Análise de Resíduos

→ investiga características que comprometem a validade do MRLS:

(1) relação entre X e Y não é linear

(2) erros não tem variância constante → homoscedasticidade

(3) erros correlacionados

(4) erros não são normalmente distribuídos

(5) modelo não ajusta bem a uma ou mais observações → prováveis dados atípicos (avaliar sua influência)

(6) uma ou mais covariáveis não foram incluídas no modelo

Gráficos de Resíduos

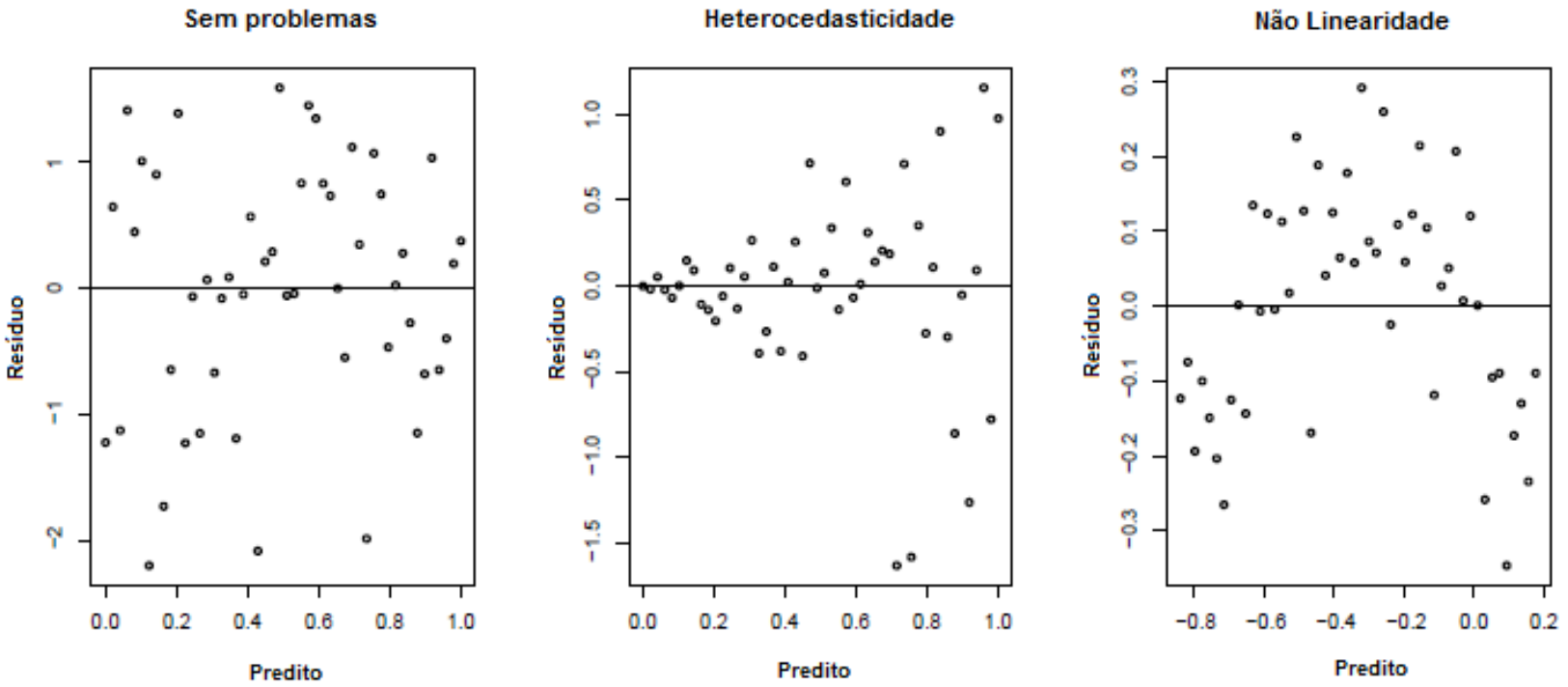
1) diagrama de dispersão de resíduo e predito

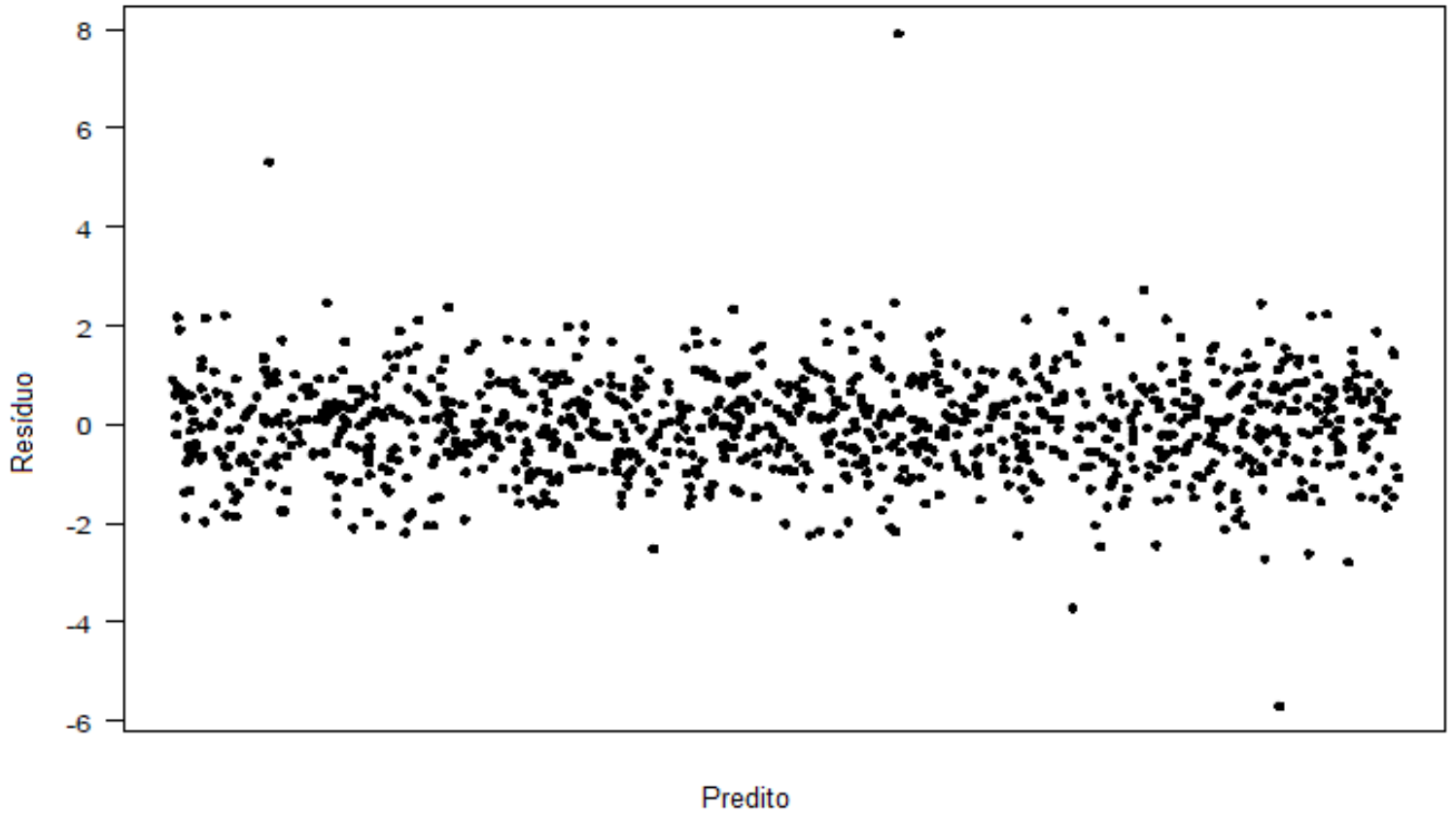
→ detectar heterocedasticidade de ε_i

→ detectar não-linearidade entre X e Y

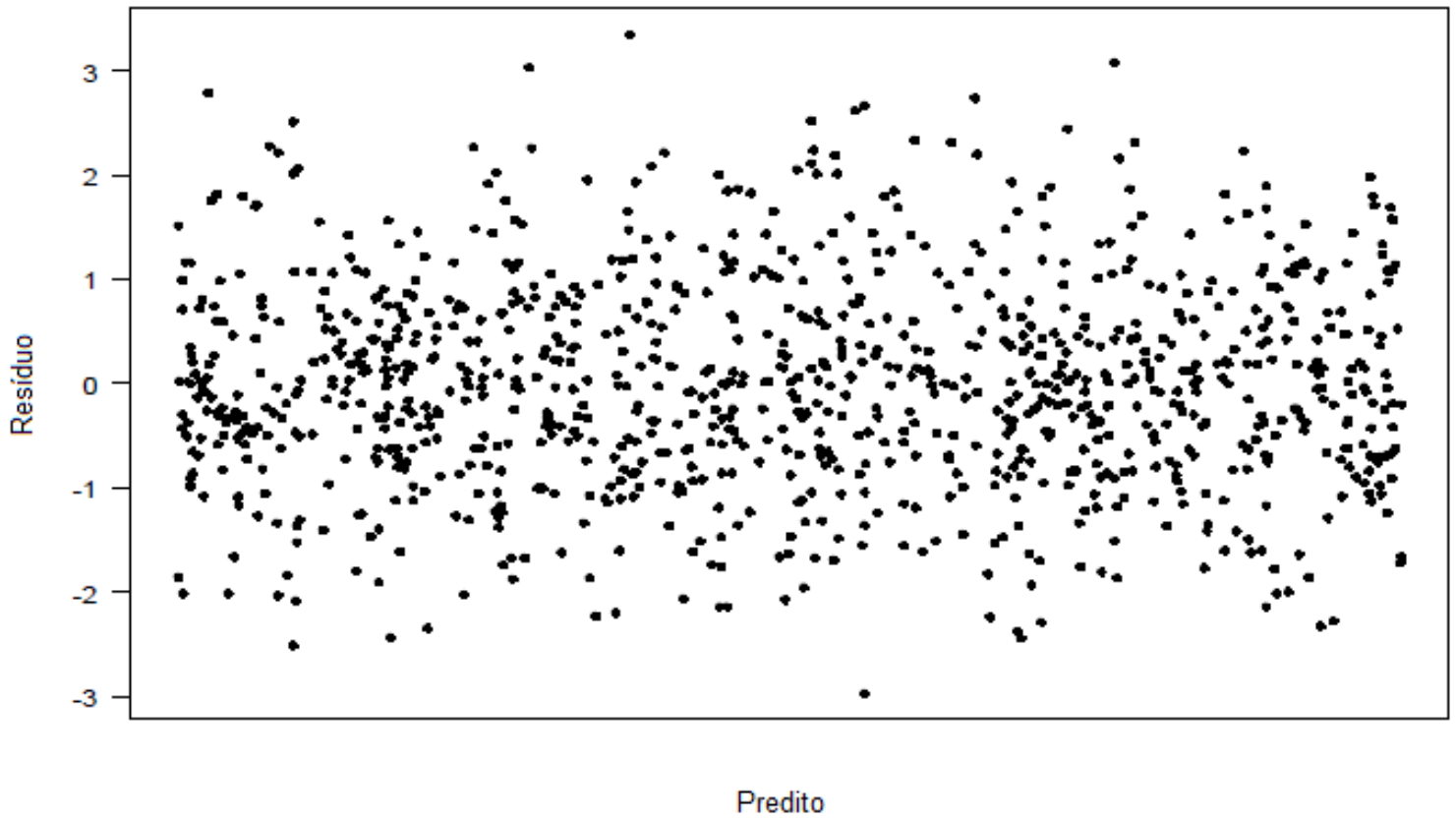
→ detectar prováveis dados atípicos

modelo bem ajustado: resíduos dispersos aleatoriamente em torno de zero, com variância constante, concentrados entre -2 e 2 e pouquíssimos pontos acima de 3 ou abaixo de -3

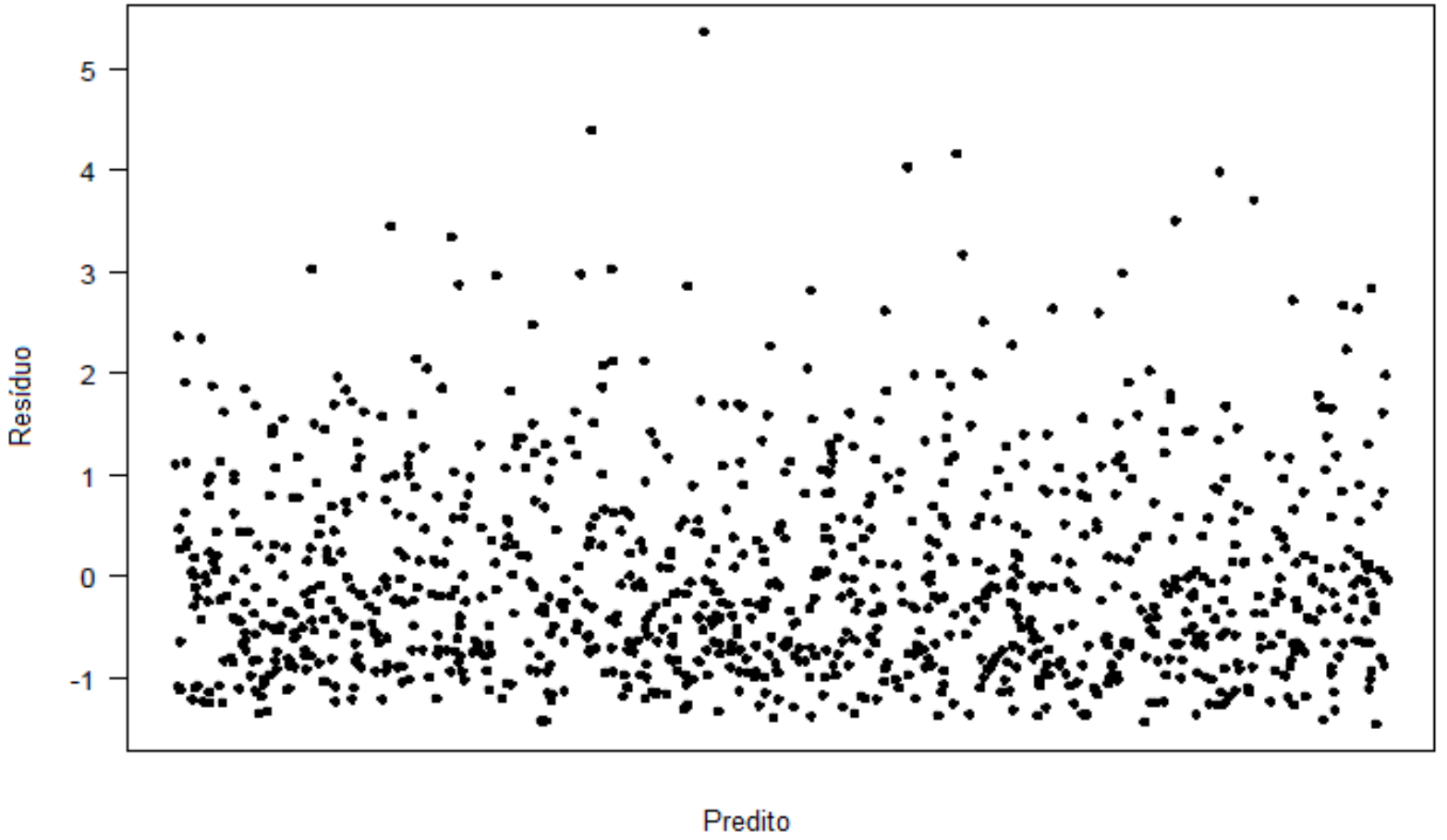

Sem problemas
Resíduo
Predito
Heterocedasticidade
Resíduo
Predito
Não Linearidade
Resíduo
Predito



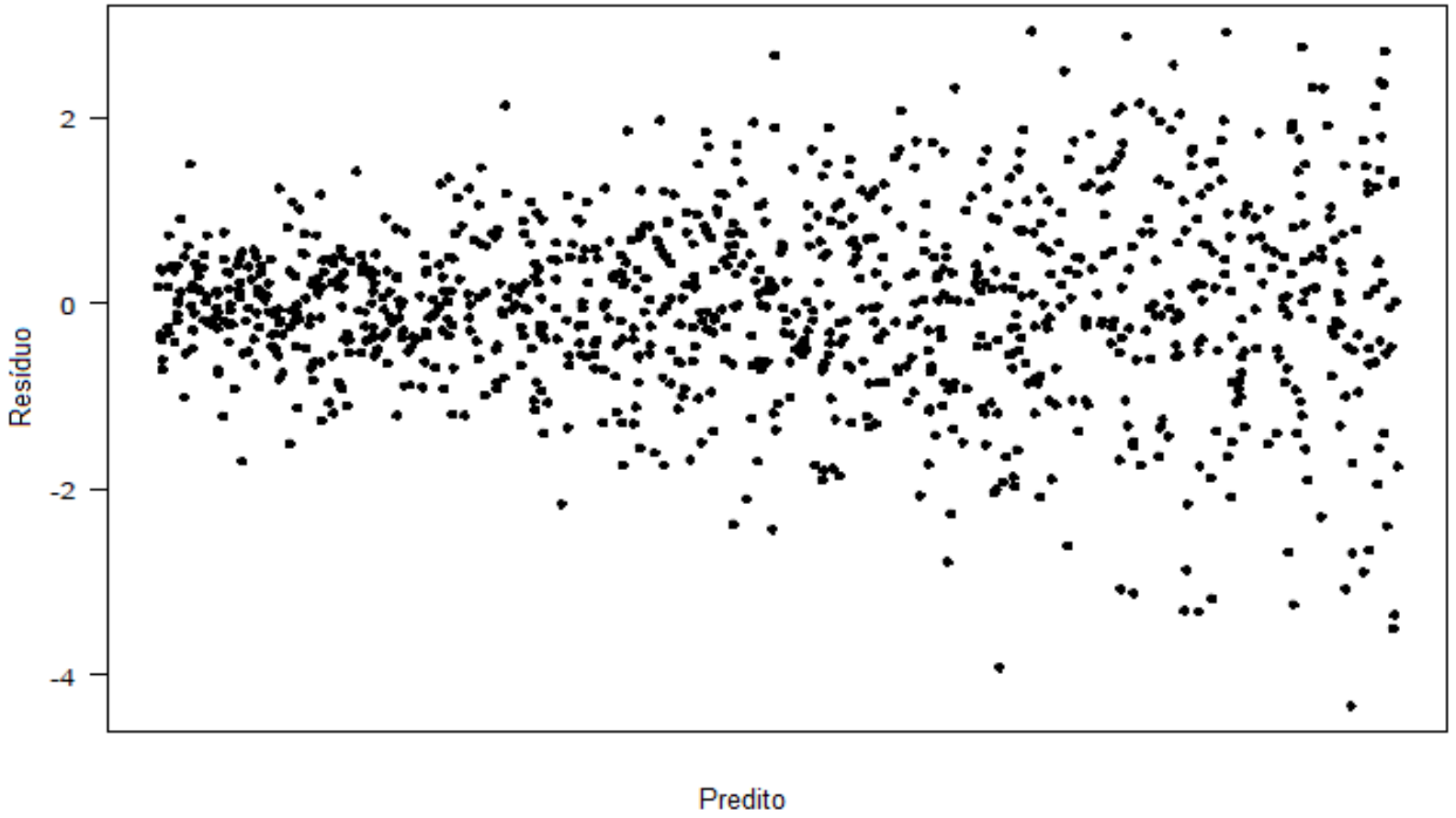
Presença de alguns resíduos extremos (observações mal ajustadas)



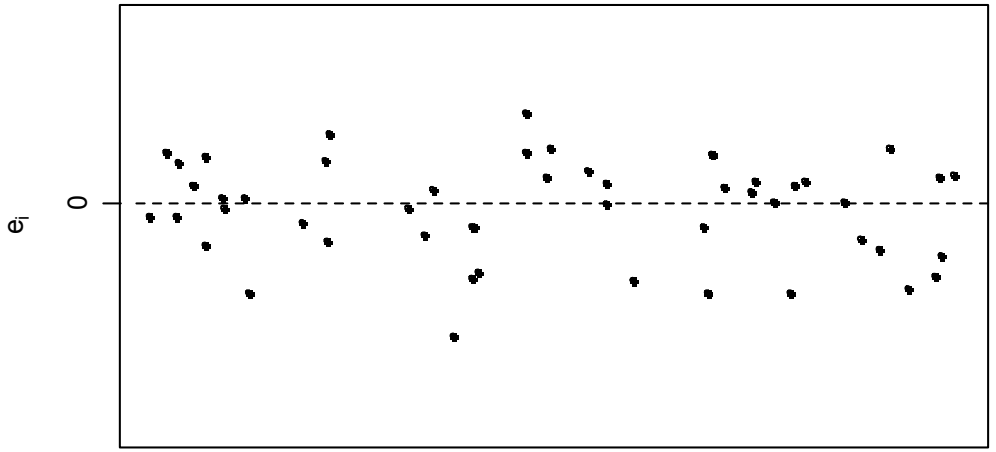
Distribuição dos resíduos indica boa qualidade de ajuste



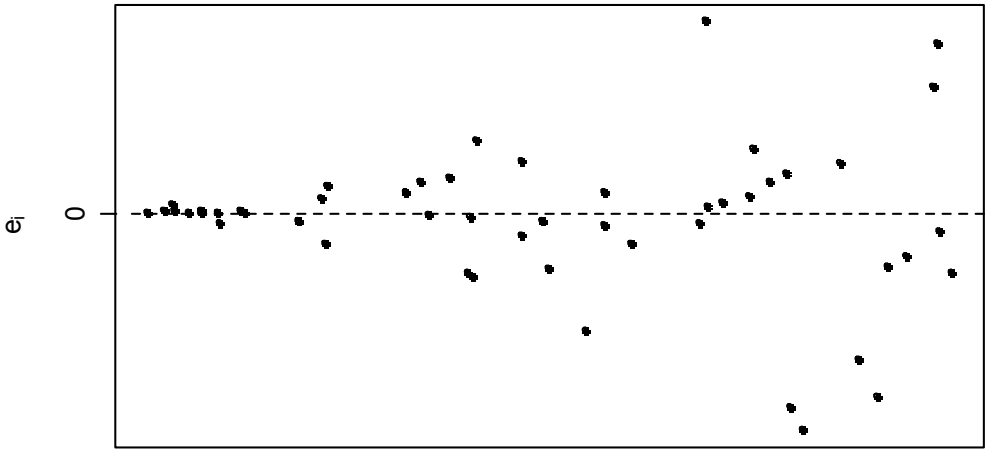
Resíduos com distribuição fortemente assimétrica



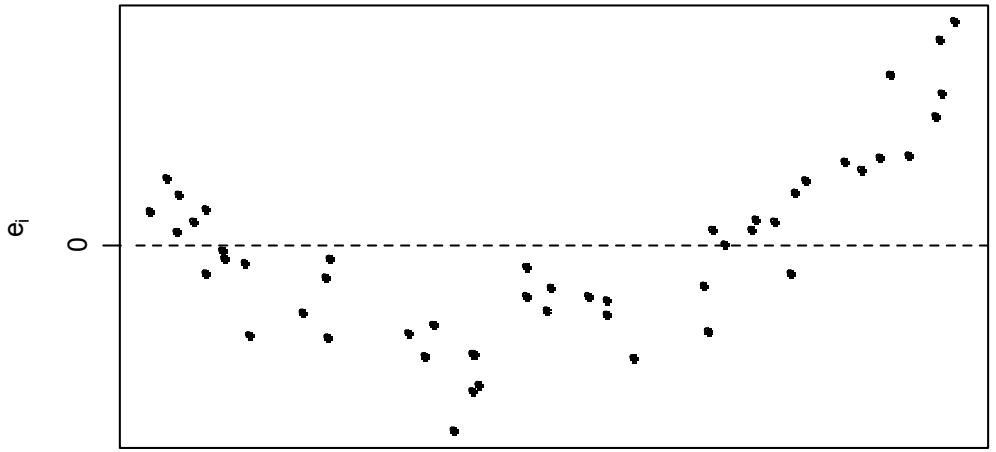
Indicação de erros heterocedásticos (variância não constante dos erros)



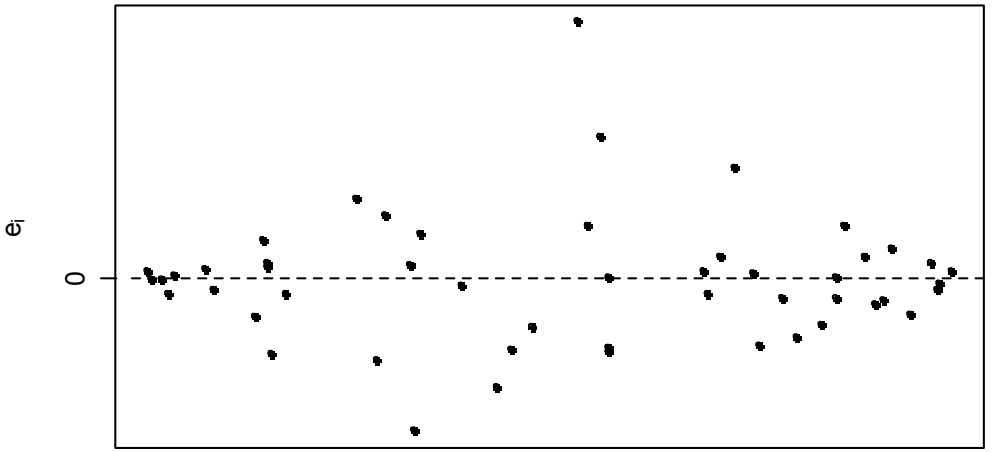
(a)



(b)



(c)



(d)

(a) Resíduos dispersos aleatoriamente em torno de zero, indica o comportamento esperado para distribuição dos erros

(b) Dispersão dos resíduos aumenta conforme o valor do predito, configurando heterogeneidade de variâncias dos erros (erros heterocedásticos); comum quando a variável resposta refere-se a contagens
solução: transformar a variável resposta ou utilizar algum modelo linear generalizado

(c) Distribuição dos resíduos apresenta uma tendência não linear (no caso, quadrática)
solução: incorporar novas variáveis explicativas ao modelo, ou considerar alguma transformação em X e/ou Y, ou utilizar algum modelo de regressão não linear

(d) Distribuição dos resíduos indica erros heterocedásticos; comum quando a variável reposta refere-se a proporções; há também uma observação com resíduo muito elevado, indicando que não é bem ajustada pela reta

solução: transformar a variável resposta ou considerar algum modelo linear generalizado; deve-se verificar inicialmente se o valor atípico foi coletado e registrado corretamente

correto

deve ser considerado na análise: investigar o motivo da discrepância e avaliar de que forma essa observação afeta os resultados (análise de influência)

incorreto

deve ser corrigido ou, caso não seja possível, descartá-lo

2) gráfico probabilístico normal dos resíduos

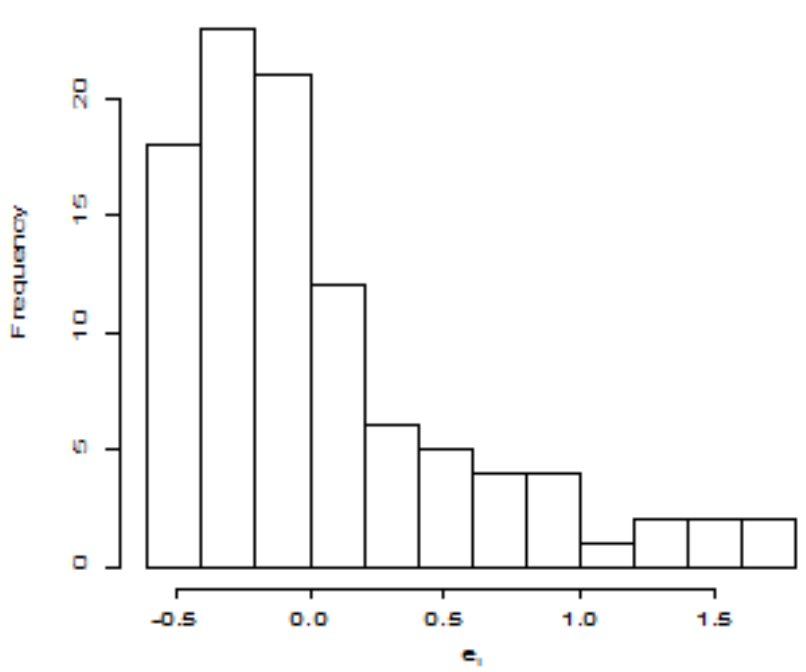
é um gráfico de pontos de quantis amostrais dos resíduos versus quantis teóricos da distribuição normal padrão
(q-q plot normal; quantil-quantil normal)

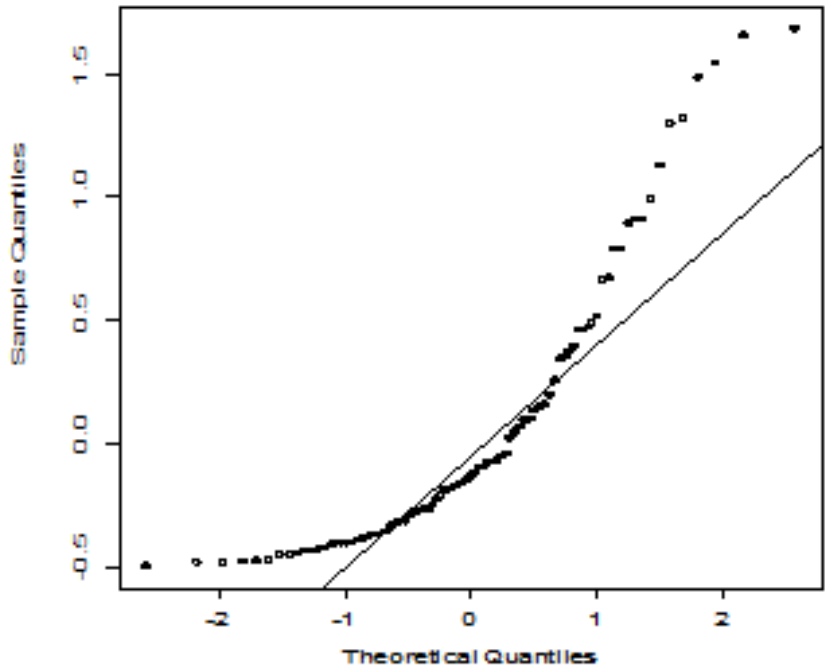
→ detectar não normalidade de ε_i

→ detectar dados atípicos

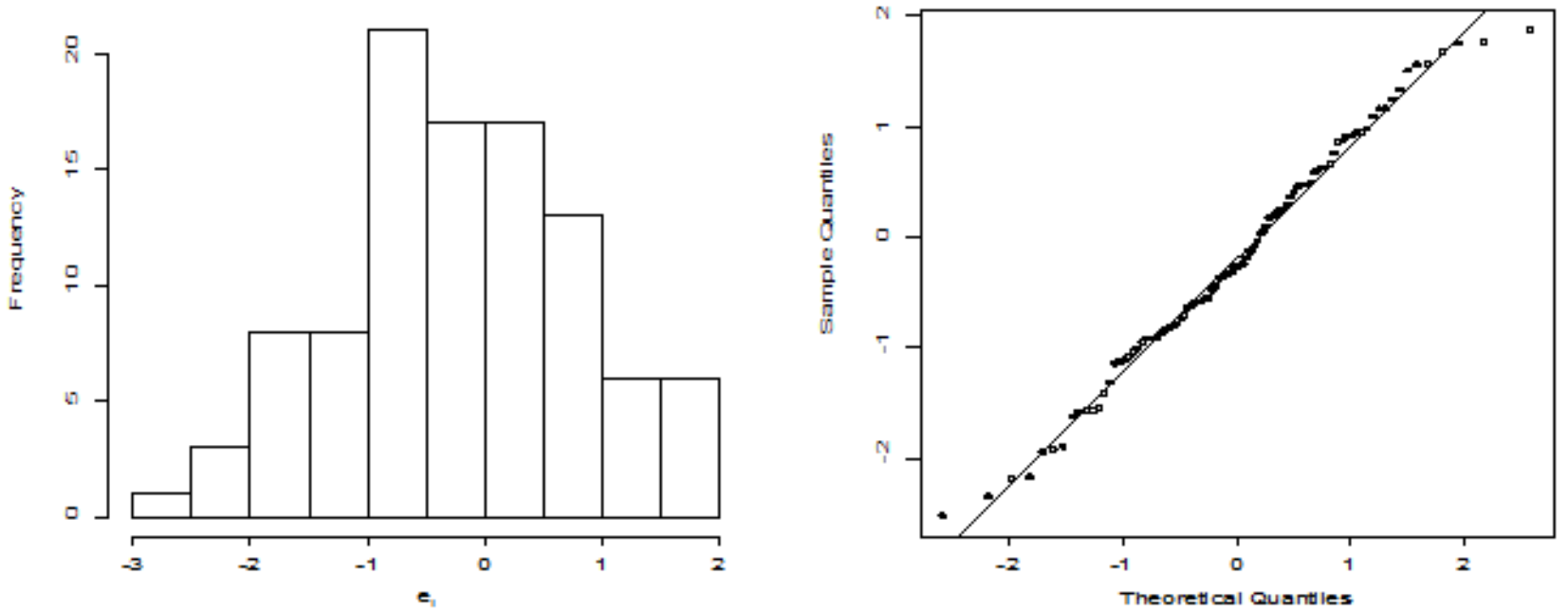
modelo bem ajustado: pontos alinhados na reta que representa a identidade dos quantis amostrais e teóricos

gráfico alternativo: histograma ou box-plot dos resíduos

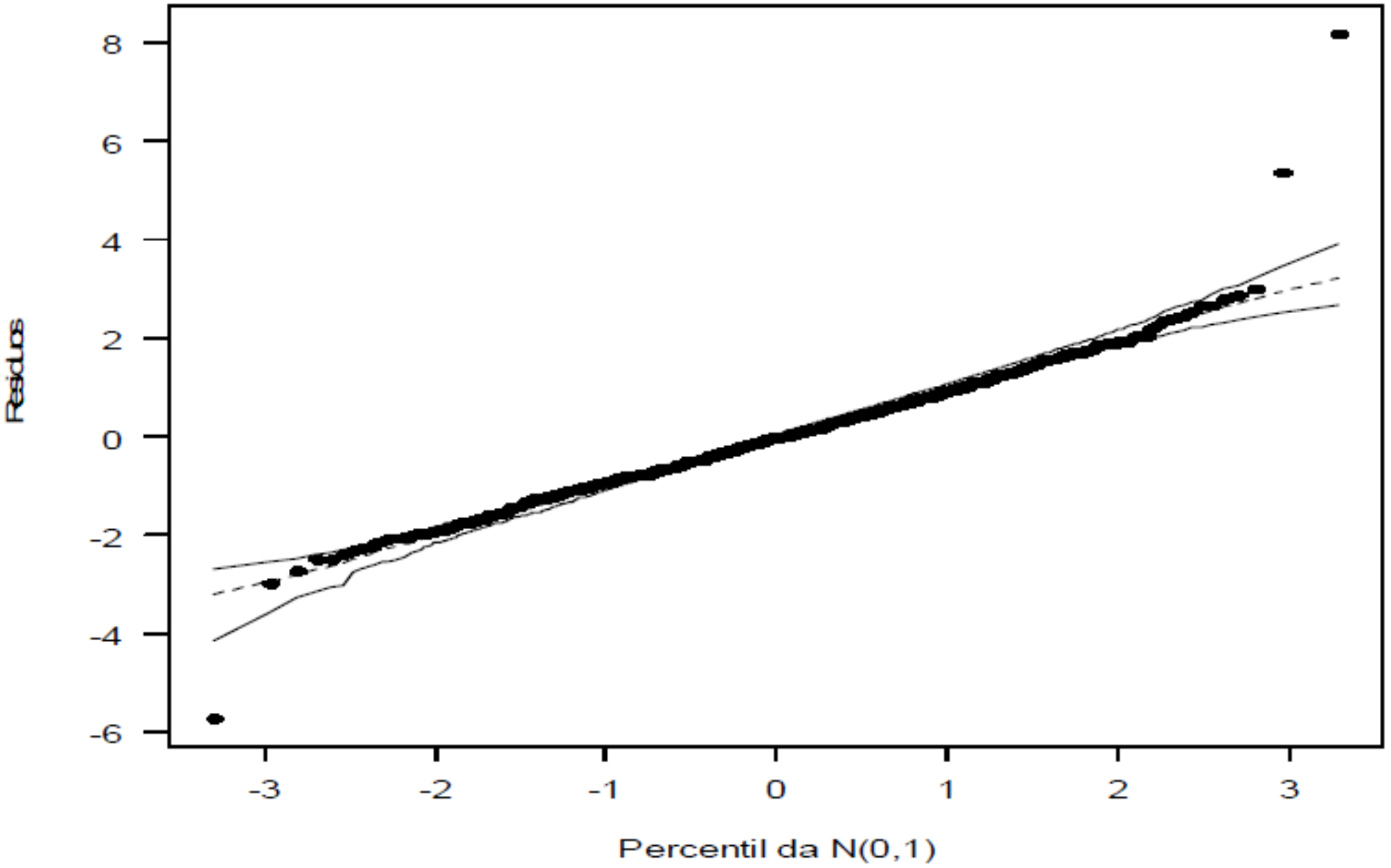




resíduos com distribuição assimétrica

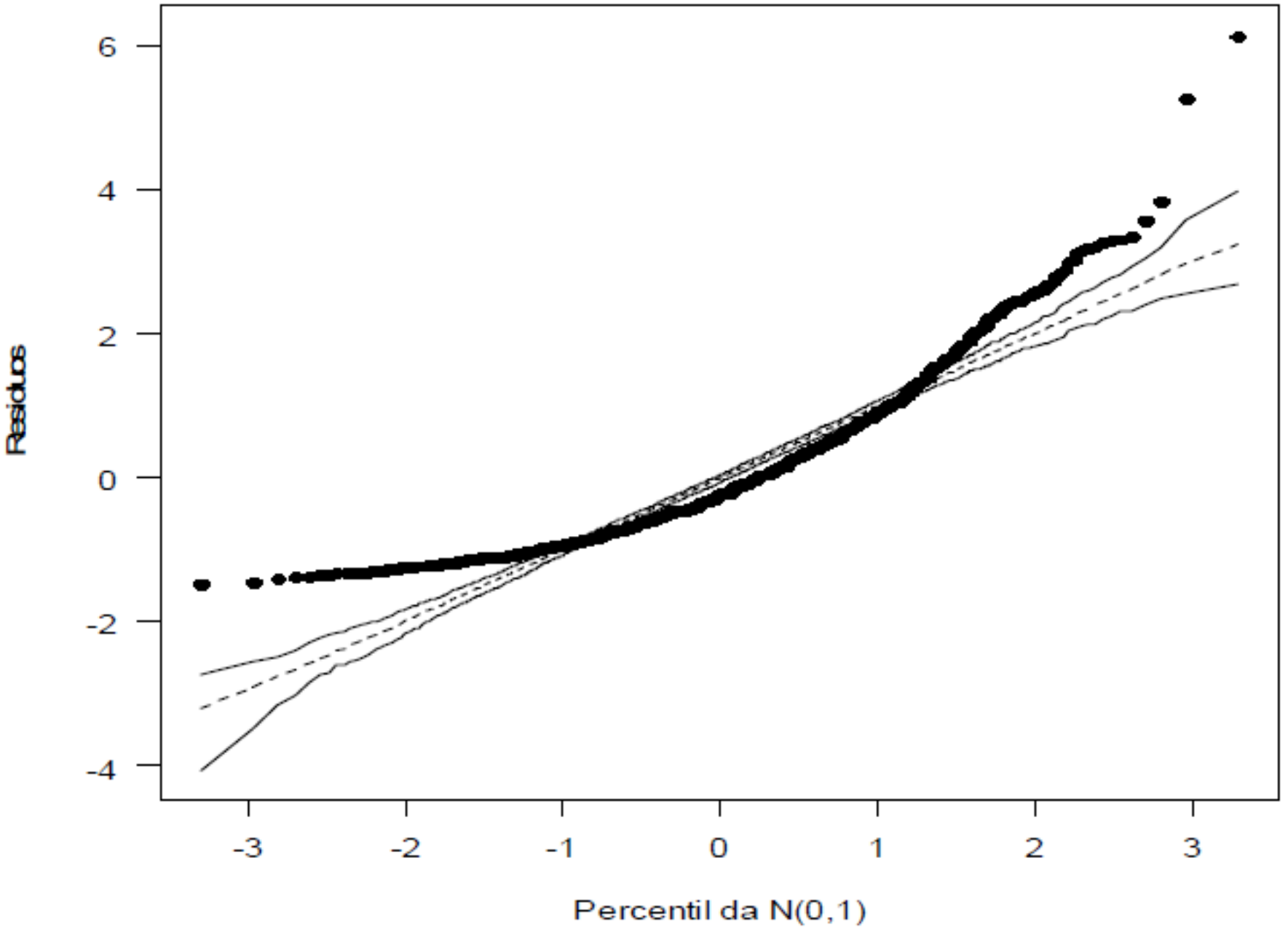


histograma e gráfico probabilístico normal dos resíduos evidenciam a normalidade dos erros



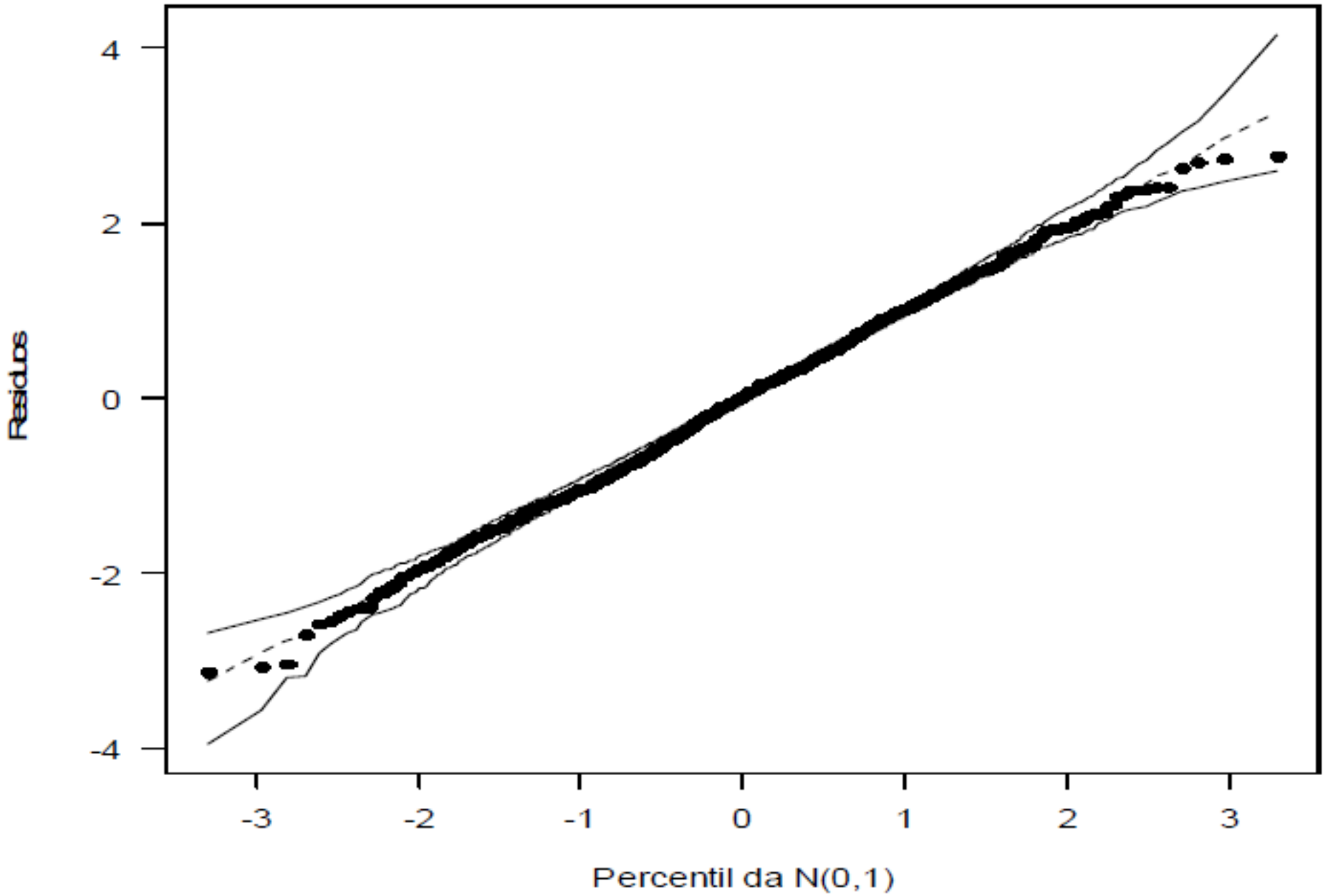
ocorrência de observações mal ajustadas

<u>solução</u>: corrigir o valor atípico, se for o caso, ou fazer análise de influência

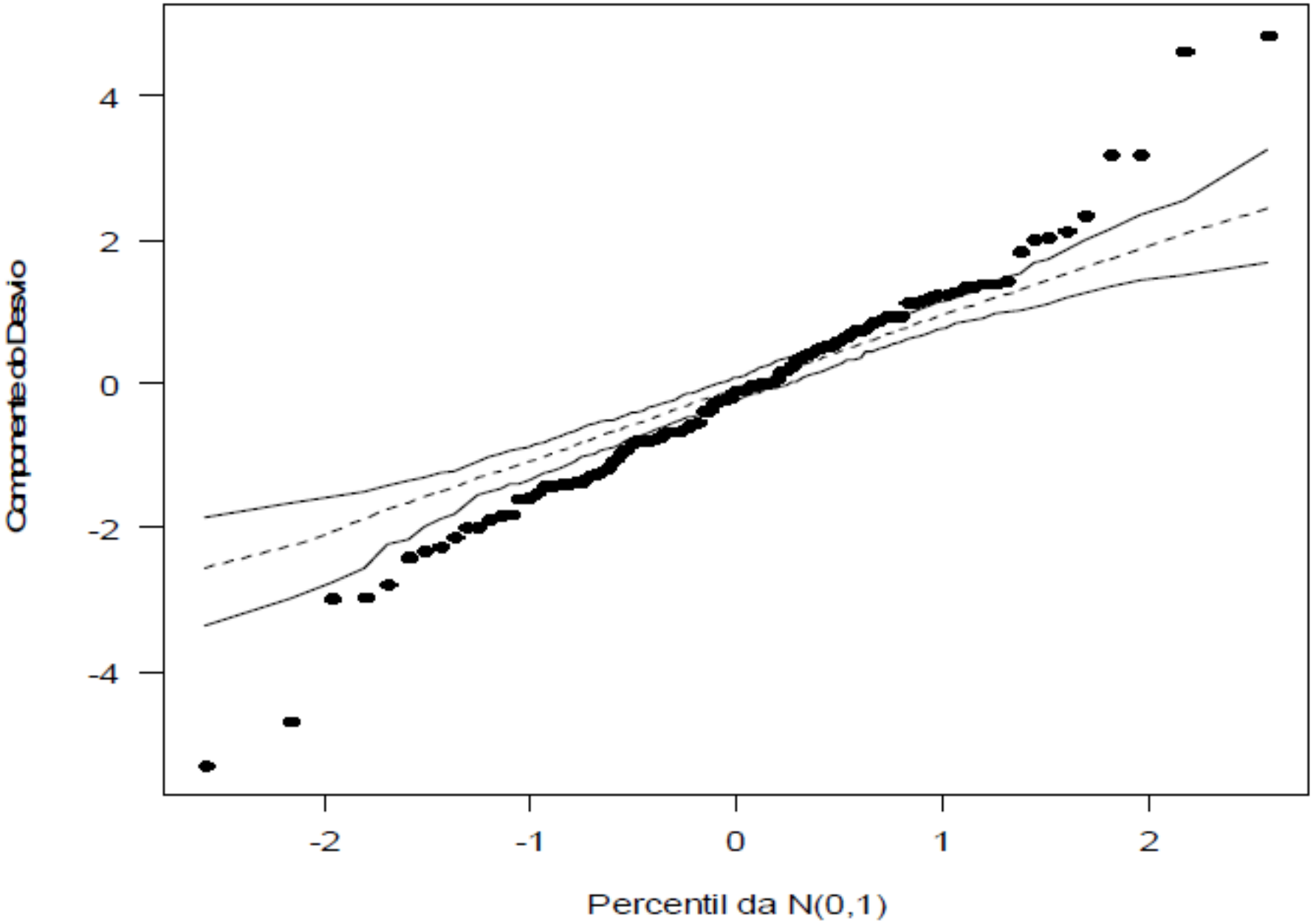


resíduos com distribuição fortemente assimétrica

<u>solução</u>: transformar a variável resposta ou utilizar algum modelo linear generalizado



indicação de bom ajuste



Distribuição dos resíduos com “caldas pesadas”

solução: transformar a variável resposta ou utilizar algum modelo linear generalizado

Se conhecida a ordem de coleta de dados:

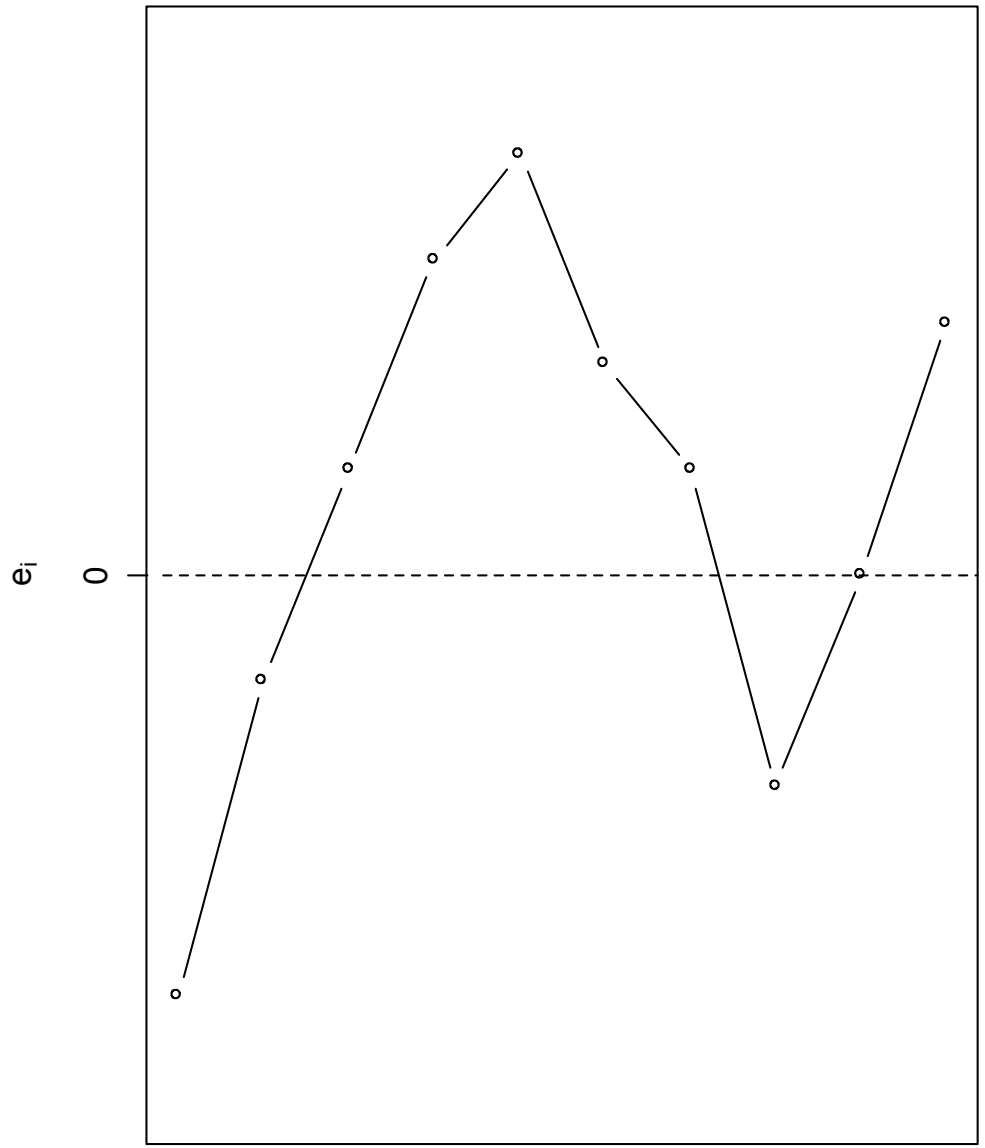
3) gráfico de resíduos versus sequencia de coleta de dados ($e_i \times i$)

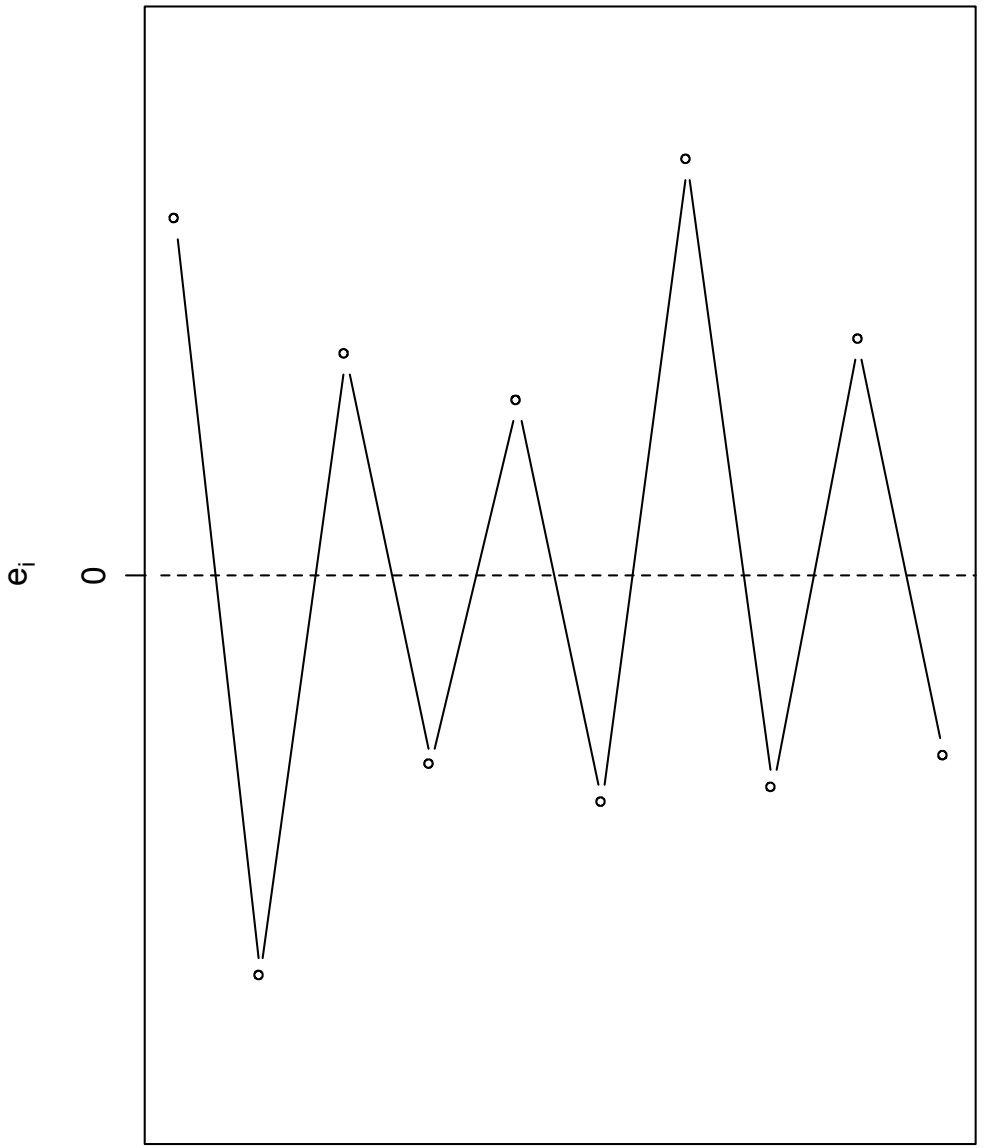
→ detectar ε_i correlacionados com a ordem de coleta dos dados

modelo bem ajustado: distribuição aleatória dos resíduos em torno do zero

* a presença de algum padrão sistemático pode indicar dependência com relação à ordem de coleta

gráfico alternativo: gráfico de resíduos versus posição da observação no tempo ou espaço





gráficos (a) e (b) evidenciam, em sua maneira, que os erros estão correlacionados

solução: análise de séries temporais com covariáveis ou análise de dados longitudinais

4) diagrama de dispersão de resíduo e covariável

→ detectar variação na magnitude de σ^2 em relação a X

→ detectar não-linearidade entre X e Y

→ detectar prováveis dados atípicos

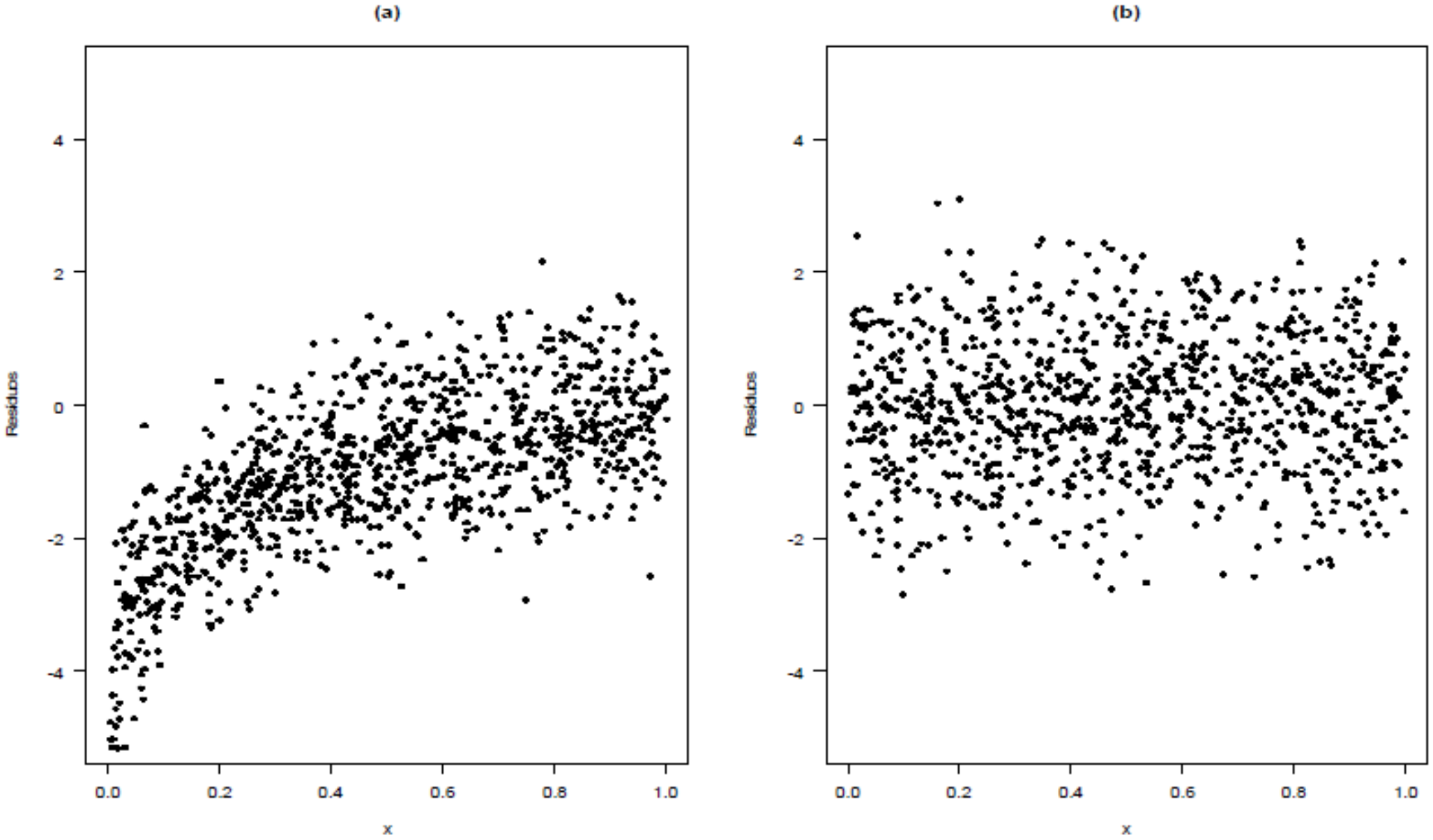
modelo bem ajustado: resíduos aleatoriamente dispersos em torno de zero

* a presença de algum padrão sistemático indica que a variável em questão não foi incluída no modelo numa escala correta

Se conhecido os valores da covariável omitida:

5) diagrama de dispersão de resíduo e covariável omitida

* qualquer padrão sistemático indica a necessidade de se incorporar a variável ao modelo



(a) indica que a variável *x* deve ser inserida de outra forma (ou deve ser incluída) no modelo; ou utilizar algum modelo de regressão não linear
(b) não se tem indicativo da mudança de escala de *x* (ou da necessidade de inclusão)

CORREÇÃO

$$Var(e_i) = Var(y_i - \hat{y}_i) = Var(y_i) + Var(\hat{y}_i) - 2\,Cov(y_i, \hat{y}_i)$$

$$= \sigma^2 + \sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right) - 2\left[\sigma^2\left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}}\right)\right]$$

$$= \sigma^2\left(1 - \frac{1}{n} - \frac{(x_i - \bar{x})^2}{S_{xx}}\right)$$

CORREÇÃO

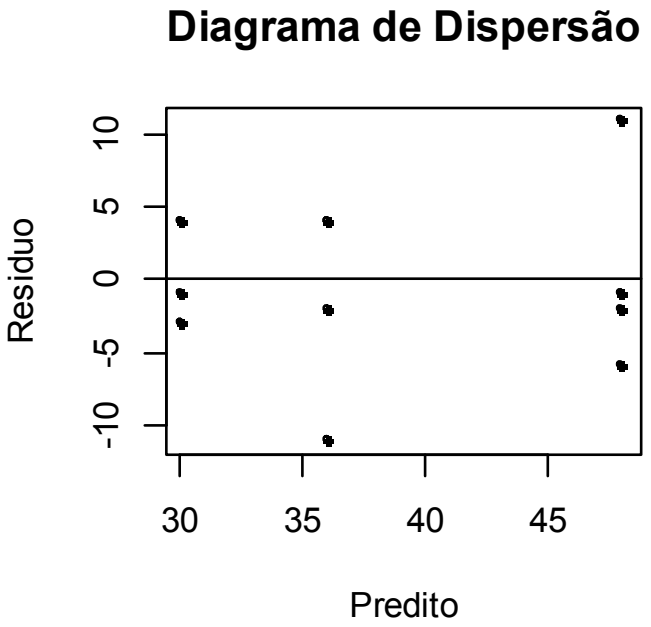
Como $h_{ii} = \left(\frac{1}{n} + \frac{(x_i - \bar{x})^2}{S_{xx}} \right)$ então: $Var(e_i) = \sigma^2 (1 - h_{ii})$

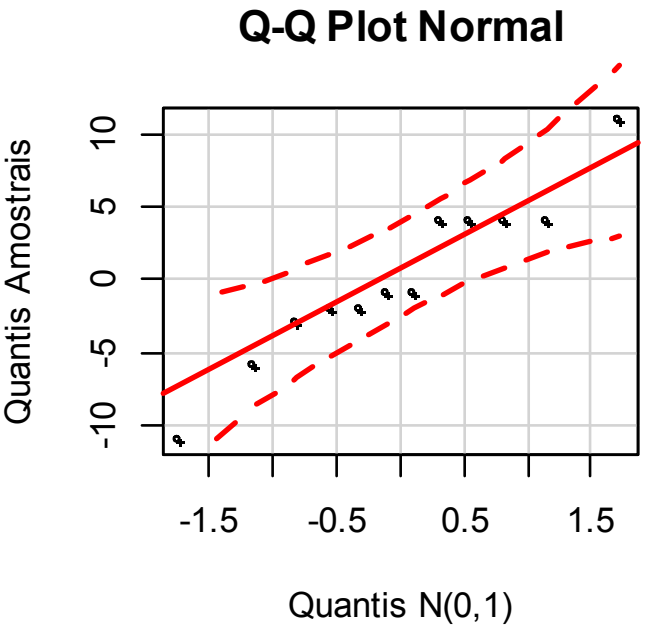
$$e_i \sim N(0, \sigma^2 (1 - h_{ii})) \quad ; \; i = 1,2,\ldots,n$$

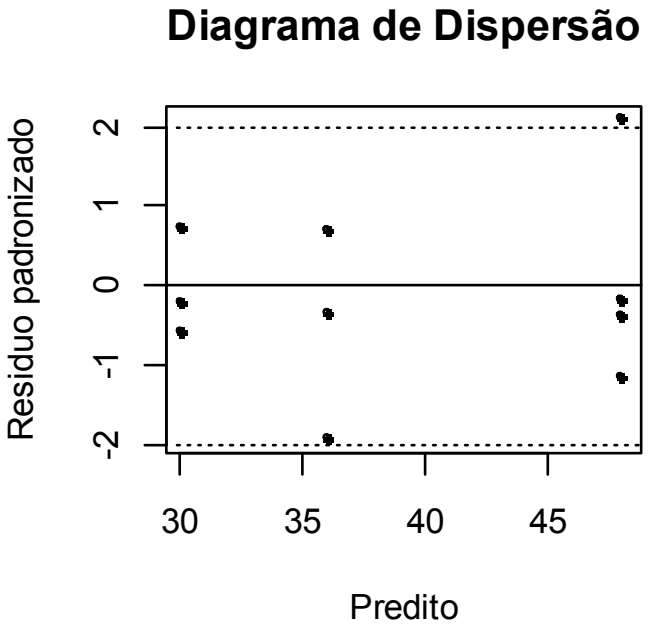
$$z_i = \frac{e_i}{\sqrt{\sigma^2 (1 - h_{ii})}}$$

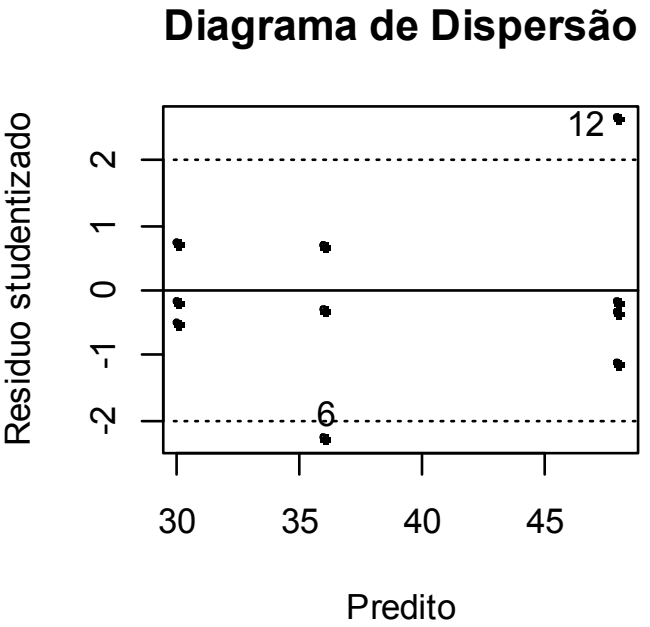
$$z_i^* = \frac{e_i}{\sqrt{\sigma_{(i)}^2 (1 - h_{ii})}}$$

Análise de Resíduos do Exemplo 1









```
###dados - Exemplo 1
x <- c(rep(1.35,4),rep(1.4,4),rep(1.5,4))
y <- c(34,34,29,27,40,25,40,34,46,42,47,59)

#ajuste de MQ
reta<- lm(y~x)

# Análise de Resíduos
predito <- reta$fit
residuo <- reta$res
cbind(predito,residuo)
# transformacoes dos residuos
z <- rstandard(reta)    # residuos padronizados
zstudent <- rstudent(reta)    # residuos studentizados
cbind(z,zstudent)
# graficos de residuos
par(mfrow=c(2,2))
# residuo vs predito
plot(predito, residuo, pch=20, main="Diagrama de Dispersão", xlab="Predito", ylab="Residuo")
abline(h=0)
#q-q plot normal envelope
require(car)    #para instalar o pacote use: install.packages()
require(MASS)
qqPlot(residuo, pch=20, main="Q-Q Plot Normal", xlab="Quantis N(0,1)", ylab="Quantis Amostrais")
# residuo transformado vs predito
plot(predito, z, pch=20, main="Diagrama de Dispersão", xlab="Predito", ylab="Residuo padronizado")
abline(h=0)
abline(h=2, lty=3)
abline(h=-2, lty=3)
plot(predito, zstudent, pch=20, main="Diagrama de Dispersão", xlab="Predito", ylab="Residuo studentizado")
abline(h=0)
abline(h=2, lty=3)
abline(h=-2, lty=3)
#identificar n pontos clicando próximo aos pontos
identify(predito,zstudent,n=2)
```