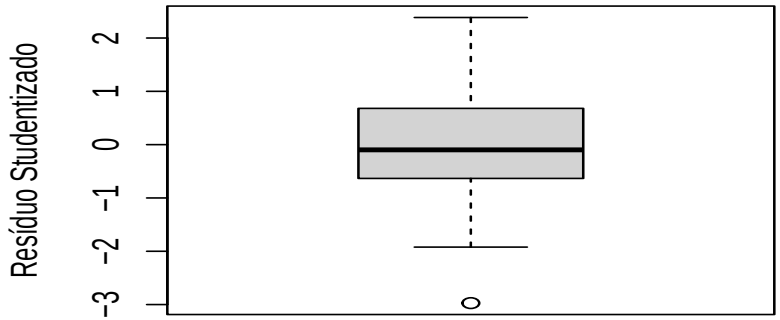


Modelo 3: diagnóstico

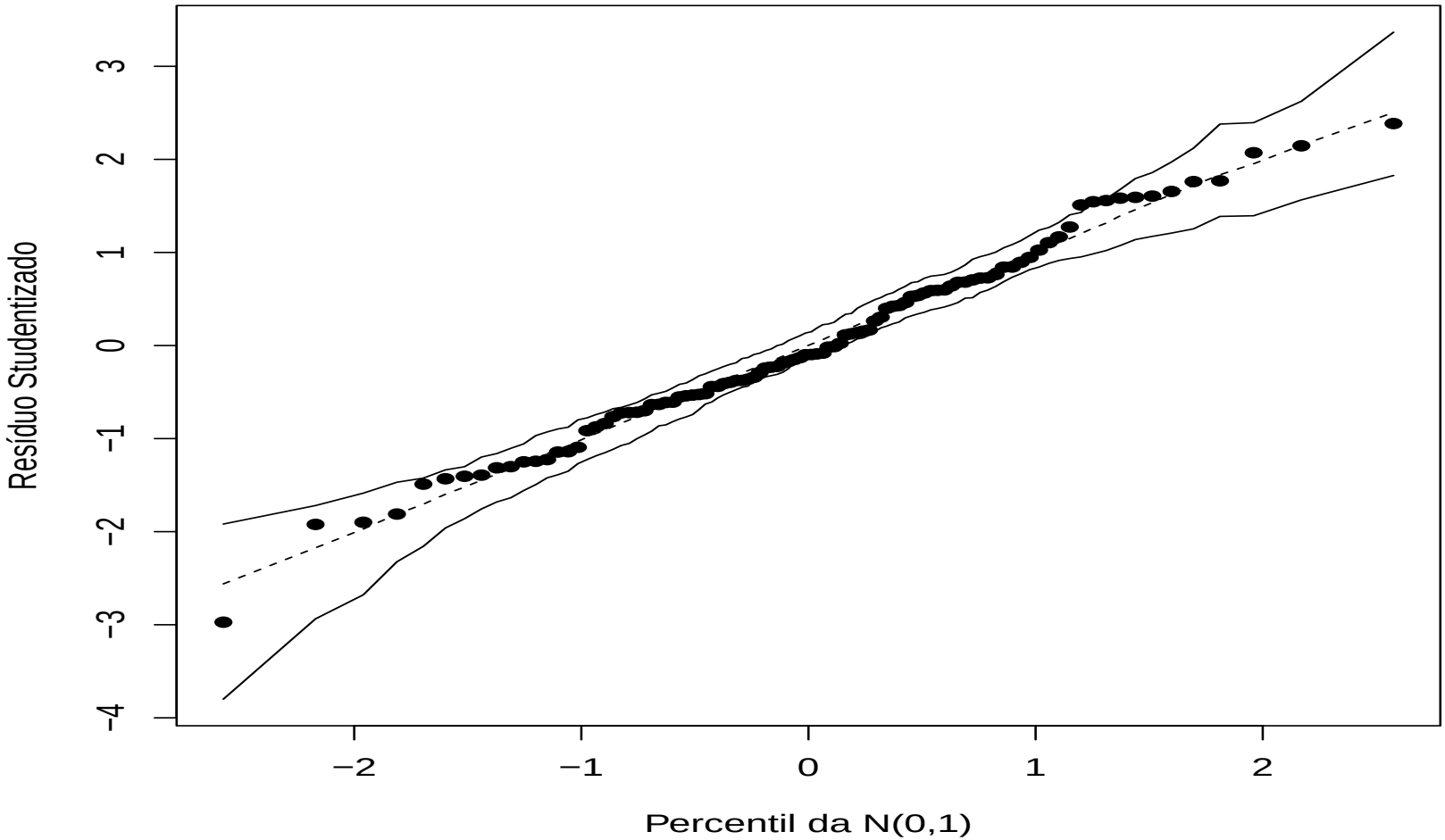




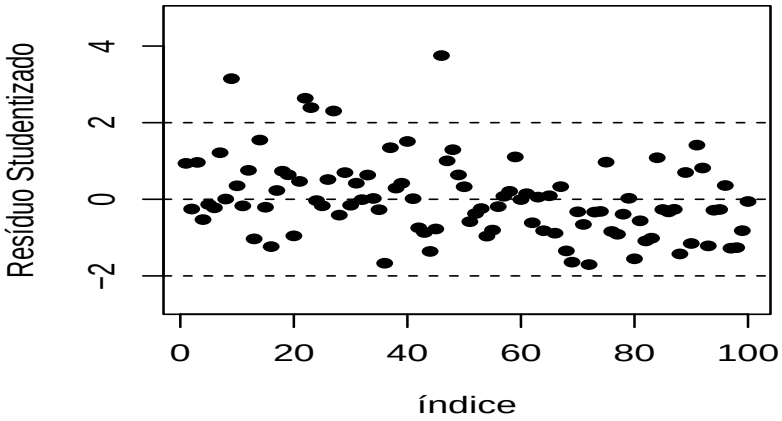


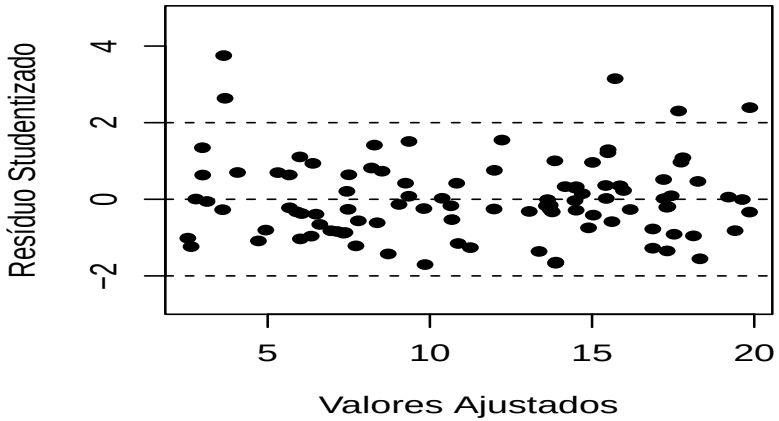


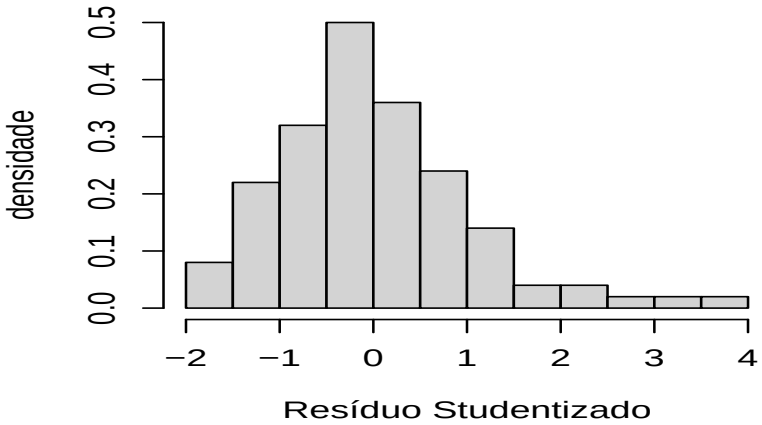
Modelo 3: envelope

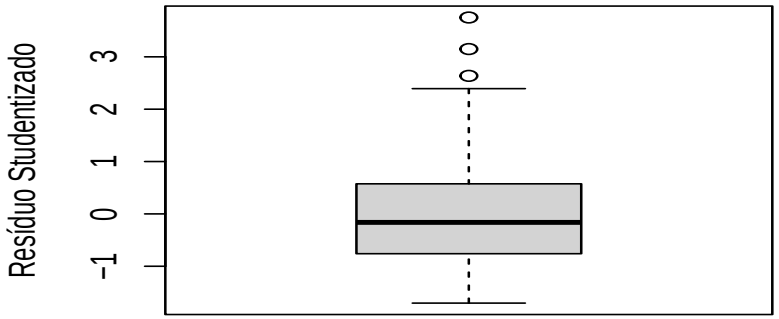


Modelo 4: diagnóstico

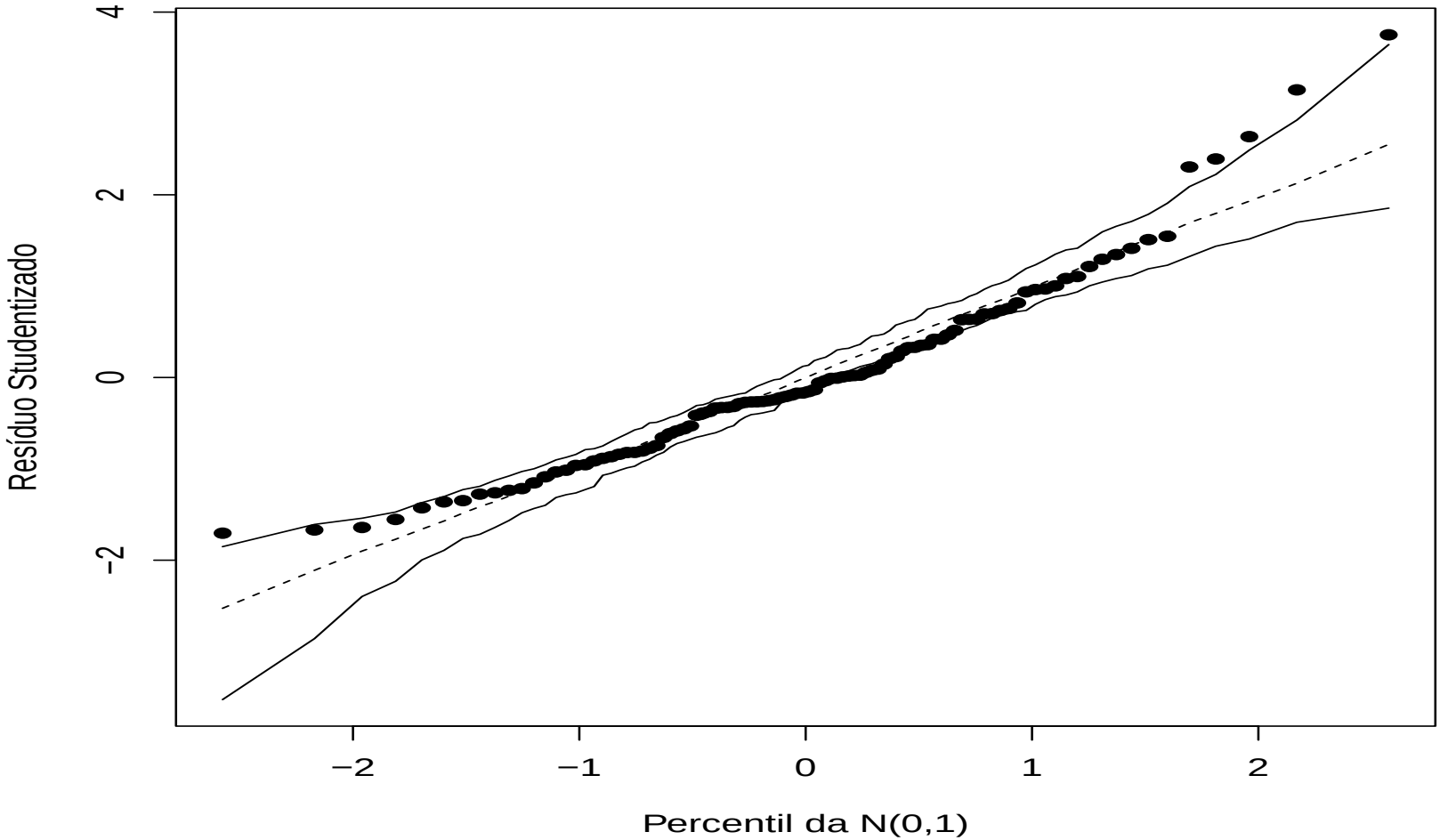








Modelo 4: envelope

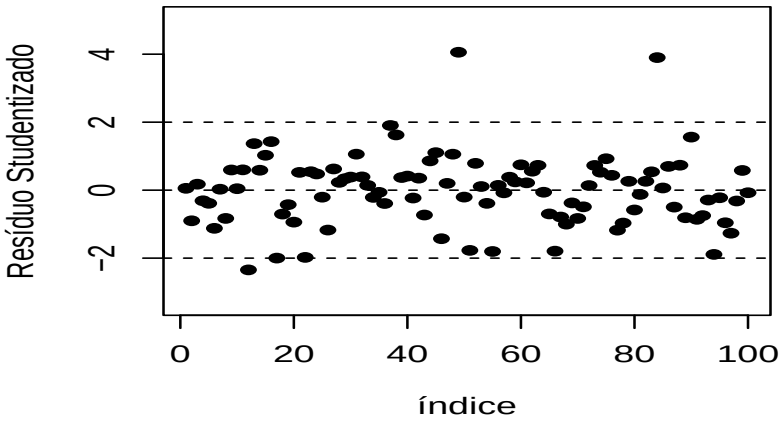


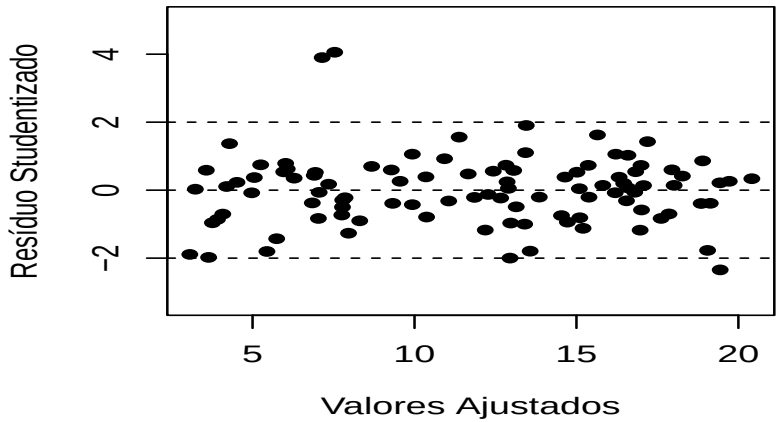
Ausência de normalidade

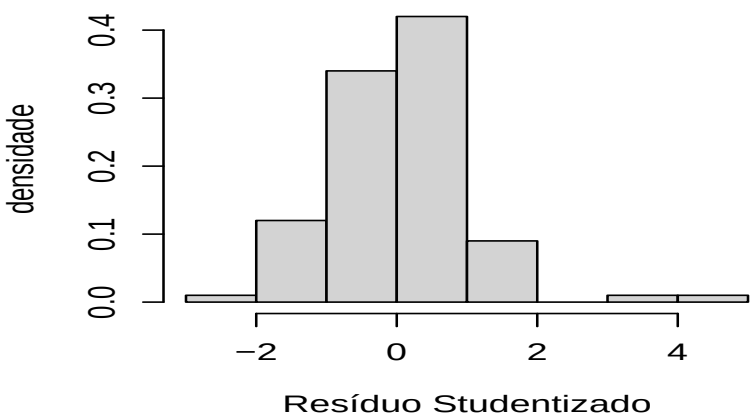
- Modelo 1 (M1): $Y_i = 1 + 2x_i + \xi_i$, $i = 1, 2, ..., 100$, $x_i \overset{i.i.d.}{\sim} U(1, 10)$ e $\xi \overset{i.i.d.}{\sim} N(0, 4)$.
- Modelo 2 (M2): M1 com $\xi_i \overset{ind.}{\sim} t_{(4)}$ (caudas pesadas).
- Modelo 3 (M3): M1 com $\xi_i \overset{ind.}{\sim} NA(0, 2, 20)$ (assimetria positiva).
- Modelo 4 (M4): M1 com $\xi_i \overset{ind.}{\sim} NA(0, 2, -20)$ (assimetria negativa).
- OBS: $NA(\mu, \psi, \lambda)$ representa uma distribuição normal assimétrica (na parametrização usual) com parâmetro de localização μ, de dispersão ψ e de assimetria λ.

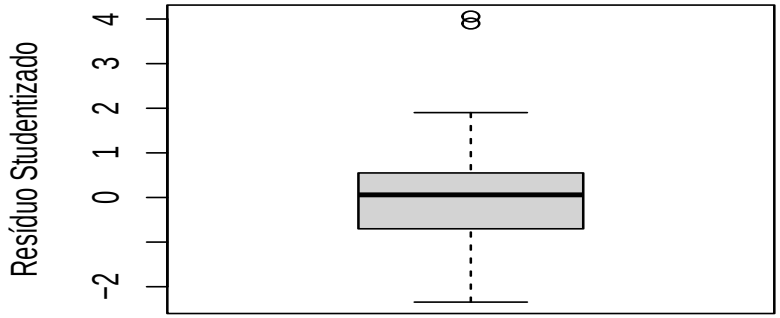
link

Modelo 1: diagnóstico

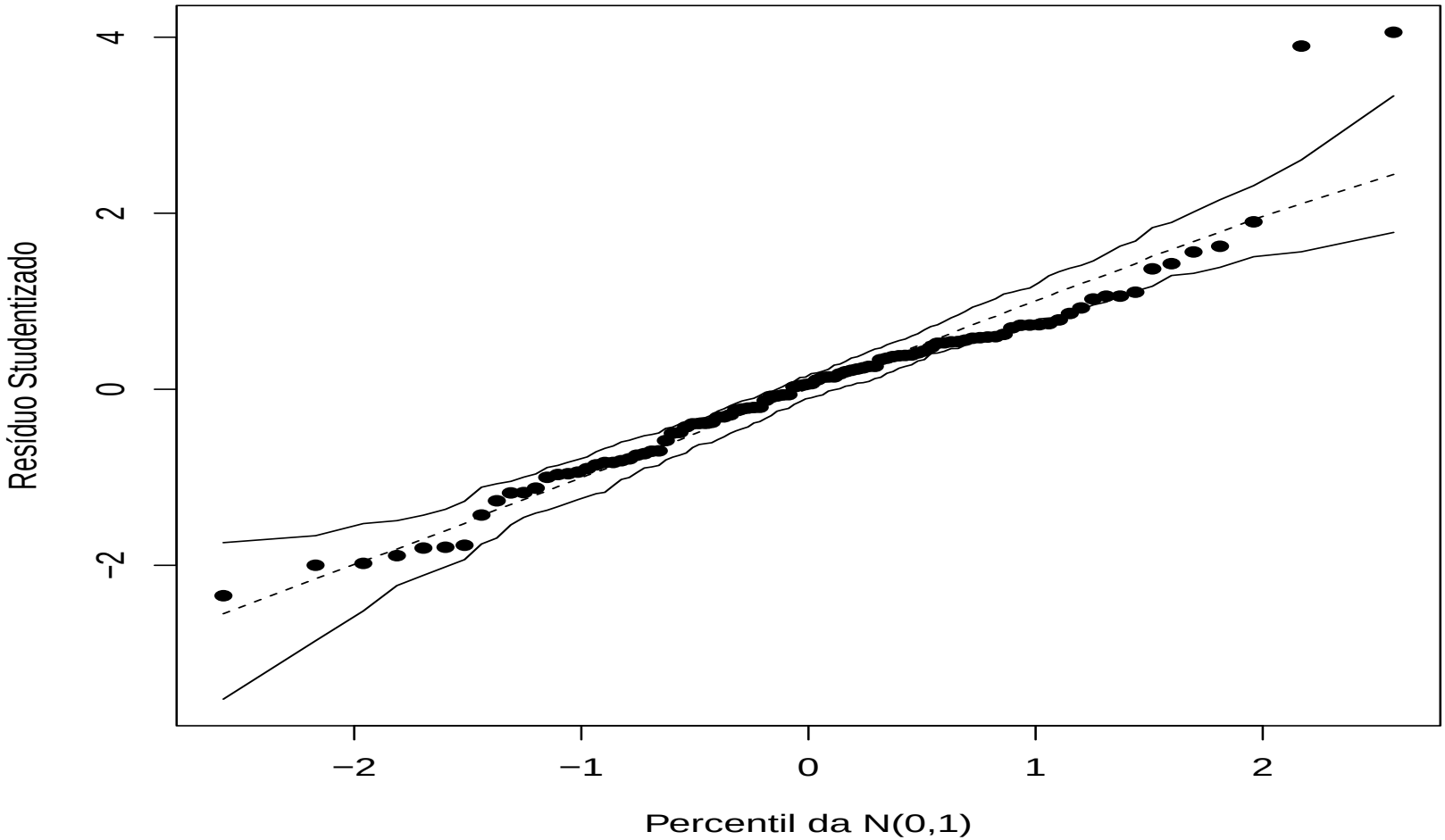




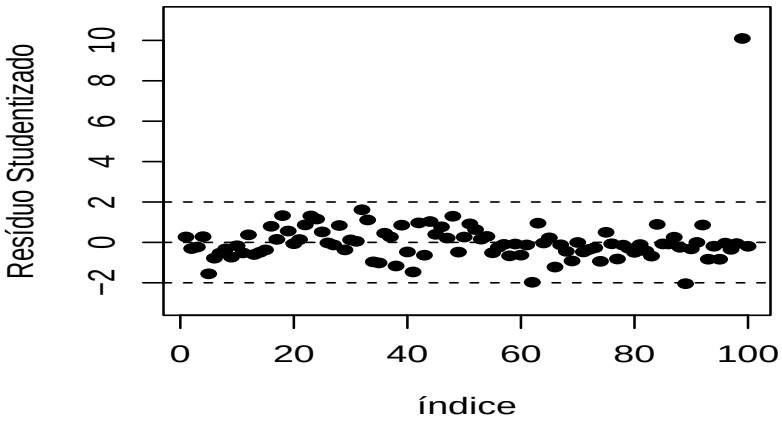


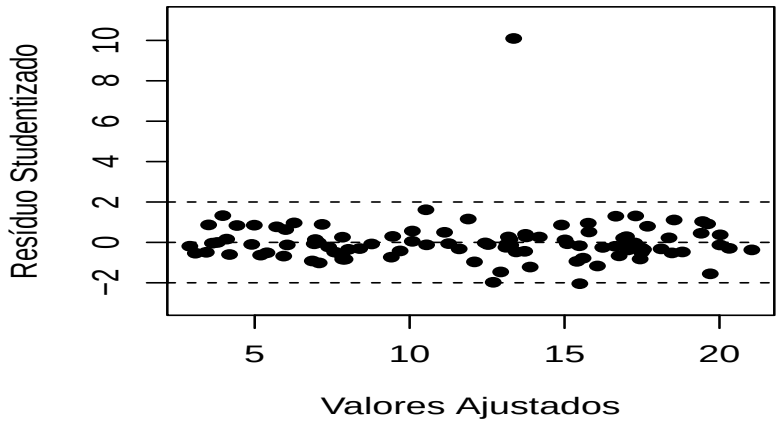


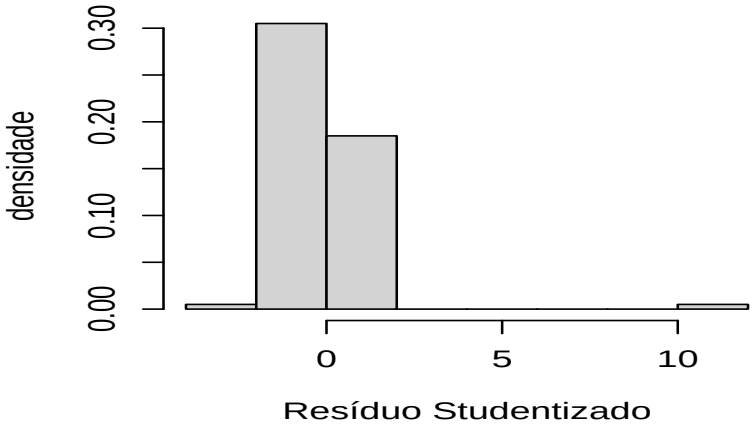
Modelo 1: envelope

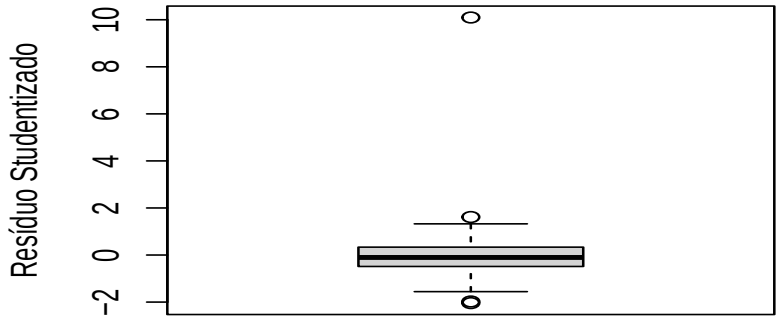


Modelo 2: diagnóstico

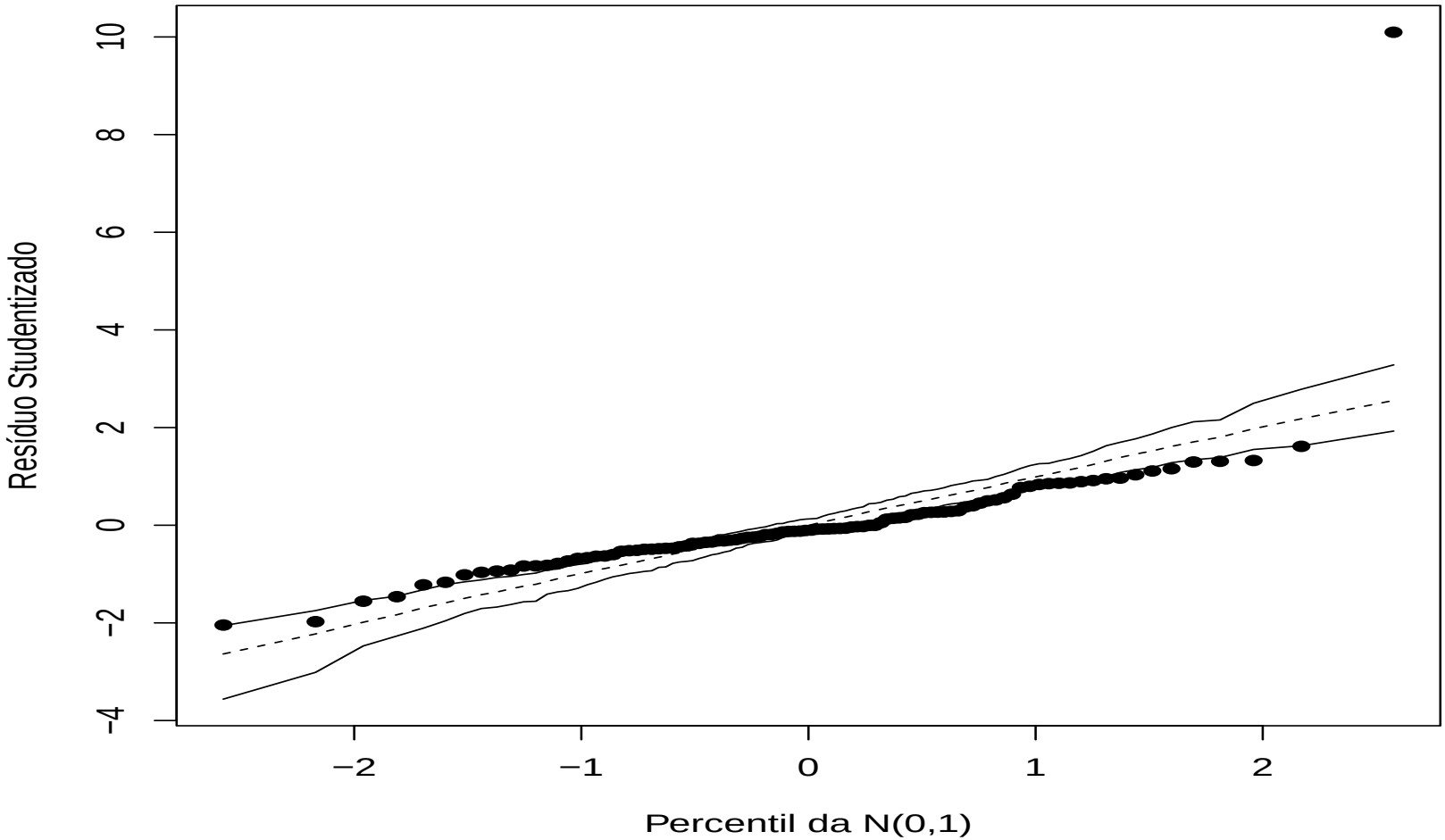




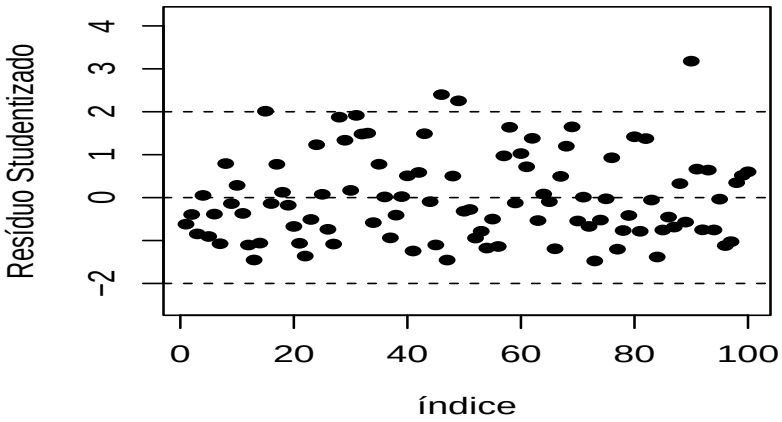


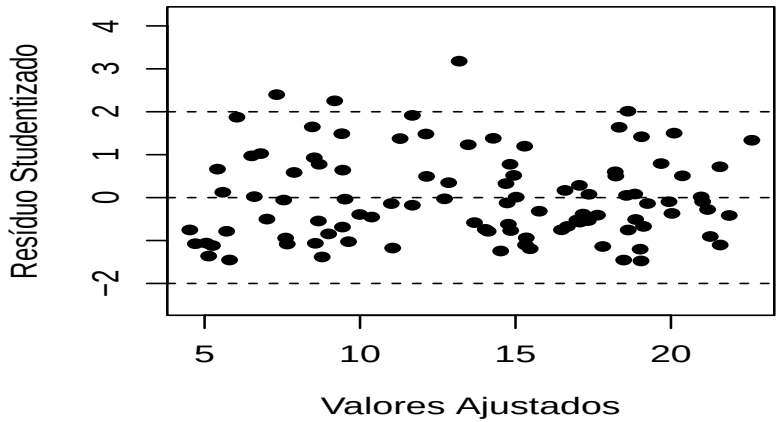


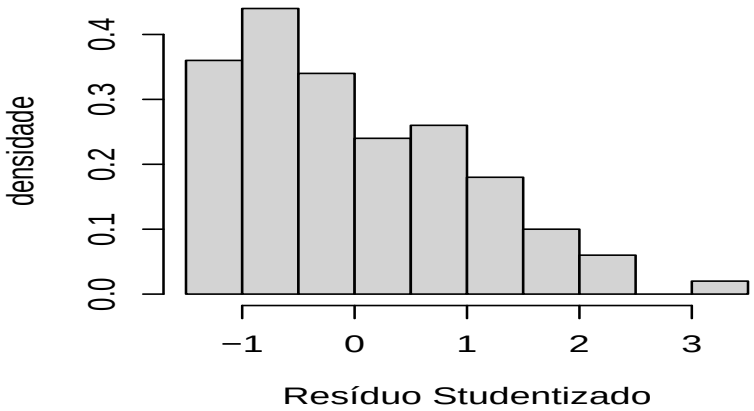
Modelo 2: envelope

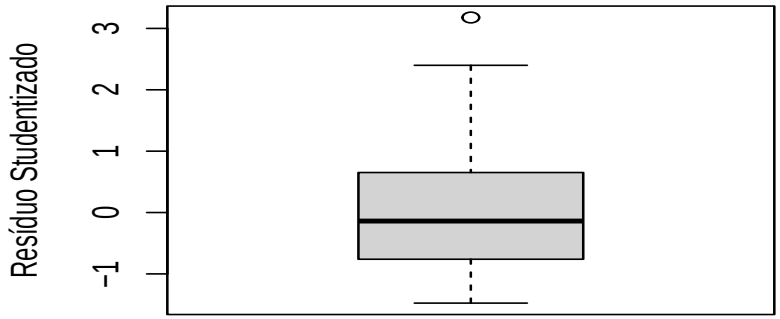


Modelo 3: diagnóstico

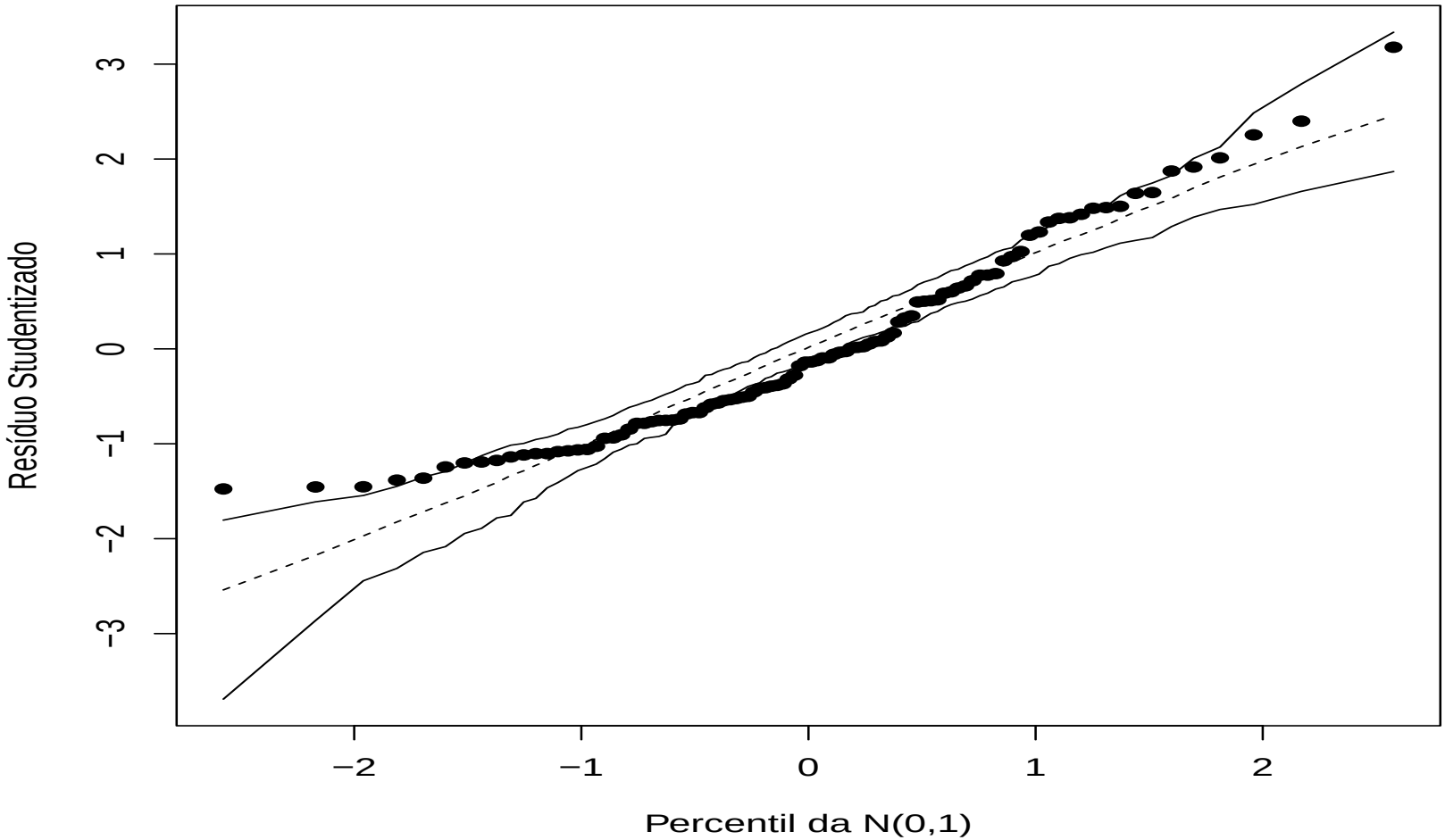




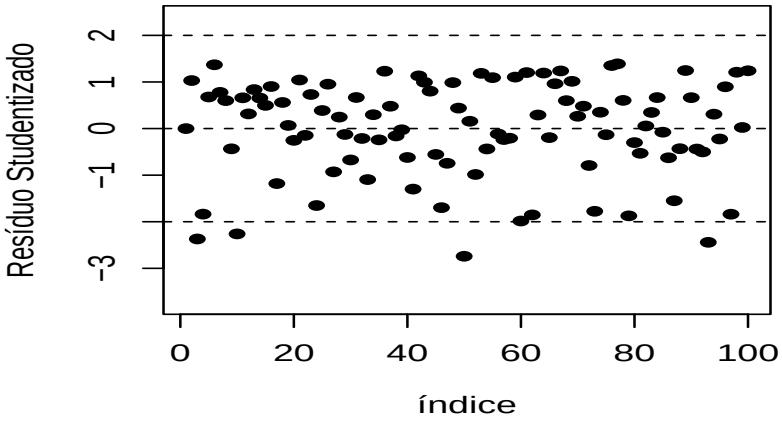


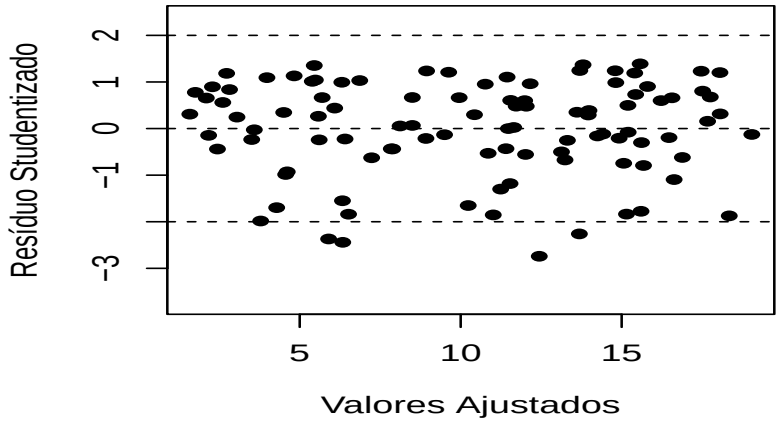


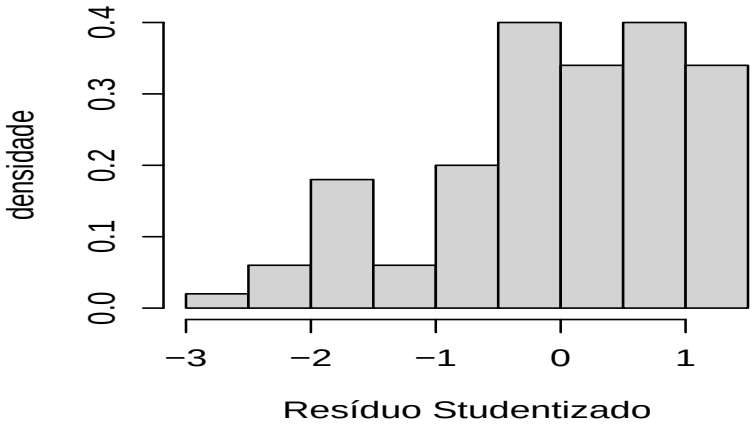
Modelo 3: envelope

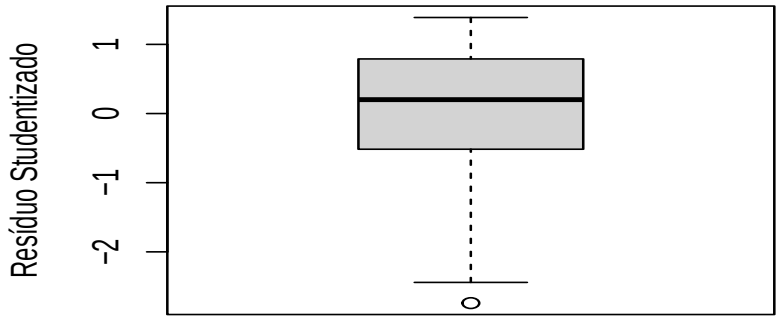


Modelo 4: diagnóstico









Modelo 4: envelope