Modelos de regressão para dados correlacionados

Cibele Russo

cibele@icmc.usp.br

ICMC USP

Mini-curso oferecido no
Workshop on Probabilistic and Statistical Methods

28 a 30 de janeiro de 2013

Conteúdo da aula

- Análise de diagnóstico
- Seleção de modelos

Análise de diagnóstico

A análise de diagnóstico em modelos estatísticos visa verificar a coerência das suposições de um modelo e avaliar a qualidade do ajuste resultante.

Análise de diagnóstico

Entre as técnicas de diagnóstico mais utilizadas, estão

- Análise de resíduos
- Diagnóstico de influência, que inclui
  - Exclusão de casos
  - Influência local

Modelo de regressão linear (sem efeitos aleatórios)

No modelo com p covariáveis

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon} \text{ com } \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 I_n), \text{ temos o}$$

Vetor de resíduos:

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{Y}}$$

em que $\mathbf{y}$ é o vetor observado de $\mathbf{Y}$ e $\widehat{\mathbf{Y}}$ é o vetor ajustado pelo modelo de regressão.

Podemos reescrever $\mathbf{e}$ como

$$\mathbf{e} = (I - H)\mathbf{Y}$$

com

$$H = X(X'X)^{-1}X' \text{ (matriz chapéu ou hat).}$$

Modelo de regressão linear (sem efeitos aleatórios)

Temos que

$$E(\mathbf{e}) = (I - H)E(\mathbf{Y}) = 0 \text{ e}$$

$$Var(\mathbf{e}) = \sigma^2(I - H)$$

pois

$$Var(\mathbf{e}) = Var((I - H)\mathbf{Y}) = (I - H)Var(\mathbf{Y})(I - H)'.$$

Como $(I - H)$ é simétrica e idempotente,

$$Var(\mathbf{e}) = Var((I-H)\mathbf{Y}) = (I-H)\sigma^2 I(I-H) = (I-H)(I-H)\sigma^2 = (I-H)\sigma^2.$$

Modelo de regressão linear (sem efeitos aleatórios)

Uma possibilidade é trabalhar com o resíduo padronizado

$$t_i = \frac{r_i}{s(1-h_{ii})^{1/2}}$$

com $s^2 = \sum_{i=1}^{n} r_i^2/(n-p)$ e h_{ii} é o i-ésimo elemento da diagonal de H.

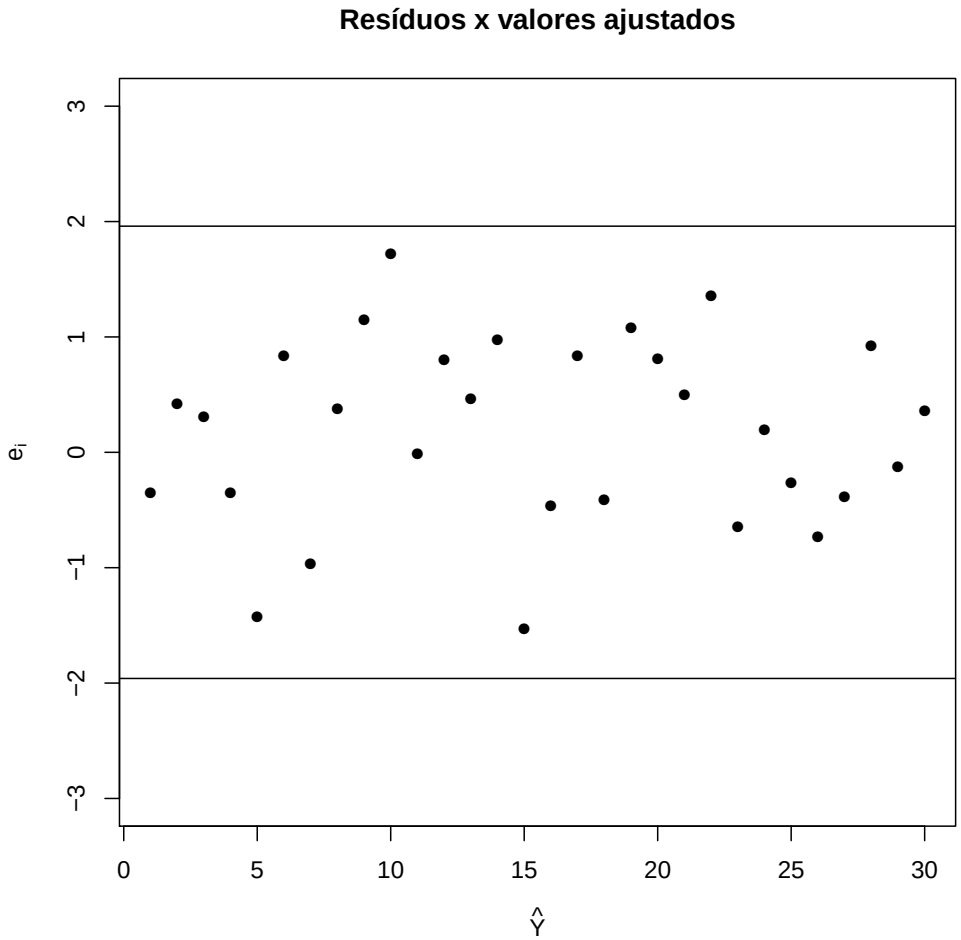
Como r_i não é independente de s^2, então uma alternativa seria considerar

$$t_i^\star = \frac{r_i}{s_{(i)}(1-h_{ii})^{1/2}}.$$

Sob a suposição de normalidade dos erros,

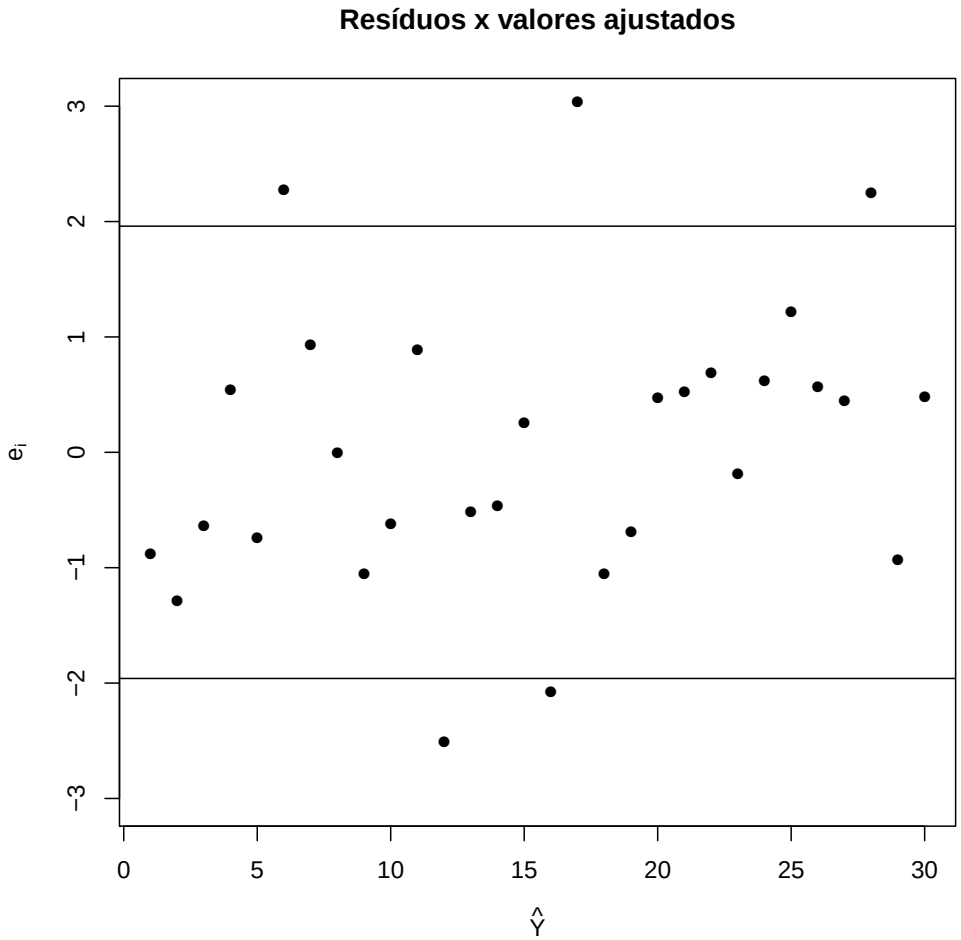
$$t_i^\star \sim t_{(n-p-1)}.$$

Verificando a presença de padrões nos resíduos



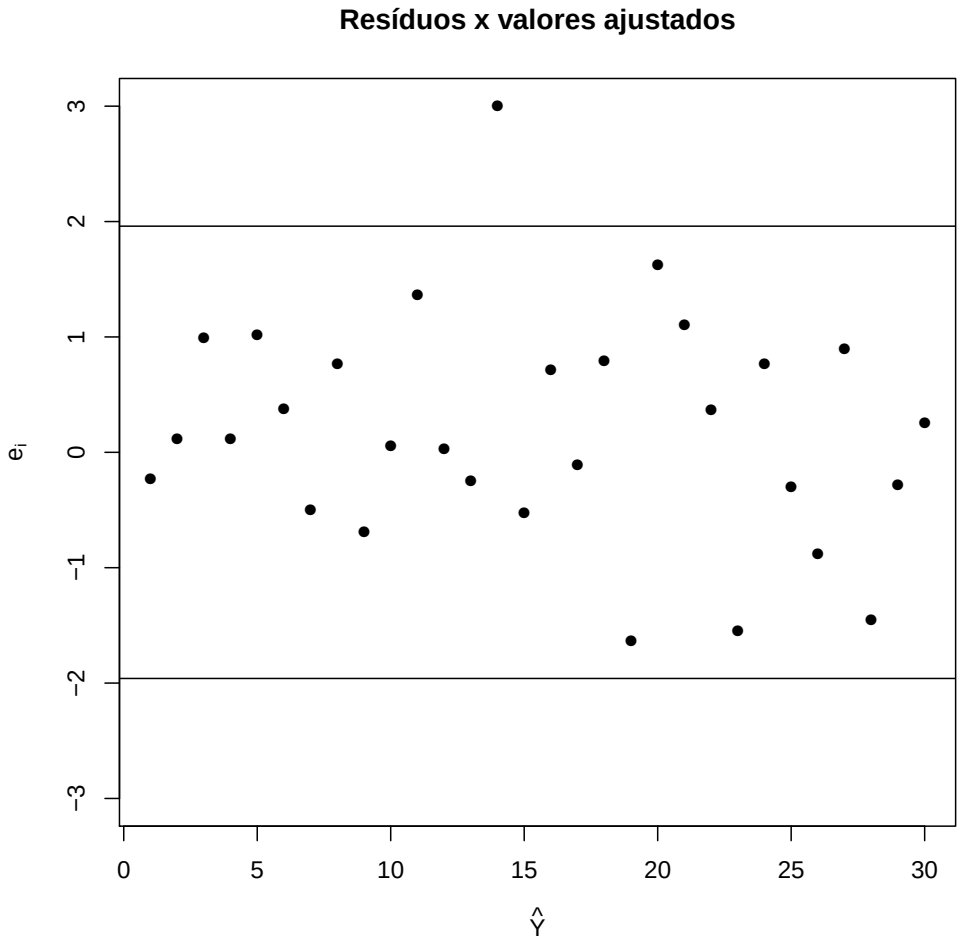
Padrão desejável nos resíduos

Verificando a presença de padrões nos resíduos



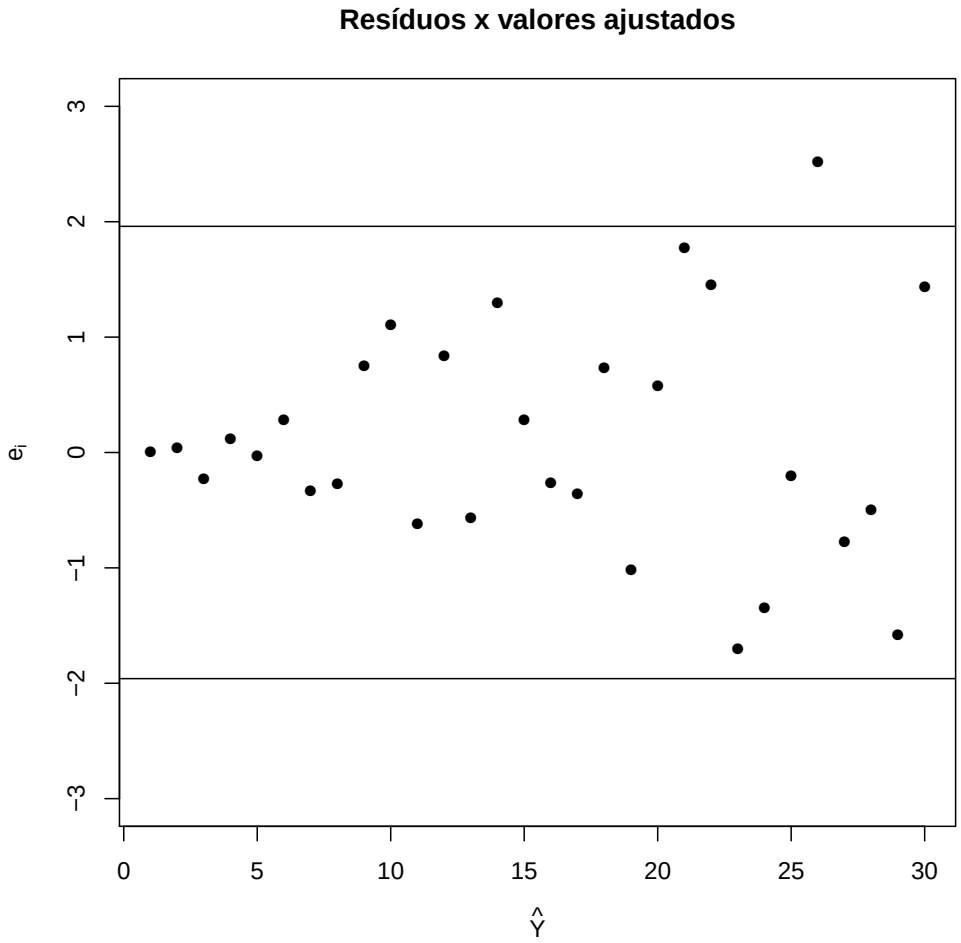
Indícios da presença de outliers ou caudas pesadas

Verificando a presença de padrões nos resíduos



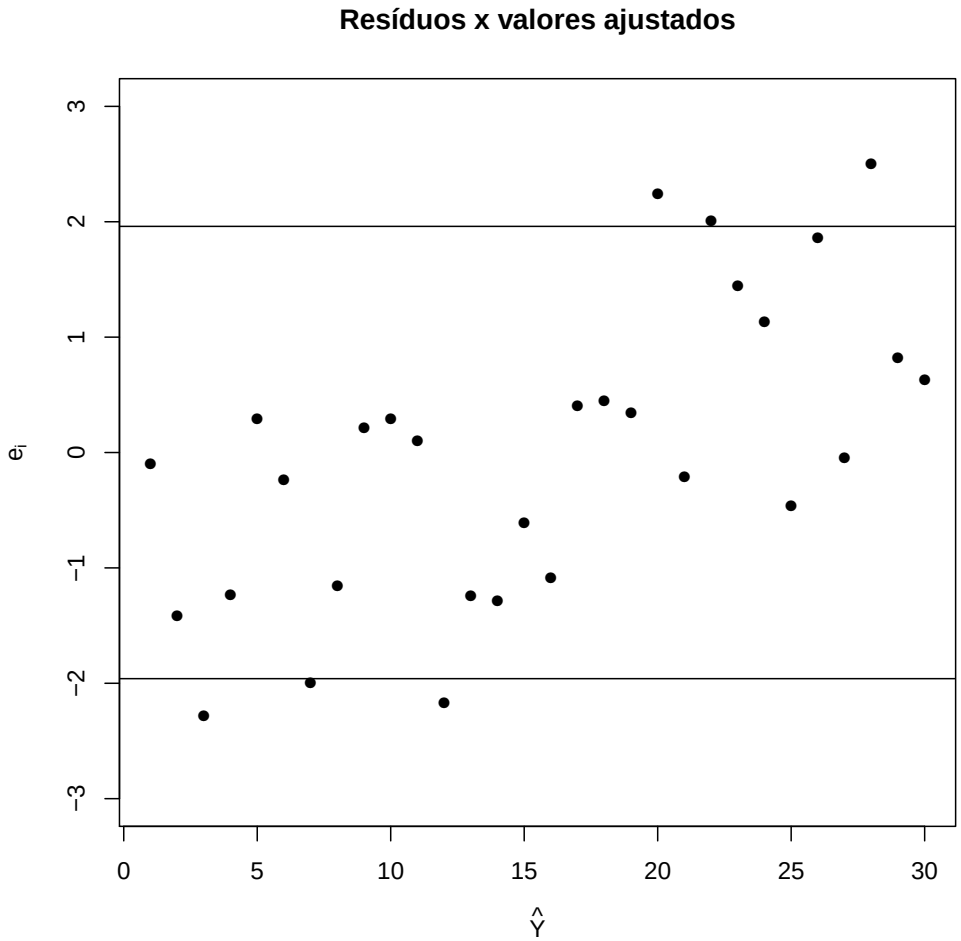
Indícios da presença de um outlier nos dados

Verificando a presença de padrões nos resíduos



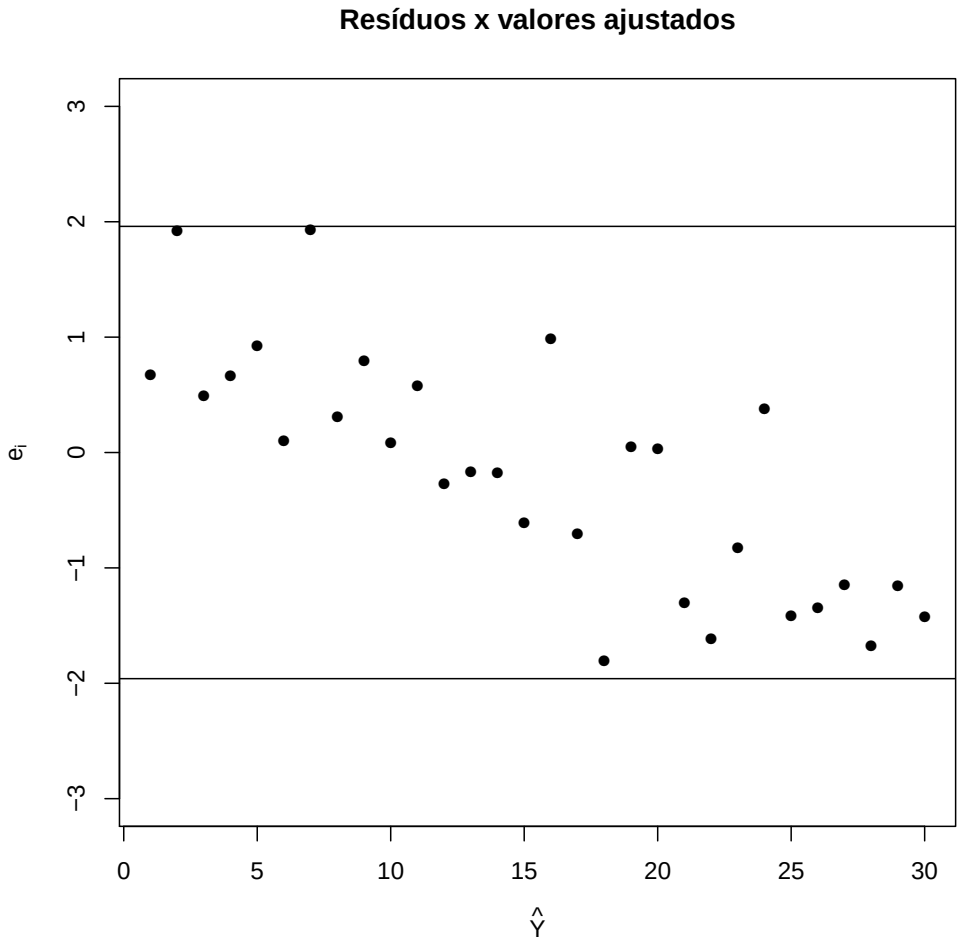
Indícios de heteroscedasticidade de variâncias dos erros

Verificando a presença de padrões nos resíduos



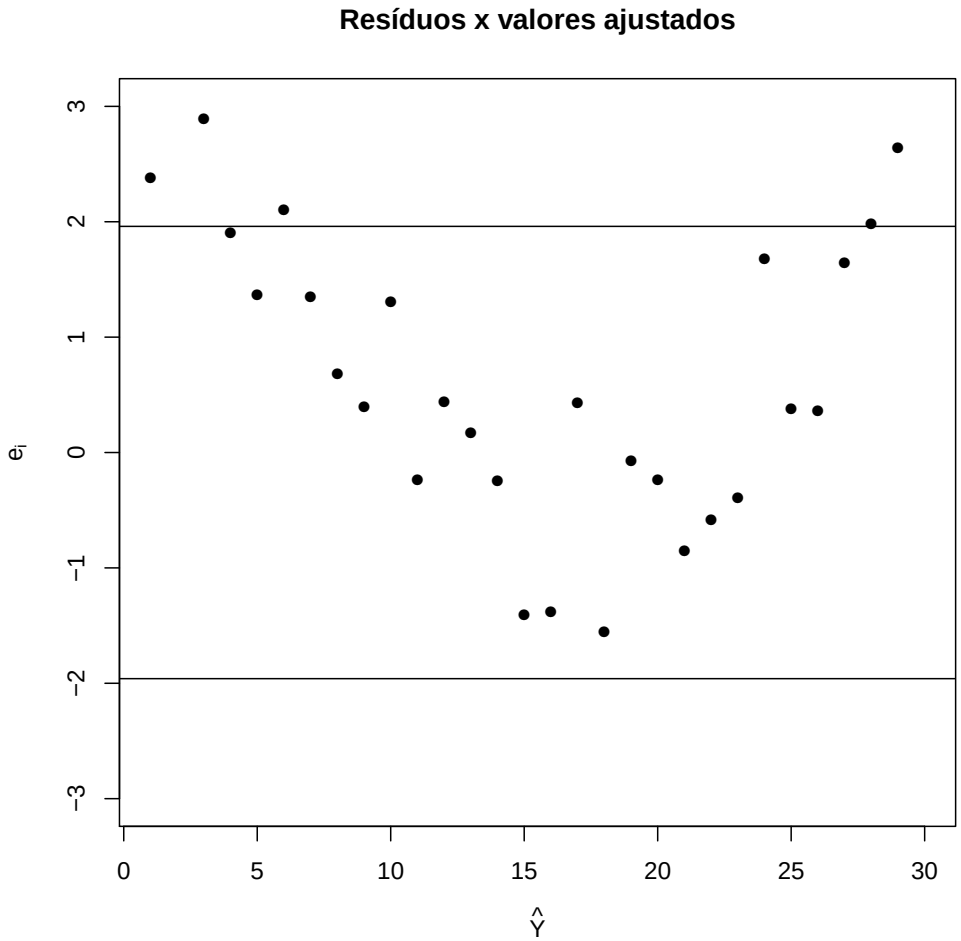
Indícios de padrão crescente - adicionar termo de primeira ordem no tempo ou outra covariável omitida

Verificando a presença de padrões nos resíduos



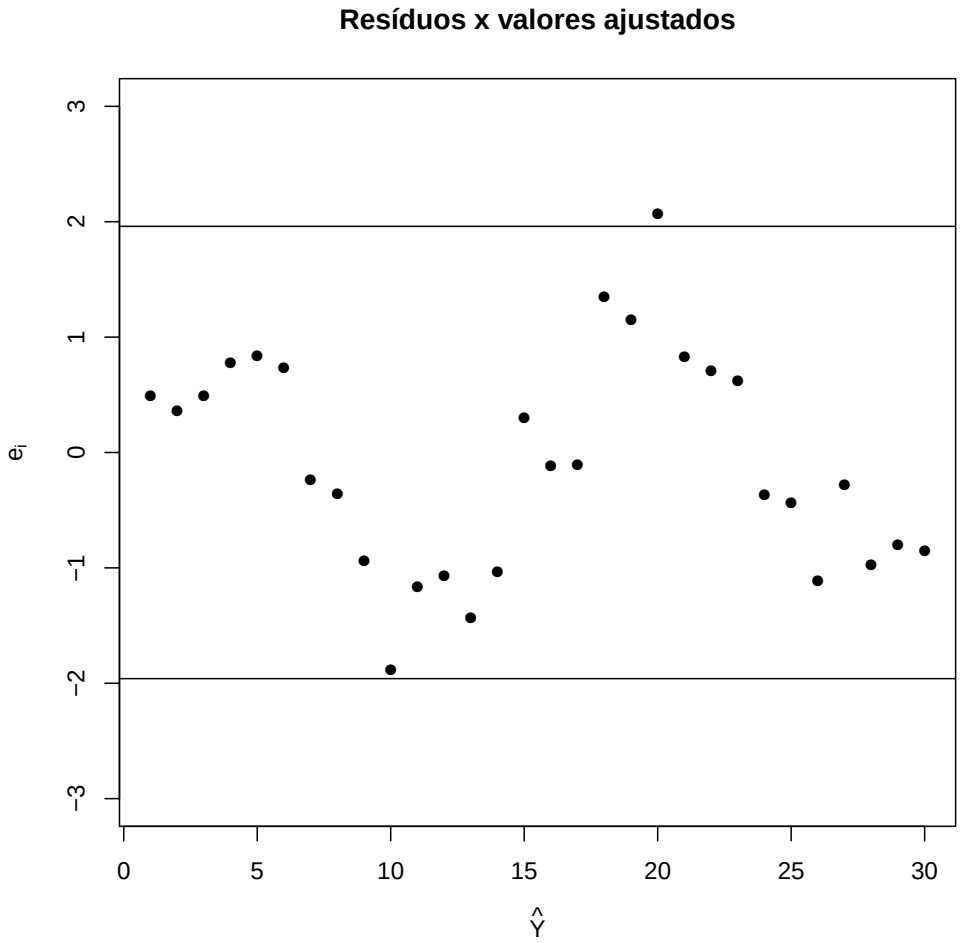
Indícios de padrão decrescente - adicionar termo de primeira ordem no tempo ou outra covariável omitida

Verificando a presença de padrões nos resíduos

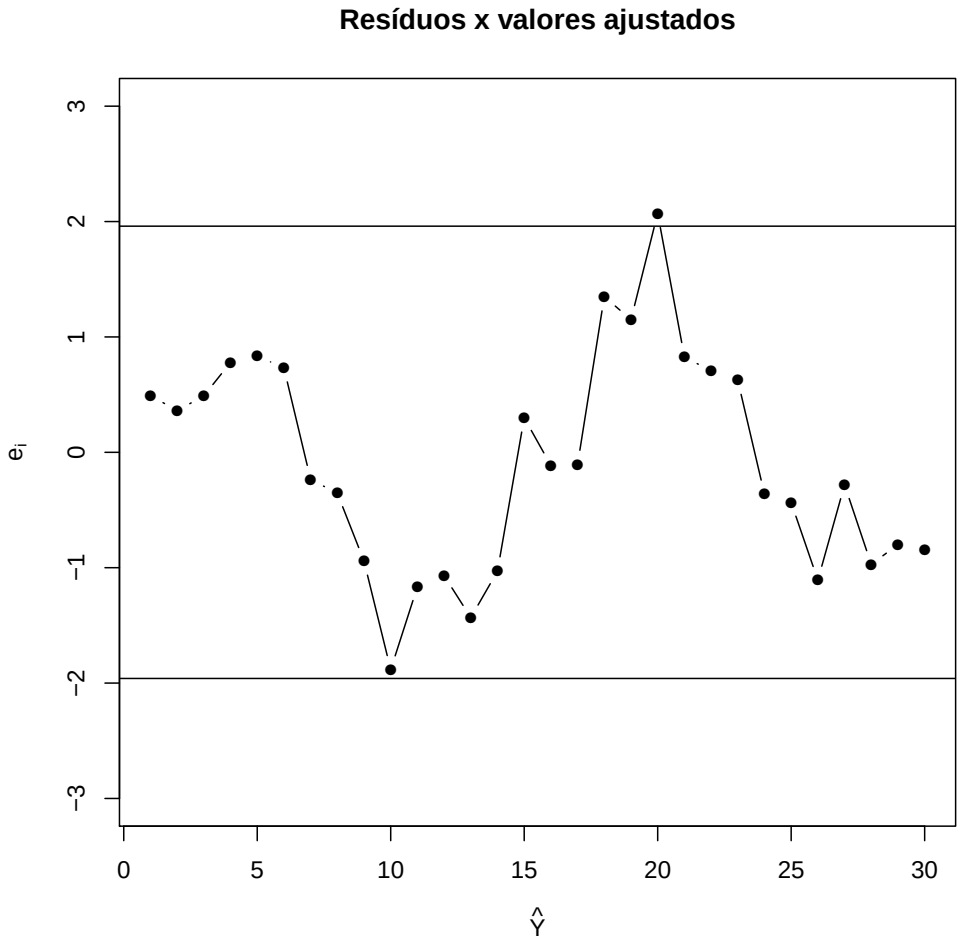


Indícios de padrão quadrático - adicionar termo quadrático no tempo ou outra covariável omitida

Verificando a presença de padrões nos resíduos

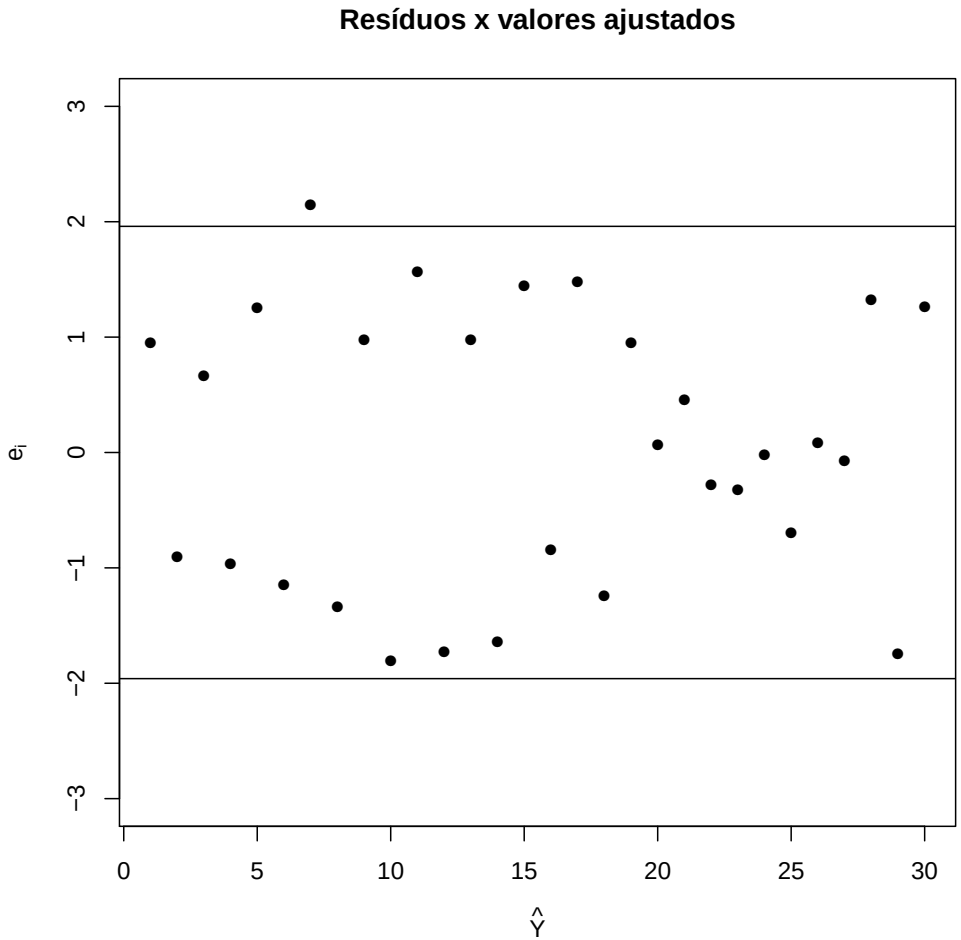


Verificando a presença de padrões nos resíduos

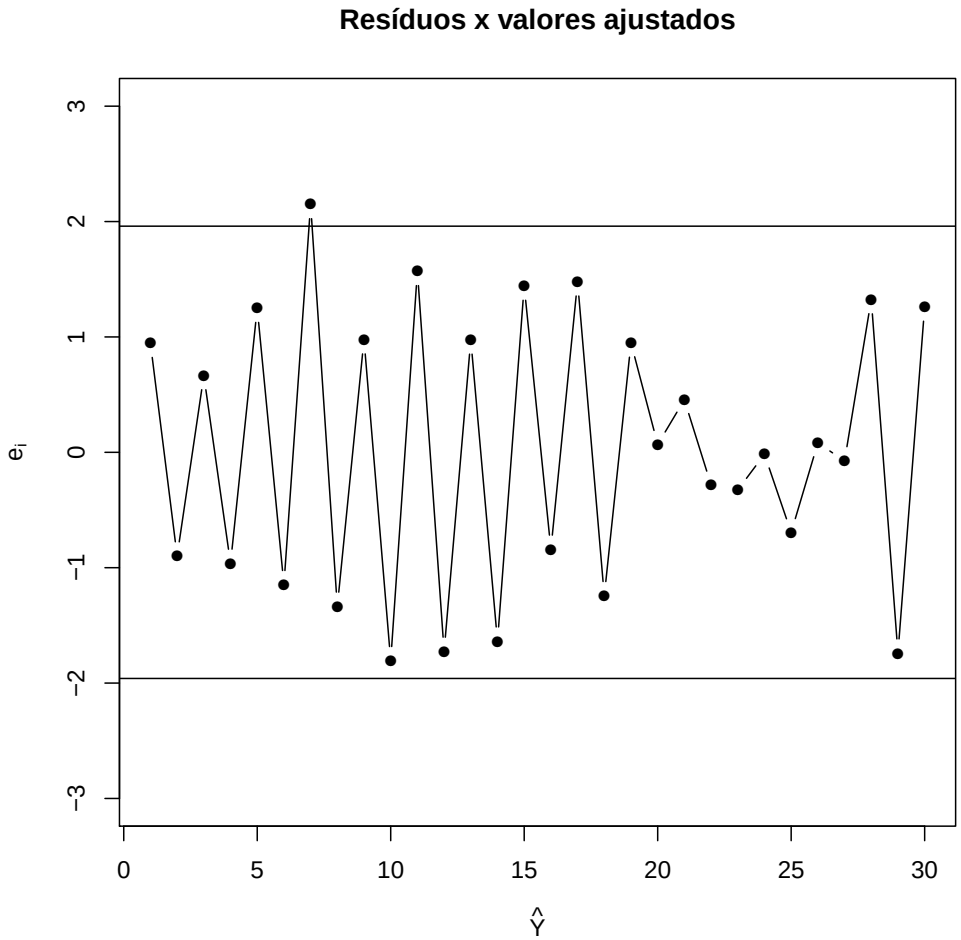


Indícios de possível autocorrelação positiva

Verificando a presença de padrões nos resíduos



Verificando a presença de padrões nos resíduos



Indícios de possível autocorrelação negativa

Diagnóstico

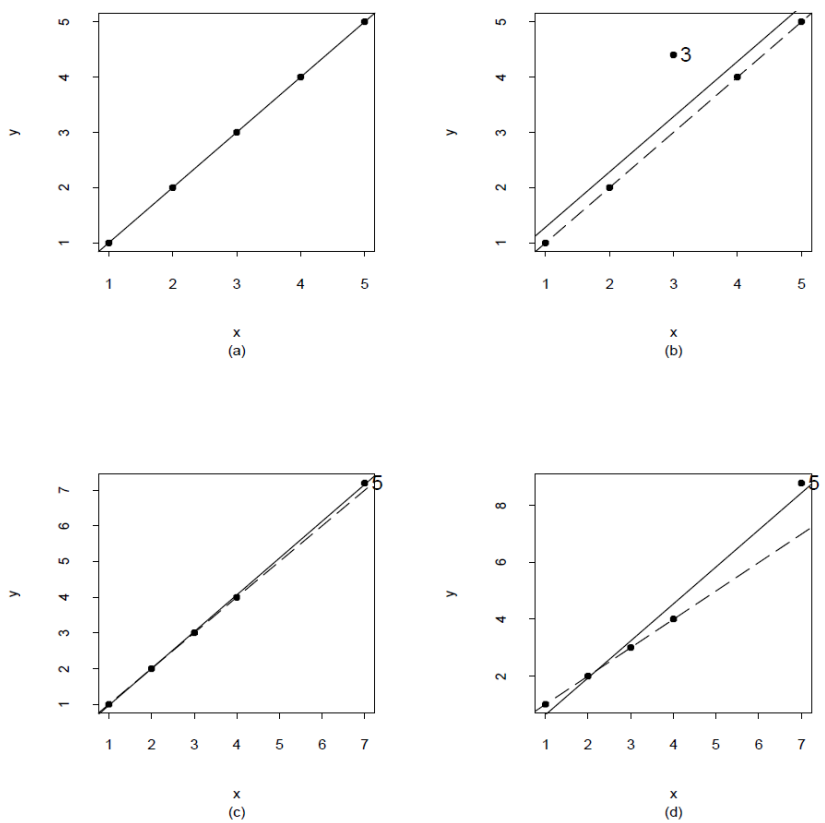


Ilustração de pontos aberrante (b), de alavanca (c) e influente (d).

Paula, G. A. 2012. Modelos de Regressão com apoio computacional. IME USP. Disponível em

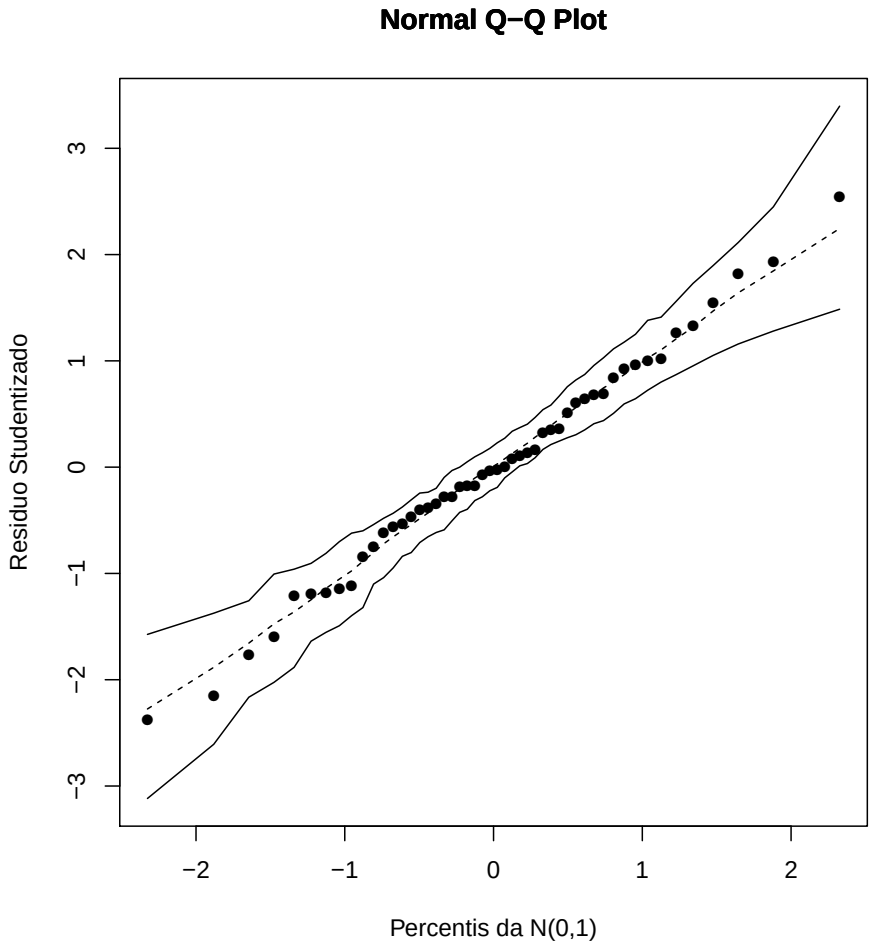
http://www.ime.usp.br/~giapaula/

Envelopes

Envelopes no modelo de regressão linear sob normalidade

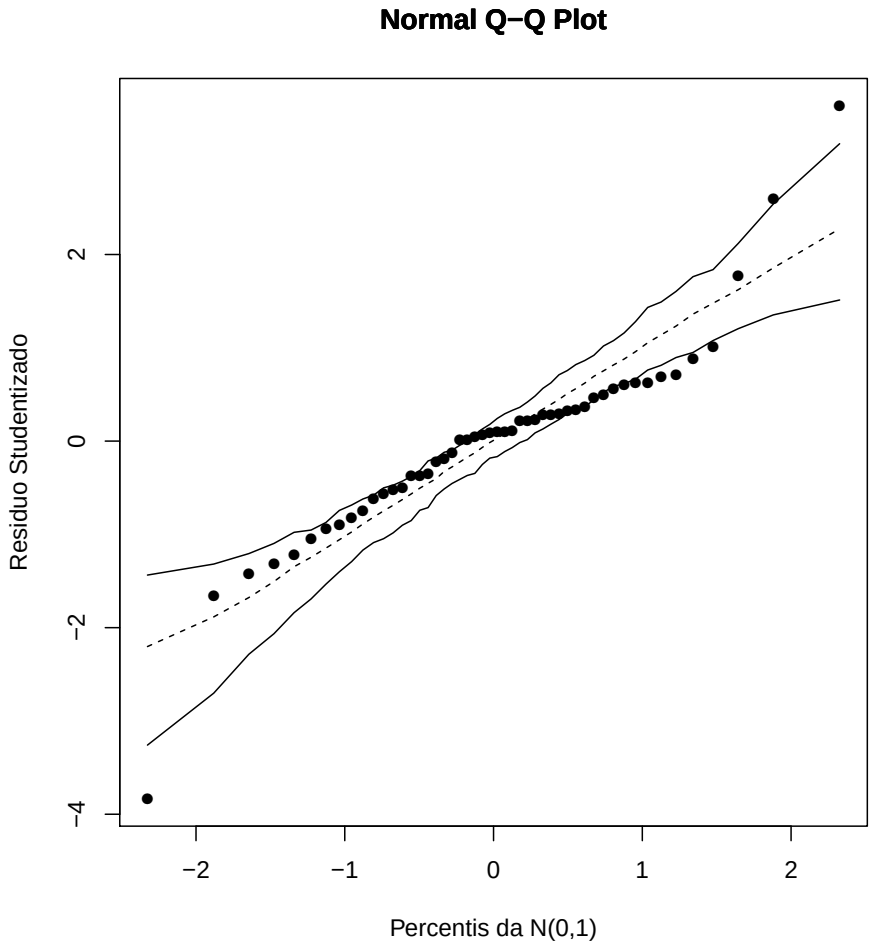
`http://www.ime.usp.br/~giapaula/envel_norm`

Envelopes para os resíduos



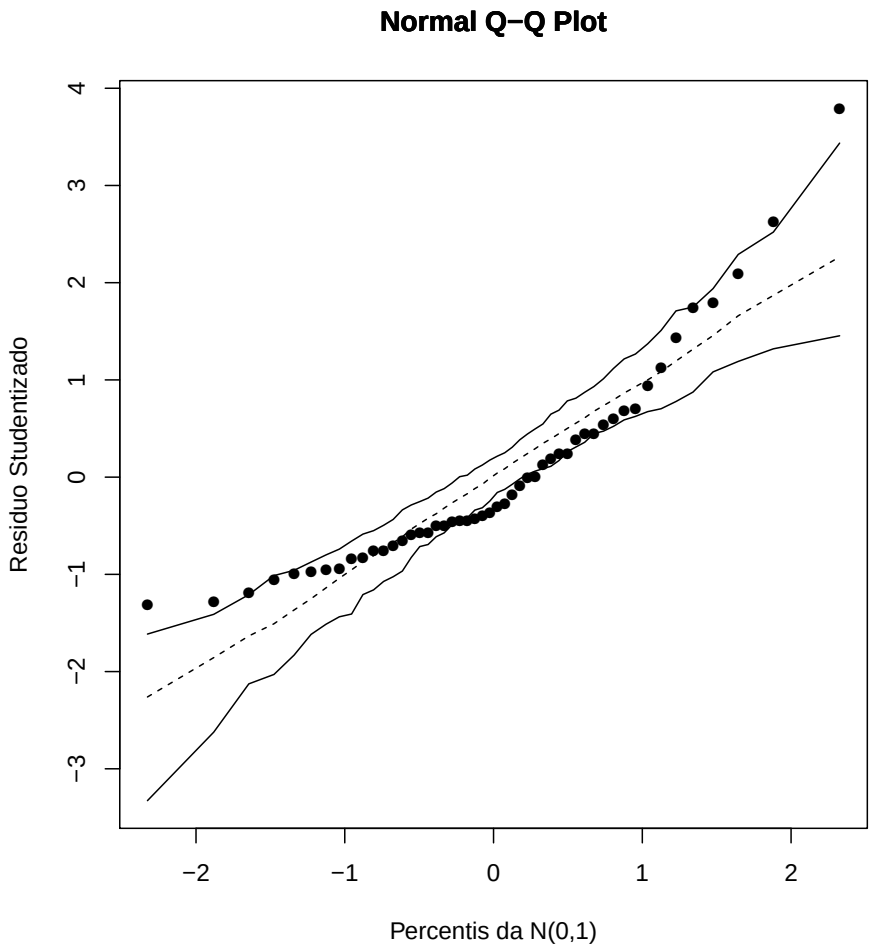
Indícios de adequabilidade

Envelopes para os resíduos



Indícios de caudas pesadas

Envelopes para os resíduos



Indícios de assimetria

Análise de resíduos em modelos lineares com efeitos mistos

Para modelos mistos, o trabalho de Nobre e Singer (2007) propõe medidas residuais para modelos com efeitos mistos.

1. Resíduo marginal

$$e_{ij} = Y_{ij} - \widehat{\mathbf{Y}}_{ij}, \quad i = 1, \ldots, n.$$

2. Resíduo condicional

$$ec_{ij} = y_{ij} - \widehat{\mathbf{Y}}_{ij} - \mathbf{Z}_{ij}^{\top}\widehat{\mathbf{b}}_i.$$

3. Distância de Mahalanobis

$$u_i = [\mathbf{Y}_i - \widehat{\mathbf{Y}}]^{\top}\boldsymbol{\Sigma}_i^{-1}[\mathbf{Y}_i - \widehat{\mathbf{Y}}], \quad i = 1, \ldots, n.$$

(Nobre, J. S. and Singer, J. M., (2007) Residual analysis for linear mixed models. Biometrical Journal, 49, 863–875)

Análise de resíduos em modelos lineares com efeitos mistos

Resultado padronizado

Uma proposta simples para padronizar resíduos em modelos lineares mistos é

$$e_{ij} = \frac{y_{ij} - \widehat{Y}_{ij}}{\widehat{\sigma}}, \ i = 1, \ldots, n; j = 1, \ldots, m_i.$$

Espera-se que não haja padrões como os de modelos lineares no gráfico de resíduos $e_{ij} \times \widehat{Y}_{ij}$.

(Pinheiro and Bates (2009) 'Mixed-effects Models in S and S-PLUS', Springer.)

Exercício: Dados ortodônticos

Modelo linear com efeitos mistos

Exercício Nos modelos ajustados para os dados ortodôntico, verifique a presença de possíveis padrões nos gráficos de resíduos $e_{ij} \times \widehat{Y}_{ij}$. Identifique possíveis observações atípicas.

Exemplo: Dados ortodônticos

Modelo linear com efeitos mistos

(1 covariável, intercepto aleatório)

Comandos em R:

```
> library(nlme)
> attach(Orthodont)
> fit.IA<-lme(distance~ age,data=Orthodont,random=~
-1+age|Subject)
> plot(fit.IA)
> plot(augPred(fit.IA),aspect="xy",grid=T)
> qqnorm(fit.IA,~ resid(.) |Sex)
```

Seleção de modelos: Critérios de informação

Os critérios de informação mais utilizados em modelos com efeitos mistos são

- AIC (Critério de informação de Akaike)
- BIC (critério de Informação de Schwartz ou critério de informação Bayesiano)

Critérios de informação

As expressões para AIC e BIC são

$$AIC = 2p - 2\log L(\widehat{\boldsymbol{\theta}})$$

com p o número de parâmetros do modelo ajustado.

$$BIC = 2p\log(N) - 2\log L(\widehat{\boldsymbol{\theta}})$$

com p o número de parâmetros do referido modelo e N o número total de observações.

Quanto menor AIC e BIC, melhor o modelo.

Análise de diagnóstico: Análise de resíduos

Exercício:

Nos dados ortodônticos, selecione o melhor entre os modelos ajustados.

Discussão

Outros modelos e análises possíveis:

- Modelos com efeitos mistos com estruturas mais gerais para a matriz de variâncias e covariâncias dos erros aleatórios
- Modelos mistos com distribuições mais gerais para efeitos e erros aleatórios
- Modelos semiparamétricos com efeitos mistos
- Modelos conjuntos (joint models)
- etc

E aqui finalizamos nosso curso...

Obrigada pela presença!