

Análise de dados e métodos de diagnóstico em modelos de regressão normais lineares (parte 1)

Prof. Caio Azevedo

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

[http : //www.ime.usp.br/~giapaula/texto_2013.pdf](http://www.ime.usp.br/~giapaula/texto_2013.pdf)

Forma matricial do MRNL

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

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

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

Suposições

- As principais suposições do MNL são:
 - Homocedasticidade (dos erros).
 - Independência (correlação nula) dos erros.
 - Normalidade dos erros.
- Como verificar as suposições 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 (ξ) não são observados (observáveis), precisamos de algum preditor apropriado para avaliar as suposições feitas sobre eles.

Resíduos

- Como os erros (ξ) não são observados (observáveis), precisamos de algum preditor apropriado para avaliar as suposições feitas sobre eles.
- Lembre-se de que $\xi \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ (não correlacionados).

Resíduos

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

Resíduos

- Como os erros (ξ) não são observados (observáveis), precisamos de algum preditor apropriado para avaliar as suposições feitas sobre eles.
- Lembre-se de que $\xi \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ (não correlacionados).
- Já definimos os resíduos: $\hat{\xi}_i = R_i = Y_i - \hat{Y}_i = Y_i - \mathbf{X}'_i \hat{\beta}$.
- Matricialmente $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$,
 $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

Resíduos

- Como os erros (ξ) não são observados (observáveis), precisamos de algum preditor apropriado para avaliar as suposições feitas sobre eles.
- Lembre-se de que $\xi \sim N_n(\mathbf{0}, \sigma^2 \mathbf{I}_n)$ (não correlacionados).
- Já definimos os resíduos: $\hat{\xi}_i = R_i = Y_i - \hat{Y}_i = Y_i - \mathbf{X}'_i \hat{\beta}$.
- Matricialmente $\mathbf{R} = \mathbf{Y} - \hat{\mathbf{Y}} = \mathbf{Y} - \mathbf{X}\hat{\beta} = (\mathbf{I}_n - \mathbf{H})\mathbf{Y}$,
 $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.
- Assim, temos que, sob as suposições do modelo,
 $\mathbf{R} \sim N_n(\mathbf{0}, \sigma^2(\mathbf{I}_n - \mathbf{H}))$ (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 \mathbf{H} .

Cont. (livro do Prof. Gilberto, pág. 48, 49 e 50)

- Defina $V_i = \frac{R_i}{\sqrt{S^2(1-h_{ii})}}$, em que $S^2 = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.

Cont. (livro do Prof. Gilberto, pág. 48, 49 e 50)

- Defina $V_i = \frac{R_i}{\sqrt{S^2(1-h_{ii})}}$, em que $S^2 = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.
- A divisão por $(1 - h_{ii})$ atenua a correlação entre os resíduos.

Cont. (livro do Prof. Gilberto, pág. 48, 49 e 50)

- Defina $V_i = \frac{R_i}{\sqrt{S^2(1-h_{ii})}}$, em que $S^2 = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.
- A divisão por $(1 - h_{ii})$ atenua a correlação entre os resíduos.
- Contudo, R_i e S^2 não são independentes (exercício).

Cont. (livro do Prof. Gilberto, pág. 48, 49 e 50)

- Defina $V_i = \frac{R_i}{\sqrt{S^2(1-h_{ii})}}$, em que $S^2 = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})' (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})$.
- A divisão por $(1 - h_{ii})$ atenua a correlação entre os resíduos.
- Contudo, R_i e S^2 não são independentes (exercício).
- Porém, $S_{(i)}^2$ e R_i o são (em que $S_{(i)}^2$ corresponde à S^2 obtido no modelo sem a i -ésima observação).

Cont. (livro do Prof. Gilberto, pág. 48, 49 e 50)

- Defina $V_i = \frac{R_i}{\sqrt{S^2(1-h_{ii})}}$, em que $S^2 = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\beta})' (\mathbf{Y} - \mathbf{X}\hat{\beta})$.
- A divisão por $(1 - h_{ii})$ atenua a correlação entre os resíduos.
- Contudo, R_i e S^2 não são independentes (exercício).
- Porém, $S_{(i)}^2$ e R_i o são (em que $S_{(i)}^2$ corresponde à S^2 obtido no modelo sem a i -ésima observação).
- Pode-se provar, além disso, que $S_{(i)}^2 = S^2 \left(\frac{n-p-V_i^2}{n-p-1} \right)$.

Cont. (livro do Prof. Gilberto, pág. 48, 49 e 50)

- Defina $V_i = \frac{R_i}{\sqrt{S^2(1-h_{ii})}}$, em que $S^2 = \frac{1}{n-p} (\mathbf{Y} - \mathbf{X}\hat{\beta})' (\mathbf{Y} - \mathbf{X}\hat{\beta})$.
- A divisão por $(1 - h_{ii})$ atenua a correlação entre os resíduos.
- Contudo, R_i e S^2 não são independentes (exercício).
- Porém, $S_{(i)}^2$ e R_i o são (em que $S_{(i)}^2$ corresponde à S^2 obtido no modelo sem a i -ésima observação).
- Pode-se provar, além disso, que $S_{(i)}^2 = S^2 \left(\frac{n-p-V_i^2}{n-p-1} \right)$.
- Tem-se, então, que $T_i = \frac{R_i}{\sqrt{S_{(i)}^2(1-h_{ii})}} \sim t_{(n-p-1)}$, sob a validade das hipóteses do modelo (exercício). Lembre-se de que, se $\nu \geq 30$, então $t_{(\nu)} \approx N(0, 1)$.

O que e como observar nos resíduos?

- Gráfico de dispersão dos resíduos versus o índice da observação: ausência de dependência/tendência/correlação.

O que e como observar nos resíduos?

- Gráfico de dispersão dos resíduos versus o índice da observação: ausência de dependência/tendência/correlação.
- Gráfico de dispersão dos resíduos versus os valores ajustados: homocedasticidade.

O que e como observar nos resíduos?

- Gráfico de dispersão dos resíduos versus o índice da observação: ausência de dependência/tendência/correlação.
- Gráfico de dispersão dos resíduos versus os valores ajustados: homocedasticidade.
- Boxplot e/ou gráfico de quantis-quantis: simetria, ausência de “out-liers” e multimodalidade.

O que e como observar nos resíduos?

- Gráfico de dispersão dos resíduos versus o índice da observação: ausência de dependência/tendência/correlação.
- Gráfico de dispersão dos resíduos versus os valores ajustados: homocedasticidade.
- Boxplot e/ou gráfico de quantis-quantis: simetria, ausência de “out-liers” e multimodalidade.
- Problema no gráfico de quantis-quantis: Visualmente, muitas vezes, é complicado avaliar a proximidade dos quantis.

O que e como observar nos resíduos?

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

Procedimento para se gerar o gráfico de envelopes

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

Procedimento para se gerar o gráfico de envelopes

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

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

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):

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

- 6) Obtem-se os limites $t_{(i)l}^* = \min_{1 \leq j \leq m} t_{(i)j}^*$ e $t_{(i)s}^* = \max_{1 \leq j \leq m} t_{(i)j}^*$, $j = 1, 2, \dots, m$.

Procedimento para se gerar o gráfico de envelopes

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

- Além disso, consideramos como a linha de referência

$$t_{(i)}^* = \frac{1}{m} \sum_{j=1}^m t_{(i)j}^*, i = 1, 2, \dots, n.$$

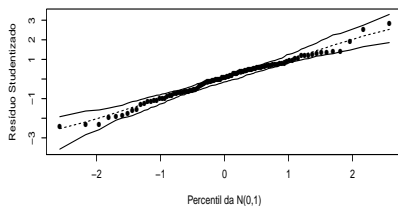
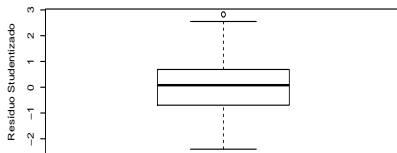
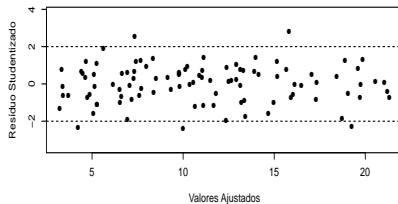
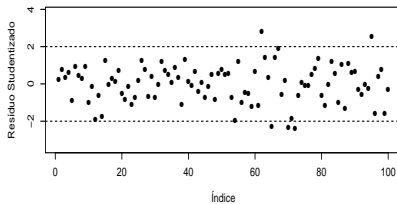
Estudo de simulação

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

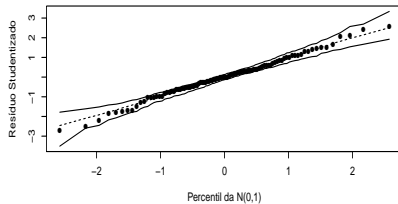
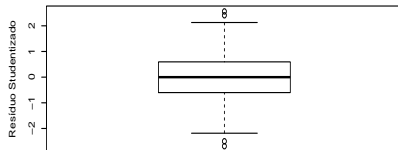
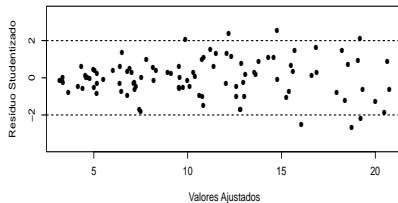
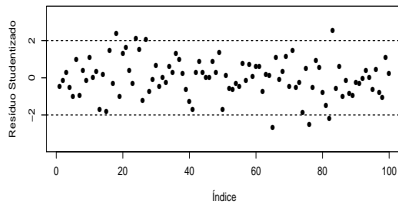
Heterocedasticidade

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

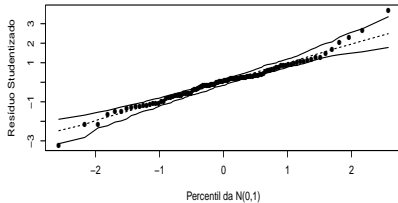
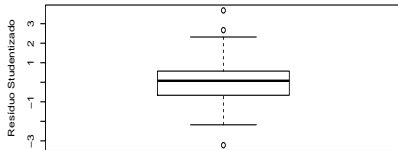
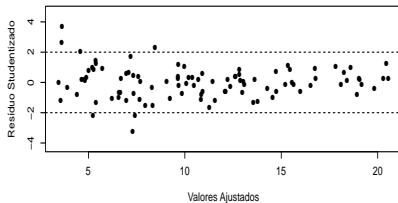
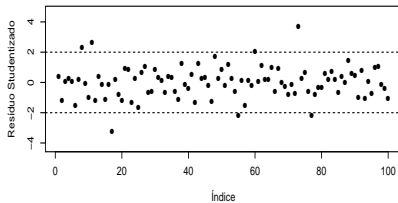
Modelo 1



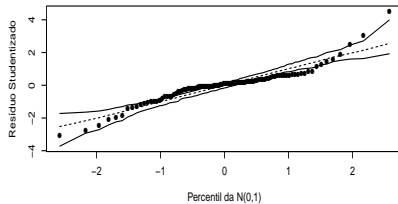
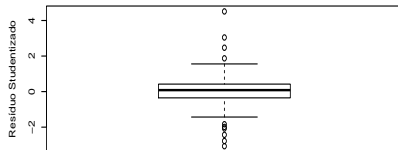
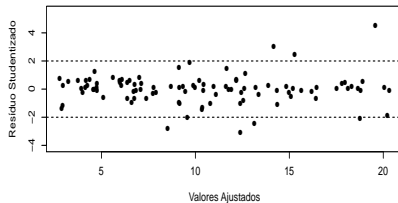
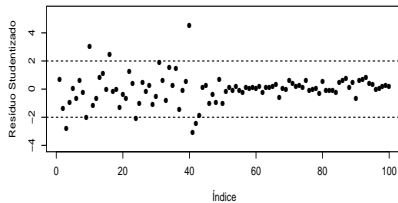
Modelo 2



Modelo 3



Modelo 4



Dependência

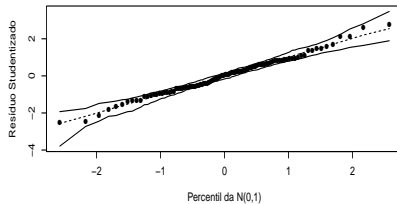
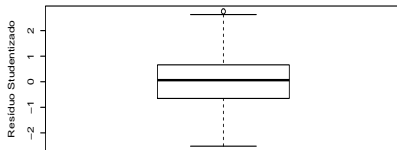
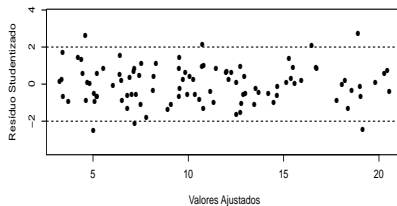
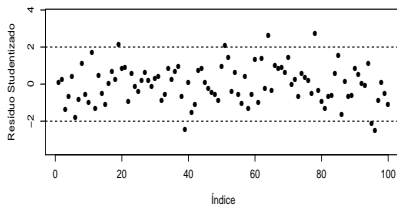
- Modelo 1 (M1): $Y_i = 1 + 2x_i + \xi_i, i = 1, 2, \dots, 100, x_i \stackrel{i.i.d.}{\sim} U(1, 10)$ e $\xi \stackrel{i.i.d.}{\sim} N(0, 4)$.
- Modelo 2 (M2): $Y_i, i = 1, \dots, 100$ segue um processo AR(1) com $\rho = 0,8$ (correlação entre as observações).
- Modelo 3 (M3): M1 com $(\xi_i, \xi_{i+1})' \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 3.2 \\ 3.2 & 4 \end{bmatrix} \right), i=1, \dots, 99$.

Dependência

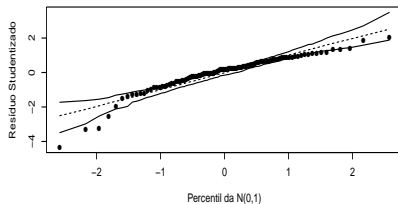
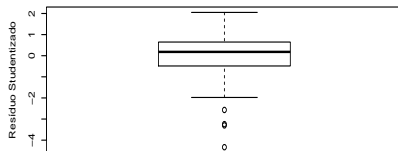
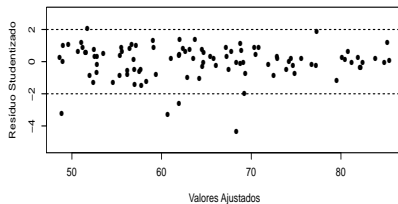
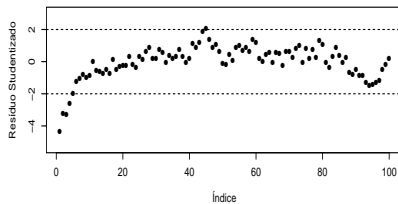
- Modelo 4 (M4): M1 com $\xi_1 \sim N_{50}(\mathbf{0}, \Sigma_1)$ e $\xi_2 \sim N_{50}(\mathbf{0}, \Sigma_2)$, em que $\xi_1 = (\xi_1, \dots, \xi_{50})'$ e $\xi_2 = (\xi_{51}, \dots, \xi_{100})'$.

$$\Sigma_1 = \begin{bmatrix} 4 & 3.2 & \dots & 3.2 \\ 3.2 & 4 & \dots & 3.2 \\ \vdots & \vdots & \ddots & \vdots \\ 3.2 & 3.2 & \dots & 4 \end{bmatrix}; \Sigma_2 = \begin{bmatrix} 4 & 3.6 & \dots & 3.6 \\ 3.6 & 4 & \dots & 3.6 \\ \vdots & \vdots & \ddots & \vdots \\ 3.6 & 3.6 & \dots & 4 \end{bmatrix}$$

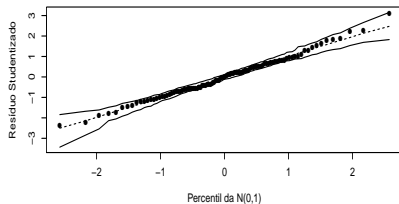
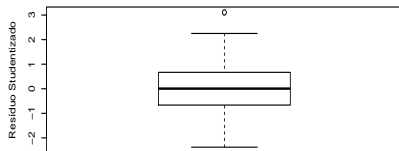
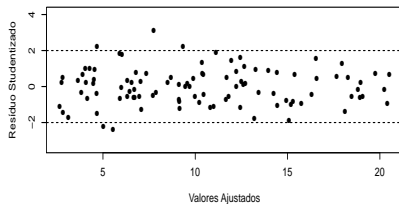
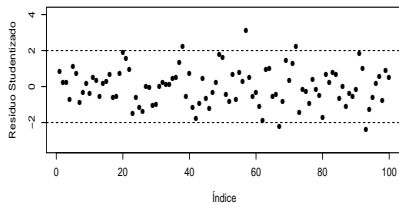
Modelo 1



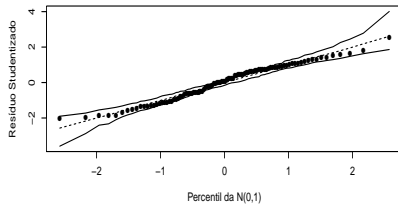
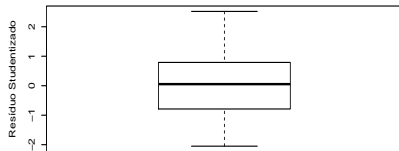
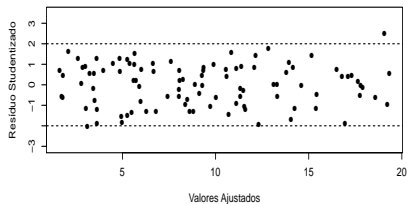
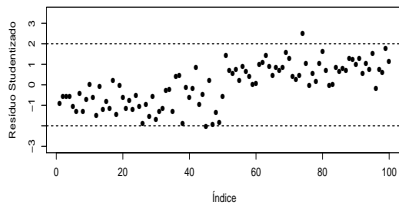
Modelo 2



Modelo 3



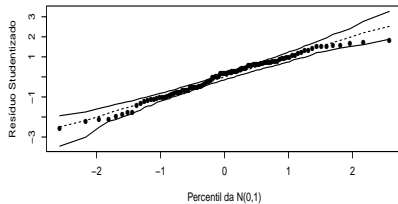
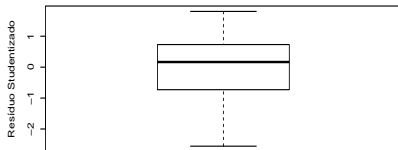
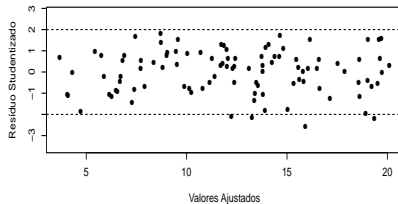
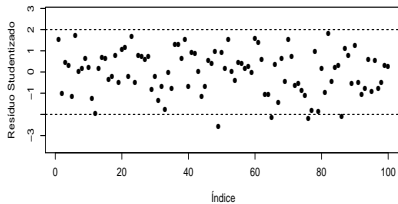
Modelo 4



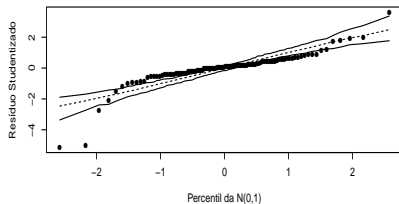
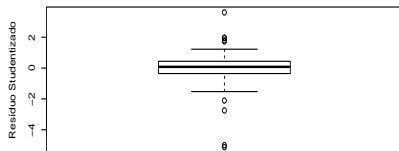
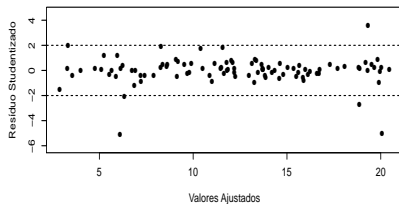
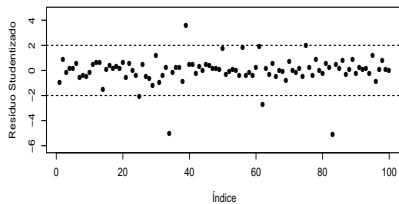
Ausência de normalidade

- Modelo 1 (M1): $Y_i = 1 + 2x_i + \xi_i, i = 1, 2, \dots, 100, x_i \stackrel{i.i.d.}{\sim} U(1, 10)$ e $\xi \stackrel{i.i.d.}{\sim} N(0, 4)$.
- Modelo 2 (M2): M1 com $\xi_i \stackrel{ind.}{\sim} t_{(4)}$ (caudas pesadas).
- Modelo 3 (M3): M1 com $\xi_i \stackrel{ind.}{\sim} NA(0, 2, 20)$ (assimetria positiva).
- Modelo 4 (M4): M1 com $\xi_i \stackrel{ind.}{\sim} NA(0, 2, -20)$ (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 .

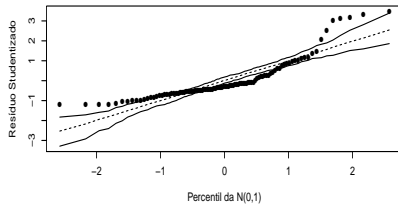
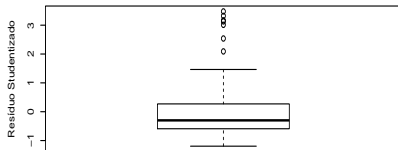
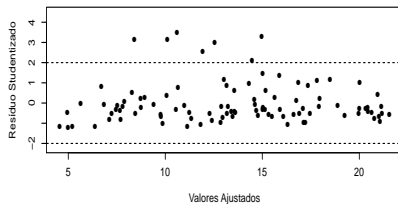
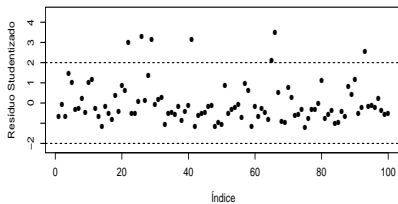
Modelo 1



Modelo 2



Modelo 3



Modelo 4

