Análise de resíduos em modelos hierárquicos/multiníveis

Prof. Caio Azevedo (grande parte do material apresentado foi extraído do livro Modelos de regressão com apoio computacional do Prof. Gilberto A. Paula
link e da Dissertação de Mestrado do Prof. Juvêncio Nobre: link)

Forma matricial do MRNL

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

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

- Suposição: $\boldsymbol{\xi} \sim N_n(0, \sigma^2 \boldsymbol{I}_n)$.
- $\boldsymbol{Y}$ 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 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.
- 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 $\boldsymbol{\xi} \sim N_n(0, \sigma^2 \boldsymbol{I}_n)$ (não correlacionados).
- Podemos considerar: $\widehat{\xi}_i = R_i = Y_i - \widehat{Y}_i = Y_i - \boldsymbol{X}_i' \widehat{\beta}$ (resíduo ordinário).
- Referência : Cox and Snell (1968)

Resíduos

- Matricialmente $\boldsymbol{R} = \boldsymbol{Y} - \widehat{\boldsymbol{Y}} = \boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ (resíduo ordinário).
- Assim, temos que, sob as suposições do modelo, $\boldsymbol{R} \sim N_n(0, \sigma^2(\boldsymbol{I}_n - \boldsymbol{H}))$ (são correlacionados). Mais especficamente, $r_i \sim N(0, \sigma^2(1 - h_{ii}))$ 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 $S^2 = \frac{1}{n-p}\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)'\left(\boldsymbol{Y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)$ (resíduo padronizado).
- A divisão por $(1 - h_{ii})$ atenua a correlação entre os resíduos.

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

- 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).
- Pode-se provar, além disso, que $S^2_{(i)} = S^2 \left(\frac{n-p-V_i^2}{n-p-1} \right)$.
- 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 (exercício). Lembre-se de que, se $\nu \geq 30$, então $t_{(\nu)} \approx N(0,1)$ (resíduo studentizado).

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 outliers e pesos das caudas.
- Histograma: simetria, ausência 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 $t_i^* = \frac{r_i^*}{\sqrt{1-h_{ii}}}$.
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).

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

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 $t^*_{(i)I} = \frac{t^*_{(i)(2)}+t^*_{(i)(3)}}{2}$ 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, $i = 1, 2, ..., n$.

- Além disso, consideramos como a linha de referência $t^*_{(i)} = \frac{1}{m}\sum_{j=1}^{m} t^*_{(i)j}, i = 1, 2, ..., n$.

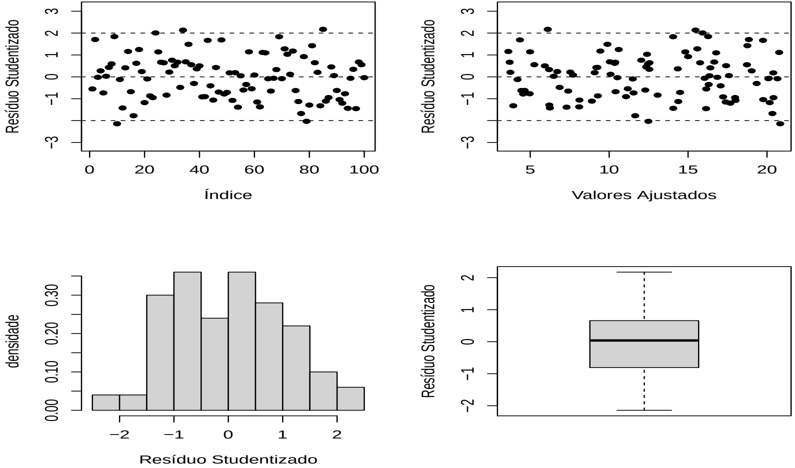
Estudo de simulação

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

Heterocedasticidade

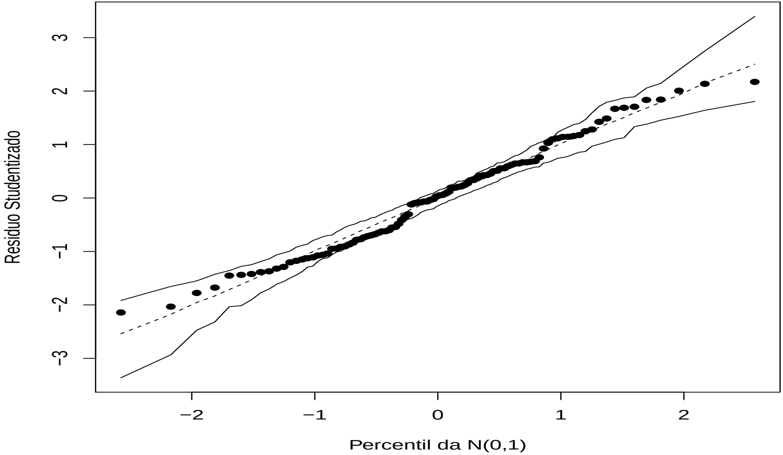
- Modelo 1 (M1): $Y_i = 1 + 2x_i + \xi_i, i = 1, 2, ..., 100,\ x_i \overset{i.i.d.}{\sim} U(1, 10)$ e $\xi_i \overset{i.i.d.}{\sim} N(0, 4)$.
- Modelo 2 (M2): M1 com $\xi_i \overset{ind.}{\sim} N(0, 4x_i)$.
- Modelo 3 (M3): M1 com $\xi_i \overset{ind.}{\sim} N(0, 4x_i^{-1})$.
- Modelo 4 (M4): M1 com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

Modelo 1

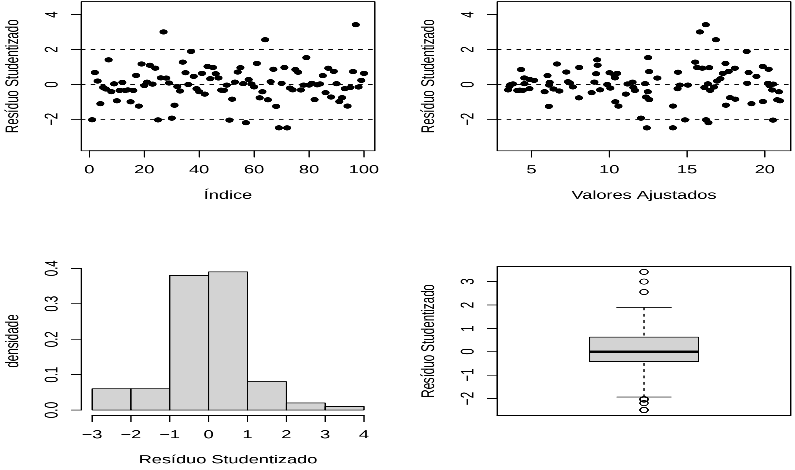


Resíduo Studentizado
Índice
Resíduo Studentizado
Valores Ajustados
densidade
Resíduo Studentizado
Resíduo Studentizado

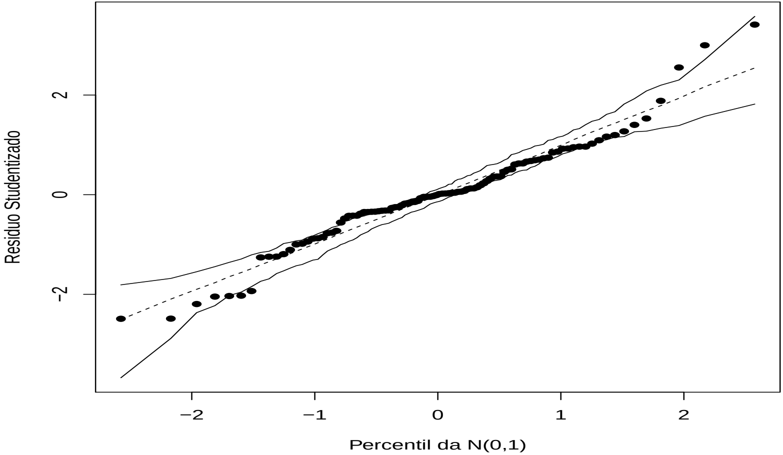
Modelo 1



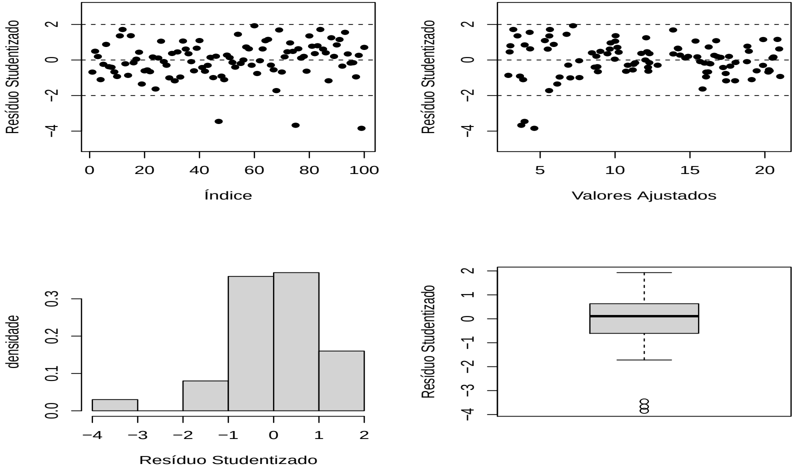
Modelo 2


Resíduo Studentizado
Índice
Valores Ajustados
densidade
Resíduo Studentizado
Resíduo Studentizado

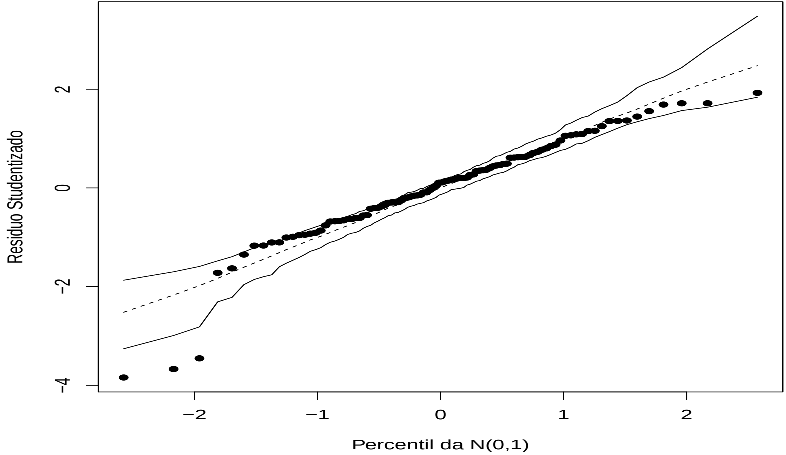
Modelo 2



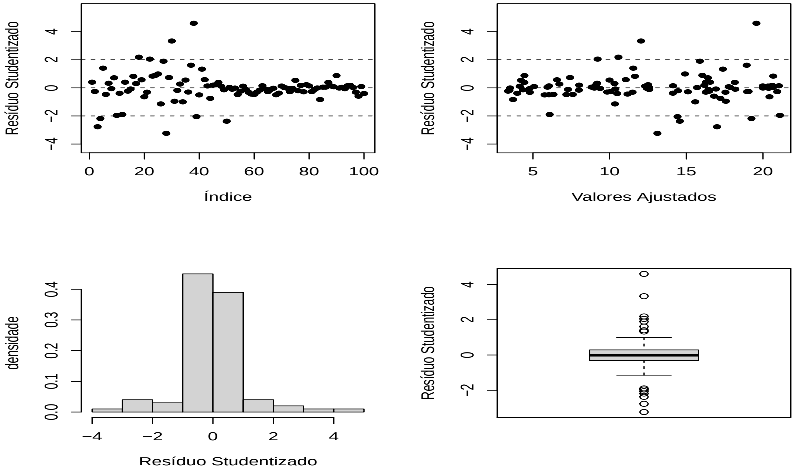
Modelo 3


Resíduo Studentizado
Índice
Resíduo Studentizado
Valores Ajustados
densidade
Resíduo Studentizado
Resíduo Studentizado

Modelo 3

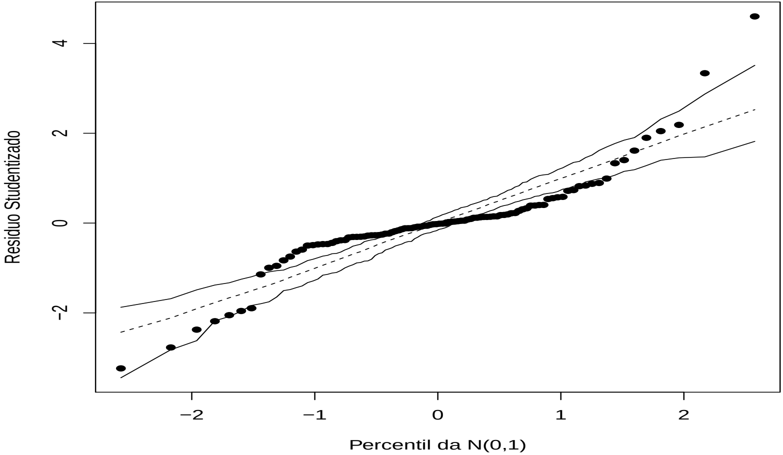


Modelo 4



Resíduo Studentizado
Índice
Valores Ajustados
densidade
Resíduo Studentizado

Modelo 4



Dependência

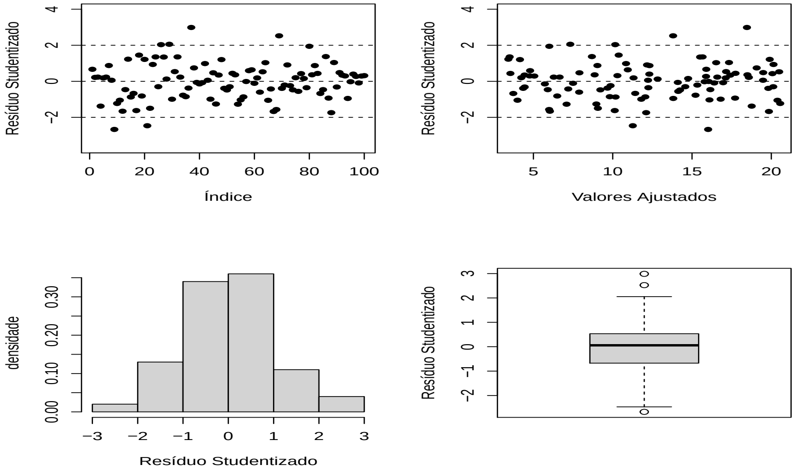
- Modelo 1 (M1): $Y_i = 1 + 2x_i + \xi_i, i = 1, 2, ..., 100$, $x_i \overset{i.i.d.}{\sim} U(1, 10)$ e $\xi_i \overset{i.i.d.}{\sim} N(0, 4)$.
- Modelo 2 (M2): $Y_i, i = 1, ..., 100$ segue um processo AR(1) com $\rho = 0,8$ (correlação entre as observações).
- Modelo 3 (M3): M1 com $(\xi_i, \xi_{i+1})' \sim N_2 \left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 4 & 3.2 \\ 3.2 & 4 \end{bmatrix} \right)$, i=1, 3, 5, 7...,99.

Dependência

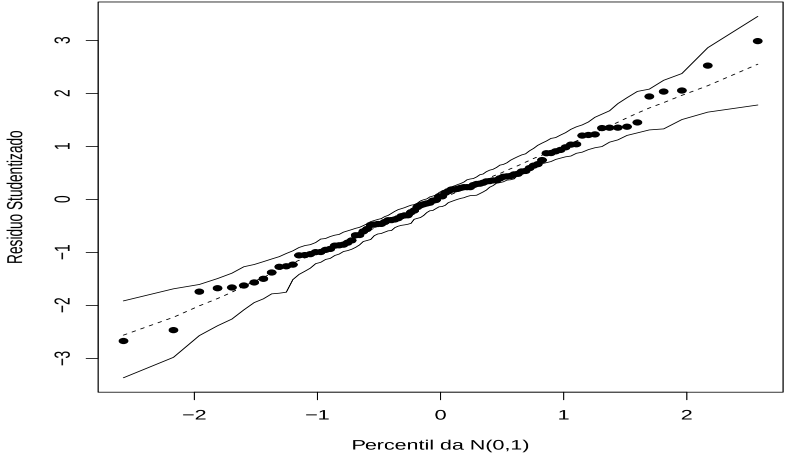
- Modelo 4 (M4): M1 com $\boldsymbol{\xi}_1 \sim N_{50}(0, \boldsymbol{\Sigma}_1)$ e $\boldsymbol{\xi}_2 \sim N_{50}(0, \boldsymbol{\Sigma}_2)$, em que $\boldsymbol{\xi}_1 = (\xi_1, ..., \xi_{50})'$ e $\boldsymbol{\xi}_2 = (\xi_{51}, ..., \xi_{100})'$.

$$\boldsymbol{\Sigma}_1 = \begin{bmatrix} 4 & 3.2 & \dots & 3.2 \\ 3.2 & 4 & \dots & 3.2 \\ \vdots & \vdots & \ddots & \vdots \\ 3.2 & 3.2 & \dots & 4 \end{bmatrix}; \boldsymbol{\Sigma}_2 = \begin{bmatrix} 4 & 3.6 & \dots & 3.6 \\ 3.6 & 4 & \dots & 3.6 \\ \vdots & \vdots & \ddots & \vdots \\ 3.6 & 3.6 & \dots & 4 \end{bmatrix}$$

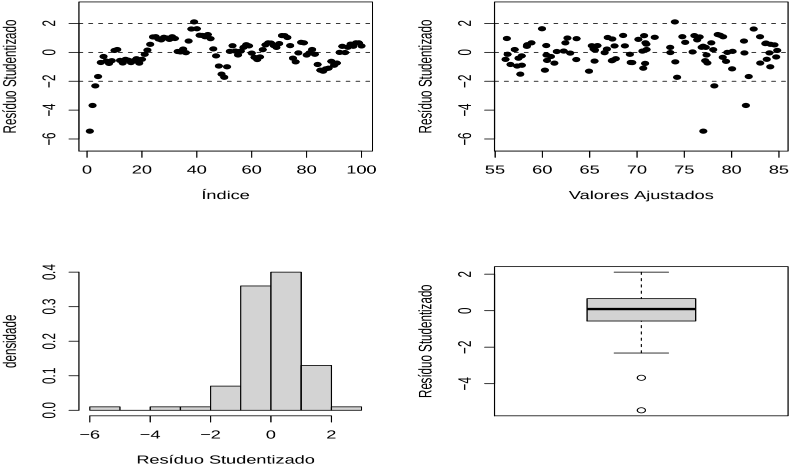
Modelo 1


Resíduo Studentizado
Índice
Valores Ajustados
densidade
Resíduo Studentizado

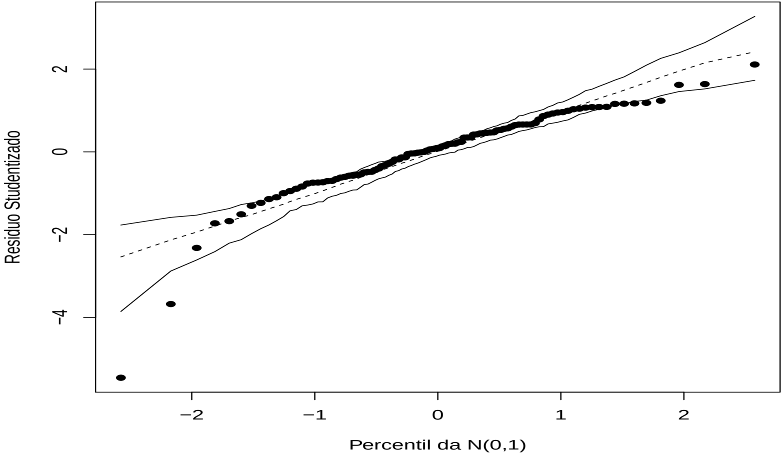
Modelo 1



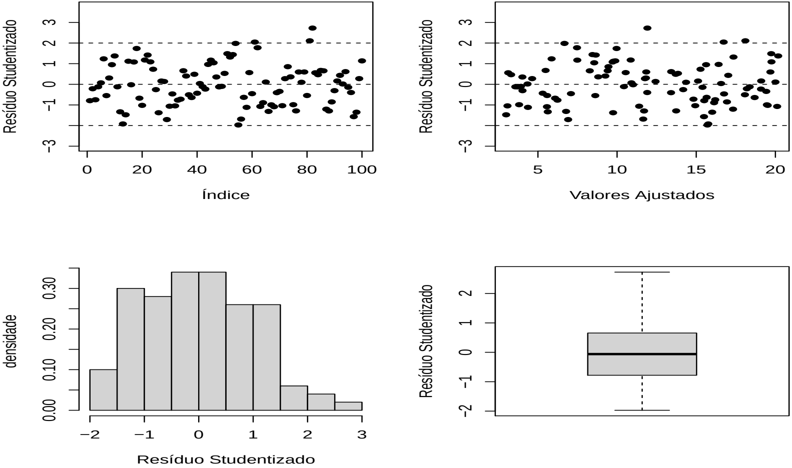
Modelo 2


Resíduo Studentizado
Índice
Valores Ajustados
densidade
Resíduo Studentizado

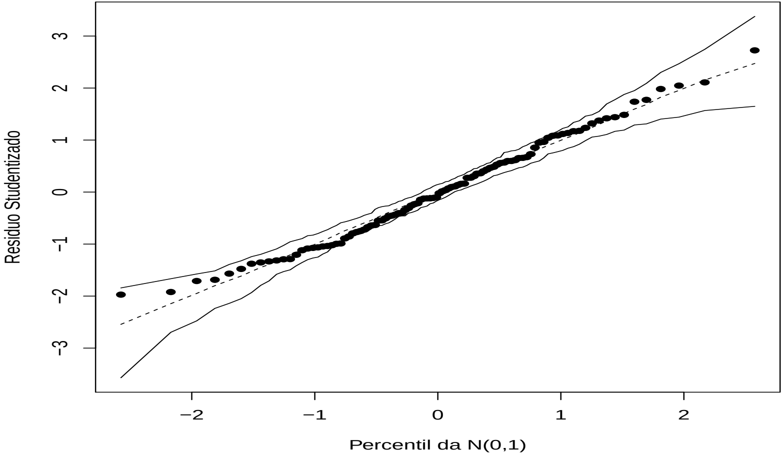
Modelo 2



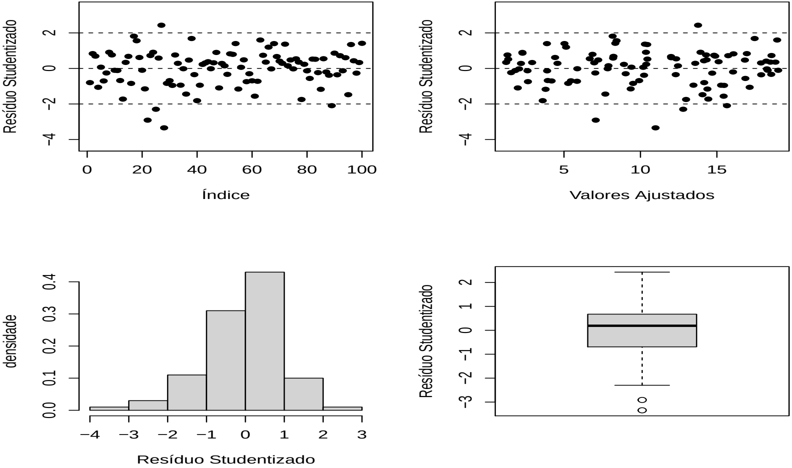
Modelo 3


Resíduo Studentizado
Índice
Resíduo Studentizado
Valores Ajustados
densidade
Resíduo Studentizado
Resíduo Studentizado

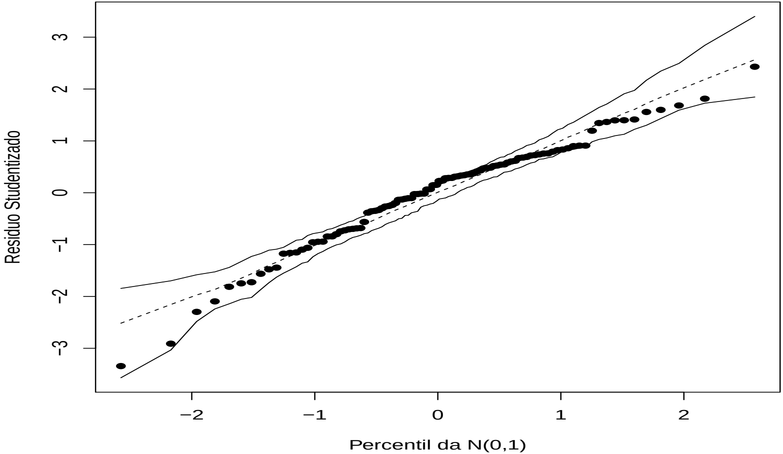
Modelo 3



Modelo 4


Resíduo Studentizado
Índice
Valores Ajustados
densidade
Resíduo Studentizado
0 20 40 60 80 100
5 10 15
−4 −3 −2 −1 0 1 2 3
0.0 0.1 0.2 0.3 0.4
-4 -2 0 2
-3 -2 -1 0 1 2

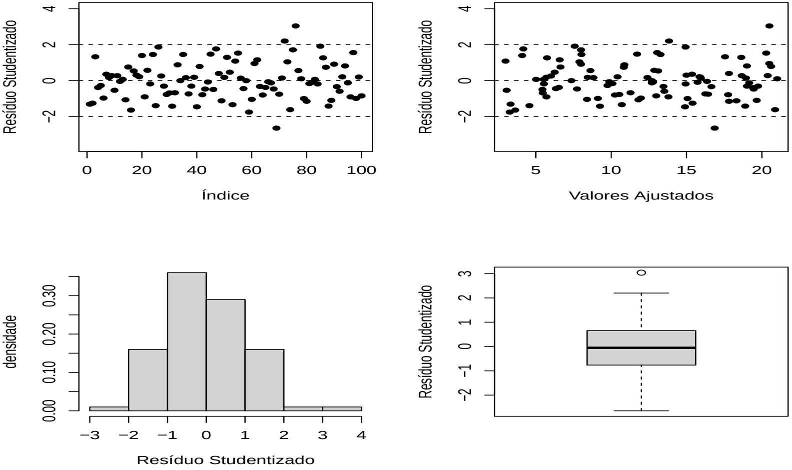
Modelo 4



Ausência de normalidade

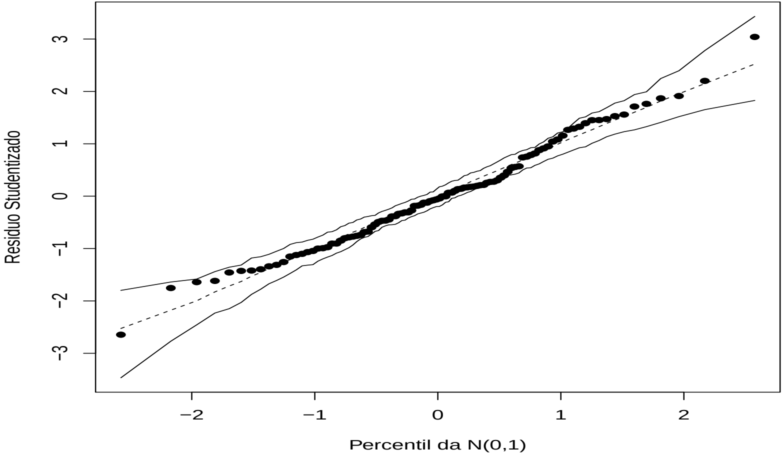
- Modelo 1 (M1): $Y_i = 1 + 2x_i + \xi_i, i = 1, 2, ..., 100$, $x_i \overset{i.i.d.}{\sim} U(1, 10)$ e $\xi_i \overset{i.i.d.}{\sim} N(0, 4)$.
- Modelo 2 (M2): M1 com $\xi_i \overset{ind.}{\sim} t_{(4)}$ (caudas pesadas).
- Modelo 3 (M3): M1 com $\xi_i \overset{ind.}{\sim} NA(0, 2, 20)$ (assimetria positiva).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

Modelo 1

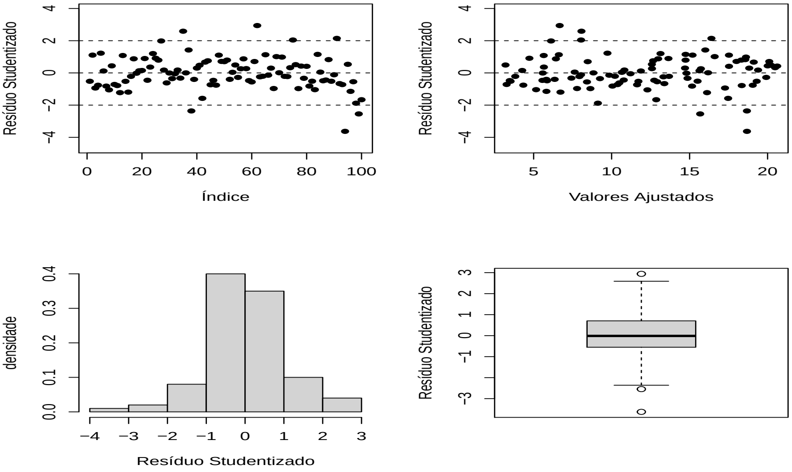


link

Modelo 1

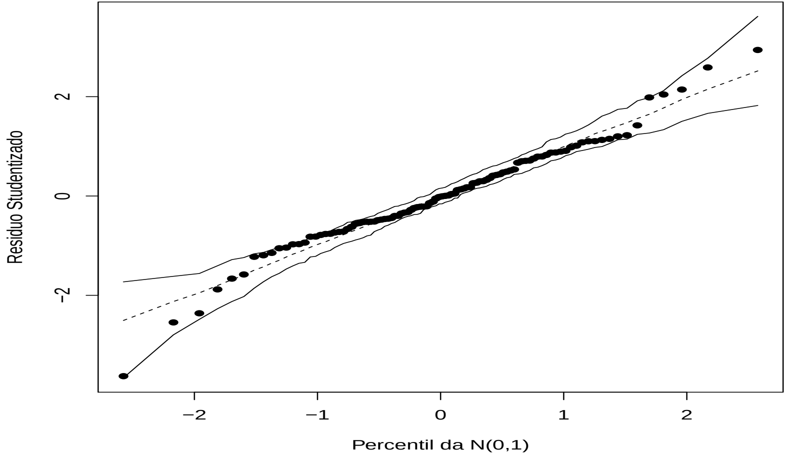


Modelo 2

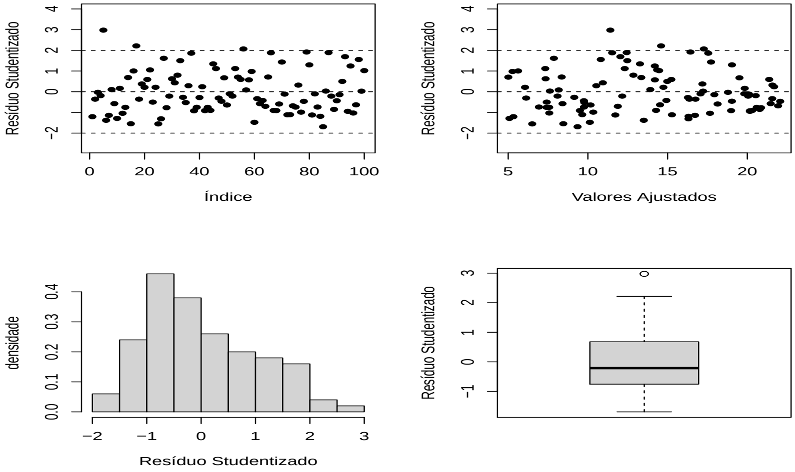


Resíduo Studentizado
Índice
Resíduo Studentizado
Valores Ajustados
densidade
Resíduo Studentizado
Resíduo Studentizado

Modelo 2

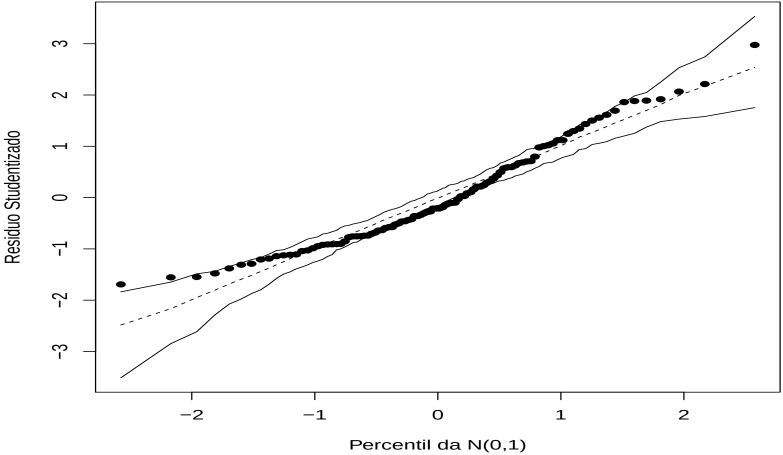


Modelo 3

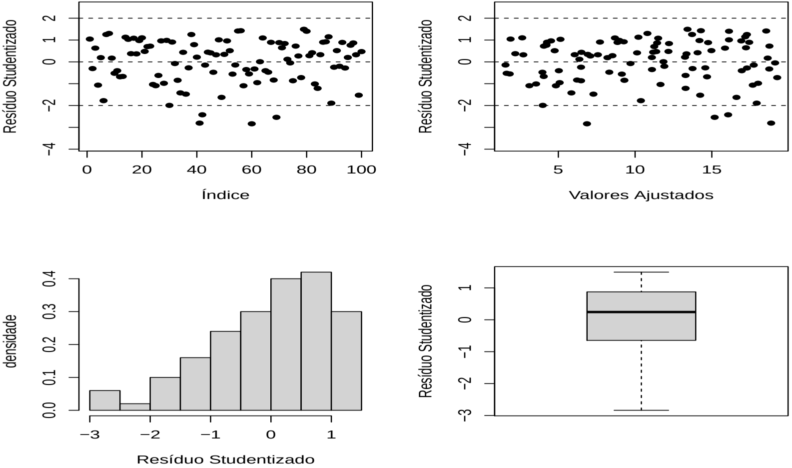


Resíduo Studentizado
Índice
Valores Ajustados
densidade
Resíduo Studentizado

Modelo 3

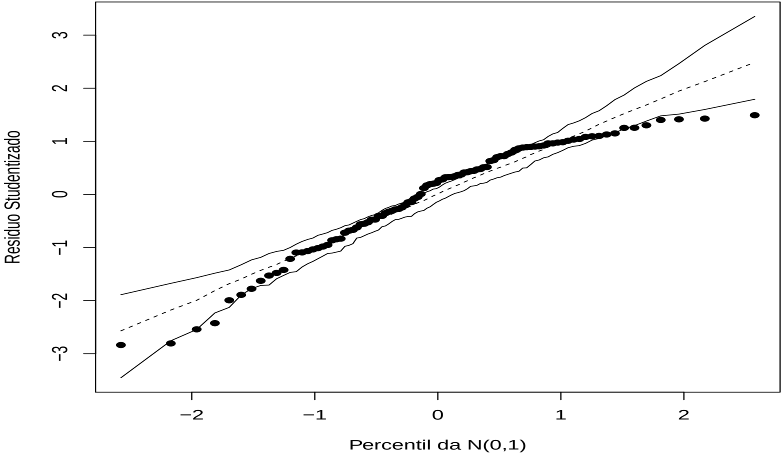


Modelo 4



Resíduo Studentizado
Índice
Resíduo Studentizado
Valores Ajustados
densidade
Resíduo Studentizado
Resíduo Studentizado

Modelo 4



Análise de resíduos para MH: Referências

- Nobre (2004). Métodos de diagnóstico para modelos lineares mistos. Dissertação de Mestrado. IME-USP.
- Nobre & Singer (2007). Residual analysis for linear mixed models. Biometrical Journal, 49, 6, 863-875.
- Veja também aqui e aqui.
- Lembremos que existe uma relação entre os modelos hierárquicos de dois níveis e os modelos mistos (link).

Análise de resíduos para modelos mistos

- Existem duas fontes de variação: os efeitos aleatórios $\boldsymbol{u}$ e os erros $\boldsymbol{\xi}$.
- Tipos de erros:
  - Erros condicionais: $\boldsymbol{\xi}_j = \boldsymbol{Y}_j - \boldsymbol{Z}_j\gamma - \boldsymbol{X}_j\boldsymbol{u}_j$
  - Erros marginais: $\boldsymbol{\epsilon}_j = \boldsymbol{Y}_j - \boldsymbol{Z}_j\gamma = \boldsymbol{X}_j\boldsymbol{u}_j + \boldsymbol{\xi}_j$
  - Efeitos aleatórios: $\boldsymbol{X}_j\boldsymbol{u}_j = \mathcal{E}(\boldsymbol{Y}_j|\boldsymbol{u}_j) - \mathcal{E}(\boldsymbol{Y}_j)$.
- Respectivos resíduos:
  - Resíduos condicionais: $\widehat{\boldsymbol{\xi}_j} = \boldsymbol{Y}_j - \boldsymbol{Z}_j\widehat{\gamma} - \boldsymbol{X}_j\widehat{\boldsymbol{u}}_j$
  - Resíduos marginais: $\widehat{\boldsymbol{\epsilon}}_j = \boldsymbol{Y}_j - \boldsymbol{Z}_j\widehat{\gamma} = \boldsymbol{X}_j\widehat{\boldsymbol{u}}_j + \widehat{\boldsymbol{\xi}_j}$

Tipos de resíduos

- Segundo Hilden-Minton (1995).
  - Resíduo puro (para um específico tipo de erro): se ele depende apenas das componentes fixas e do erro que ele pretende predizer.
  - Resíduo confundido: depende de outros tipos de erros.

link

Tipos de resíduos

- Na função "lme" (pacote (nlme))
  - Resíduo condicional normalizado: $\widehat{\boldsymbol{\xi}}_j^* = (\widehat{\boldsymbol{U}}_{(c)j})^{-1}\widehat{\boldsymbol{\xi}}_j$
  - Resíduo marginal normalizado: $\widehat{\boldsymbol{\epsilon}}_j^* = (\widehat{\boldsymbol{U}}_{(m)j})^{-1}\widehat{\boldsymbol{\epsilon}}_j$

  em que $\widehat{\boldsymbol{U}}_{(c)j}$ é a matriz triangular superior da decomposição de Cholesky de $\widehat{\boldsymbol{\Sigma}}_j = \widehat{\boldsymbol{U}}'_{(c)j}\widehat{\boldsymbol{U}}_{(c)j}$ (matriz de variâncias e covariâncias dos erros) e $\widehat{\boldsymbol{U}}_{(m)j}$ é a matriz triangular superior da decomposição de Cholesky de $\widehat{\boldsymbol{V}}_j = \widehat{\boldsymbol{U}}'_{(m)j}\widehat{\boldsymbol{U}}_{(m)j}$ (matriz de variâncias e covariâncias de $\boldsymbol{Y}_j$).

- A função "lmer" apresenta (aparentemente) mais limitações nesse sentido (embora seja possível utilizar outras funções e/ou pacotes e/ou fazer as contas de forma externa, com base no "output" dela).

Tipos de resíduos

- Duas opções (dentre outras):
  - Calcular os resíduos usando a função "lme" (requer reajustar o modelo hierárquico, com outra sintaxe veja aqui).
  - Calcular os diferentes tipos e resíduos necessários usando o "output" da função "lmer".
  - Utilizaremos os recursos disponíveis em (aqui) que, em princípio, requerem a utilização da função "lme".

Tipos de resíduos

- Segundo Pinheiro and Bates (2000), página 239, e Schabenberger (2004), respectivamente, $\widehat{\boldsymbol{\xi}}_j^*$ e $\widehat{\boldsymbol{\epsilon}}_j^*$ devem seguir, aproximadamente uma distribuição N(0,1), no caso do modelo estar bem ajustado.
- No entanto, Nobre and Singer (2007) sugerem a utilização do resíduo de confundimento mínimo proposto por Hilden-Milton (1995), veja também Nobre (2004).

Comentários

- Ambos os resíduos condicional e marginal são influenciados tanto pela distribuição dos efeitos aleatórios como pela distribuição dos erros.
- Faz-se mister definir um resíduo que dependa apenas a distribuição dos erros.

Forma matricial

$$\boldsymbol{Y} = \begin{bmatrix} \boldsymbol{Y}_1 \\ \boldsymbol{Y}_2 \\ \vdots \\ \boldsymbol{Y}_J \end{bmatrix}; \boldsymbol{Z} = \begin{bmatrix} \boldsymbol{Z}_1 \\ \boldsymbol{Z}_2 \\ \vdots \\ \boldsymbol{Z}_J \end{bmatrix}; \boldsymbol{\xi} = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_J \end{bmatrix}$$

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_1 & 0 & \dots & 0 \\ 0 & \boldsymbol{X}_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \boldsymbol{X}_J \end{bmatrix}$$

$$\boldsymbol{Y} = \boldsymbol{Z}\gamma + \boldsymbol{X}\boldsymbol{u} + \xi, \boldsymbol{u} = (\boldsymbol{u}_1, \dots \boldsymbol{u}_J)'$$

Forma matricial

- Temos que $\boldsymbol{Y} \sim N_{\sum_{j=1}^{J} n_j}\left(\boldsymbol{Z}\boldsymbol{\gamma}, \sigma^2\boldsymbol{X}\boldsymbol{D}\boldsymbol{X}' + \sigma^2\boldsymbol{R}\right)$ e $\boldsymbol{u} \sim N_{np}(0, \sigma^2\boldsymbol{D})$ $(\boldsymbol{\Sigma}_j = \sigma^2\boldsymbol{R}_j$ e $\boldsymbol{\Psi} = \sigma^2\boldsymbol{D})$, em que

$$\boldsymbol{R} = \begin{bmatrix} \boldsymbol{R}_1 & 0 & \ldots & 0 \\ 0 & \boldsymbol{R}_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \boldsymbol{R}_n \end{bmatrix}; \sigma^2\boldsymbol{D} = \begin{bmatrix} \boldsymbol{\Psi} & 0 & \ldots & 0 \\ 0 & \boldsymbol{\Psi} & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \boldsymbol{\Psi} \end{bmatrix}$$

Formas matriciais para erros e resíduos

- Tipos de erros:
  - Erros condicionais: $\boldsymbol{\xi} = \boldsymbol{Y} - \boldsymbol{Z}\gamma - \boldsymbol{X}\boldsymbol{u}$
  - Erros marginais: $\boldsymbol{\epsilon} = \boldsymbol{Y} - \boldsymbol{Z}\gamma = \boldsymbol{X}\boldsymbol{u} + \boldsymbol{\xi}$
- Respectivos resíduos:
  - Resíduos condicionais: $\widehat{\boldsymbol{\xi}} = \boldsymbol{Y} - \boldsymbol{Z}\widehat{\gamma} - \boldsymbol{X}\widehat{\boldsymbol{u}}$
  - Resíduos marginais: $\widehat{\boldsymbol{\epsilon}} = \boldsymbol{Y} - \boldsymbol{Z}\widehat{\gamma} = \boldsymbol{X}\widehat{\boldsymbol{u}} + \widehat{\boldsymbol{\xi}}$

Resultados: (Nobre (2004))

- $\widehat{\boldsymbol{\xi}} = \boldsymbol{RQ\xi} + \boldsymbol{RQXu}$.
- $\widehat{\boldsymbol{\epsilon}} - \boldsymbol{\epsilon} = -\boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{MZ})^{-1}\boldsymbol{Z}'\boldsymbol{M\epsilon}$.
- $Cov(\widehat{\boldsymbol{\xi}}) = \sigma^2 \boldsymbol{RQR}$.
- $Cov(\widehat{\boldsymbol{\epsilon}}) = \sigma^2 \boldsymbol{M}^{-1}\boldsymbol{QM}^{-1}$.

  em que $\boldsymbol{Q} = \boldsymbol{M} - \boldsymbol{MZ}(\boldsymbol{Z}'\boldsymbol{MZ})^{-1}\boldsymbol{Z}'\boldsymbol{M}$ e $\boldsymbol{M} = \boldsymbol{R}^{-1} - \boldsymbol{R}^{-1}\boldsymbol{XC}^{-1}\boldsymbol{Z}'\boldsymbol{R}^{-1}$ em que $\boldsymbol{C} = \boldsymbol{D}^{-1} + \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X}$. $\boldsymbol{Q}$ e $\boldsymbol{M}$ são matrizes simétricas.

Resíduo Marginal

- Dado que ele está associado ao modelo $\boldsymbol{Y} = \boldsymbol{Z}\gamma + \boldsymbol{\xi}$, um gráfico desse resíduo versus cada covariável (variável explicativa) ajuda a verificar se o preditor linear está corretamente especificado.
- Espera-se, em caso de correta especificação, que o gráfico tenha um comportamento aleatório em torno do zero.

Resíduo Condicional

- Seja (resíduo condicional padronizado) $\widehat{\xi}_j^* = \frac{\widehat{\xi}_j}{\widehat{\sigma}\sqrt{\widehat{p}_{jj}}}$, em que $\widehat{p}_{jj}$ é a estimativa de p_{jj} o qual, por sua vez, é o j-ésimo elemento da diagonal principal de $\boldsymbol{RQR}$.
- Presença de "outliers": $\xi_j^* \times$ índice da observação.
- Homocedasticidade dos erros condicionais: $\xi_j^* \times$ valor predito ($\widehat{\boldsymbol{Y}}_j = \boldsymbol{Z}_j\widehat{\gamma} + \boldsymbol{X}_j\widehat{\boldsymbol{u}}_j$).

Resíduo de confundimento mínimo

- Para checar a normalidade dos erros condicionais.
- Lembre que: $\widehat{\boldsymbol{\xi}} = \boldsymbol{RQ}\xi + \boldsymbol{RQXu}$. Hilden-Milton (1995) (HM) argumentam que a habilidade para avaliar a normalidade de $\widehat{\boldsymbol{\xi}}$ diminui quando $\mathcal{V}(\boldsymbol{RQXu}) = \boldsymbol{RQXDX'QRU}_j$ aumenta em relação à $\mathcal{V}(\boldsymbol{RQ}\xi) = \sigma^2\boldsymbol{RQRQR}$ (temos ainda que: $Cov(\widehat{\xi}) = \sigma^2\boldsymbol{RQR}$).
- Esse autor define a fração de confundimento para $\widehat{\boldsymbol{\xi}}_j$ como $0 \leq FC_j = \frac{\boldsymbol{U}_j'\boldsymbol{RQXDX'QRU}_j}{\boldsymbol{U}_j'\boldsymbol{RQRU}_j} = 1 - \frac{\boldsymbol{U}_j'\boldsymbol{RQRQRU}_j}{\boldsymbol{U}_j'\boldsymbol{RQRU}_j} \leq 1$ em que $\boldsymbol{U}_j$ denota a j-ésima linha da matriz $\boldsymbol{I}_n$. FC_j representa a porção da variabilidade de $\widehat{\boldsymbol{\xi}}_j$ devida ao confundimento com $\widehat{\boldsymbol{u}}$.
- Quanto maior for FC_j, maior é o grau de confundimento de $\widehat{\boldsymbol{\xi}}_j$.

Resíduo de confundimento mínimo

- HM sugere utilizar uma transformação linear, digamos, $\boldsymbol{L\xi}$, que minimize o confundimento (em algum sentido).
- Uma sugestão é minimizar o confundimento de $\boldsymbol{l}_j'\boldsymbol{\xi}$ (em que $\boldsymbol{l}_j'$ é a j-ésima linha da matriz $\boldsymbol{L}$), ou seja maximizar

$$\lambda_j = \frac{\boldsymbol{l}_j'\boldsymbol{RQRQRl}_j}{\boldsymbol{l}_j'\boldsymbol{RQRl}_j}$$

  sujeito à restrição $\mathcal{V}(\boldsymbol{l}_j'\widehat{\boldsymbol{\xi}}_j) \propto \boldsymbol{l}_j'\boldsymbol{RQRl}_j > 0$.
- A base para a obtenção de $\boldsymbol{l}_j$ é a decomposição espectral de $\boldsymbol{R}^{1/2}\boldsymbol{QR}^{1/2} = \boldsymbol{K}\Pi\boldsymbol{K}'$.

link

Resíduo de confundimento mínimo

- Após várias manipulações algébricas, chega-se que o resíduo de confundimento mínimo é dado por:

$$\boldsymbol{l}_j' \widehat{\boldsymbol{\xi}} = \sqrt{\pi_j} \boldsymbol{K}_j' \boldsymbol{R}^{-1/2} \boldsymbol{Y}$$

em que $\boldsymbol{K}_j$ é a j-ésima coluna de $\boldsymbol{K}$ e π_j é o j-ésimo autovalor (ordenado) (da matriz $\boldsymbol{\Pi}$), j=1,2,...,n-p ("n-p" é devido ao rank da matriz $\boldsymbol{Q}$)

- Pode-se provar ainda que $Cov(\boldsymbol{l}_i' \widehat{\boldsymbol{\xi}}, \boldsymbol{l}_j' \widehat{\boldsymbol{\xi}}) = \sigma^2 \delta_{ij}, i, j = 1, 2, ..., n-p$, em que $\delta_{ij} = 1$ se $i = j$ e 0 caso contrário.

Resíduo de confundimento mínimo

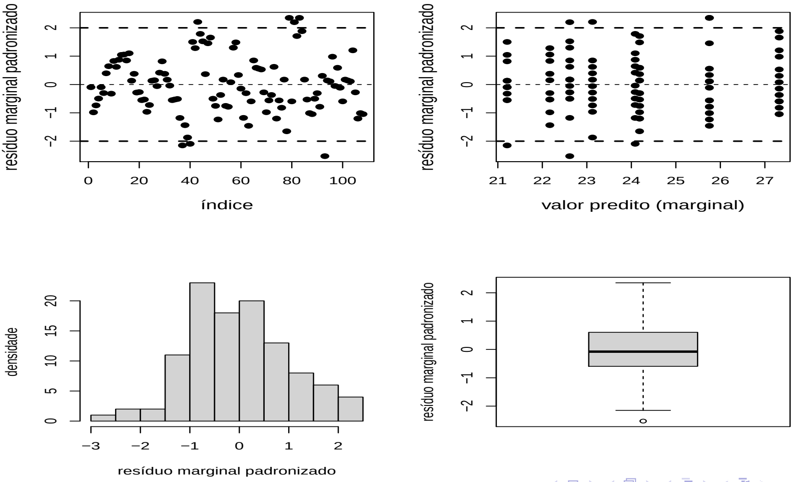
- Os resíduos de confundimento mínimo são não correlacionados, com variância constante e média zero.
- Além disso, $\frac{\boldsymbol{l}_j \widehat{\boldsymbol{\xi}}}{\widehat{\sigma}}$ tem aproximadamente distribuição N(0,1).
- Sugere-se fazer um gráfico de quantil quantil com envelopes para $\frac{\boldsymbol{l}_j \widehat{\boldsymbol{\xi}}}{\widehat{\sigma}}$.

Voltando ao Exemplo 5: (Potthoff and Roy (1964))

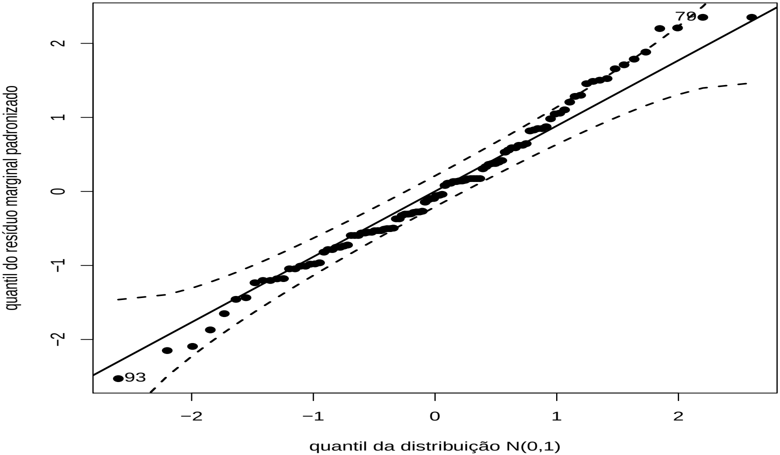
- Identificou-se, nos slides link, o modelo mais apropriado, dentre aqueles considerados.
- Vimos que a suposição de normalidade dos efeitos aleatórios parece ser não razoável, através de seus valores preditos (referências).
- Também vimos, através de resultados descritivos, que algumas das suposições sobre os erros podem não ser válidas.

link

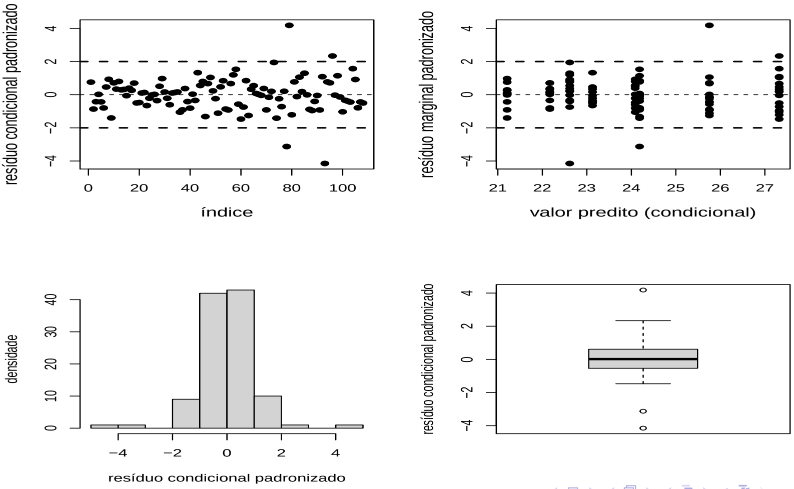
Potthof and Roy (RMP)



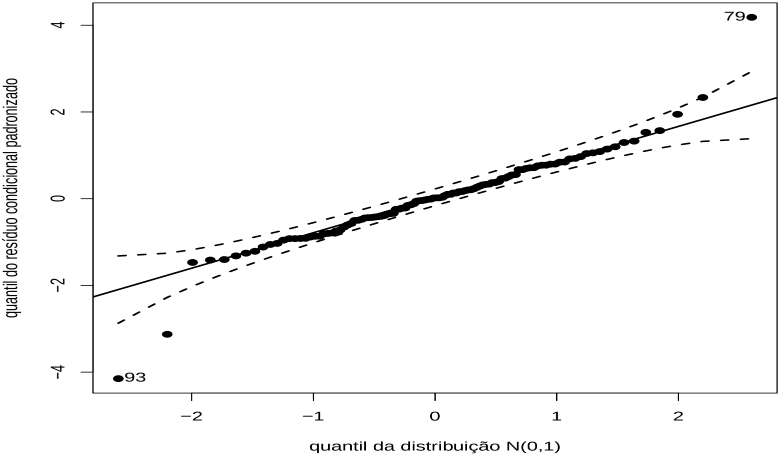
Potthof and Roy (QQplot RMP)



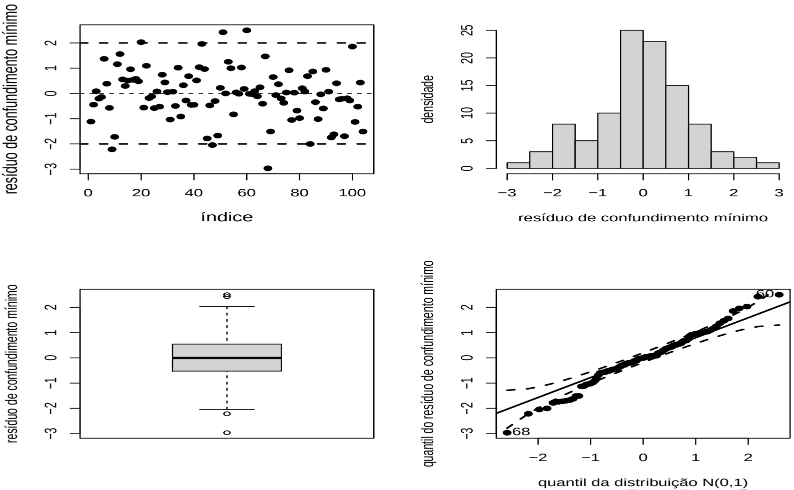
Potthof and Roy (RCP)



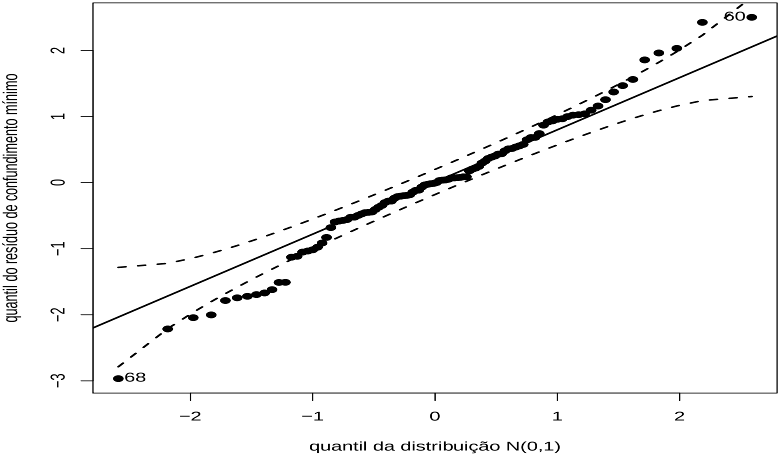
Potthof and Roy (QQplot RCP)



Potthof and Roy (RCM)



Potthof and Roy (QQplot RCM)



link