MAT02035 - Modelos para dados correlacionados

Análise de resíduos e diagnóstico

Rodrigo Citton P. dos Reis
citton.padilha@ufrgs.br

Universidade Federal do Rio Grande do Sul
Instituto de Matemática e Estatística
Departamento de Estatística

Porto Alegre, 2021

Instituto de MATEMÁTICA E ESTATÍSTICA UFRGS

Introdução

Introdução

- A análise de dados longitudinais não fica completa sem uma examinação dos **resíduos**.
  - Ou seja, a verificação das suposições impostas ao modelo e ao processo de inferência.
- As ferramentas usuais de análise de resíduos para a regressão convencional (com observações independentes) podem ser estendidas para a estrutura longitudinal.

Resíduos

Resíduos

- Defina o **vetor de resíduos** para cada indivíduo

$$r_i = Y_i - X_i\hat{\beta}.$$

- O vetor de resíduos tem média zero e fornece uma estimativa para o vetor de erros

$$e_i = Y_i - X_i\beta.$$

Resíduos

- Os resíduos r_i podem ser usados para **verificar** se há **desvios sistemáticos** do modelo quanto à **resposta média**;
  - eles também podem formar a base de uma **avaliação da adequação do modelo para a covariância**.
- Por exemplo, um gráfico de dispersão dos resíduos $r_{ij} = Y_{ij} - X'_{ij}\hat{\beta}$ *versus* a resposta média predita $\hat{\mu}_{ij} = X'_{ij}\hat{\beta}$ pode examinar a presença de qualquer tendência sistemática.
  - Em um **modelo especificado corretamente**, o gráfico de dispersão não deve exibir um padrão sistemático, com uma **dispersão** mais ou menos **aleatória** em torno de uma média constante zero.

Resíduos

Para propósitos mais práticos,os gráficos de resíduos podem ser usados para detectar discrepâncias no modelo para a resposta média ou a presença de observações distantes ("outliers") que requerem investigação adicional.

No entanto, existem duas propriedades dos resíduos de uma análise de dados longitudinais que devem ser lembradas.

1. Os componentes do vetor de resíduos r_i e não tem necessariamente variância constante.
   - Como resultado, diagnósticos de resíduos padrão para examinar a homogeneidade da variância residual ou a autocorrelação entre os resíduos **devem ser totalmente evitados**.
2. Embora os resíduos de uma regressão linear univariada não estejam correlacionados com as covariáveis, os resíduos de uma análise de regressão de dados longitudinais podem estar correlacionados com as covariáveis.
   - Como resultado, pode haver uma tendência sistemática aparente no gráfico de dispersão dos resíduos contra uma covariável selecionada.

Resíduos transformados

Resíduos transformados

- Para contornar alguns dos problemas mencionados com o uso de resíduos a partir de dados longitudinais com base em r_i, podemos **transformar os resíduos**.
- Há muitas possibilidades para transformar os resíduos.
- A transformação deve ser realizada de forma que os resíduos "imitem" aqueles da regressão linear padrão.
- Os resíduos r_i^* definidos a seguir são não-correlacionados e têm variância unitária:

$$r_i^* = L_i^{-1} r_i = L_i^{-1}(Y_i - X_i\hat{\beta}),$$

em que Li é a matriz triangular superior resultante da **decomposição de Cholesky** da matriz de covariâncias estimada $\hat{\Sigma}_i$, ou seja, $\hat{\Sigma}_i = L_i L_i'$.

Resíduos transformados: gráficos

- Dado o conjunto de resíduos transformados, r_i^*, todos os diagnósticos de resíduos usuais para regressão linear padrão podem ser aplicados.
- Por exemplo, podemos construir um gráfico de dispersão dos resíduos transformados, r_{ij}^*, *versus* os valores preditos transformados, μ_{ij}^*, em que

$$\mu_i^* = L_i^{-1}\hat{\mu}_i = L_i^{-1}X_i\hat{\beta}.$$

- Em um modelo especificado corretamente, esse gráfico de dispersão não deve exibir um padrão sistemático, com uma dispersão aleatória em torno de uma média constante zero e com um intervalo constante para variados μ_{ij}^*.

Resíduos transformados: gráficos

- Da mesma forma, podemos construir um gráfico de dispersão dos resíduos transformados *versus* covariáveis transformadas selecionadas.
- Com dados longitudinais, um gráfico de dispersão dos resíduos transformados *versus* o tempo transformado (ou idade) pode ser particularmente útil para avaliar a adequação das premissas do modelo sobre padrões de mudança na resposta média ao longo do tempo.
- Por fim, os resíduos transformados também tornam um pouco mais fácil identificar assimetria e possíveis observações atípica que requerem maior investigação.

Resíduos transformados: gráficos

- Um **gráfico de quantil normal** (ou o chamado quantil-quantil ou gráfico de Q-Q) dos resíduos transformados pode ser usado para avaliar a suposição de distribuição normal e identificar valores extremos.
  - Ou seja, com base nas posições (*ranks*) dos resíduos transformados, podemos gerar um gráfico dos quantis amostrais dos resíduos contra os quantis esperados se eles tiverem uma distribuição normal.
  - Se os resíduos se afastam discernivelmente de uma linha reta, a suposição de normalidade pode não ser sustentável.
  - A assimetria é geralmente indicada por um padrão em forma de arco no gráfico quantil normal; os outliers aparecerão como "retardatários", longe dos fins da linha reta.

Semi-variograma

Semi-variograma

- Historicamente, o **semi-variograma** tem sido amplamente utilizado em **estatística espacial** para representar a estrutura de covariância em dados geoestatísticos.
- Ao contrário dos dados espaciais bidimensionais, as coordenadas dos dados longitudinais são ao longo de uma única dimensão, a saber, o tempo.
- Para dados longitudinais, o semi-variograma é definido como **metade do quadrado da diferença esperada** entre os resíduos obtidos no mesmo indivíduo.

Semi-variograma

- O semi-variograma, denotado por $\gamma(h_{ijk})$, é dado por

$$\gamma(h_{ijk}) = \frac{1}{2}\mathrm{E}\,(r_{ij} - r_{ik})^2,$$

em que h_{ijk} é tempo decorrido entre a j-ésima e a k-ésima medidas repetidas do i-ésimo indivíduo.

- Como os resíduos tem média zero, o semi-variograma pode ser expresso como

$$\begin{aligned}
\gamma(h_{ijk}) &= \frac{1}{2}\mathrm{E}\,(r_{ij} - r_{ik})^2 \\
&= \frac{1}{2}\mathrm{E}\,(r_{ij}^2 + r_{ik}^2 - 2r_{ij}r_{ik}) \\
&= \frac{1}{2}\mathrm{Var}\,(r_{ij}) + \frac{1}{2}\mathrm{Var}\,(r_{ik}) - \mathrm{Cov}\,(r_{ij}, r_{ik}).
\end{aligned}$$

Semi-variograma

- Embora o semi-variograma possa ser usado para sugerir modelos apropriados para a covariância, aqui simplesmente o usamos como uma ferramenta de diagnóstico para avaliar a adequação de um modelo selecionado para a covariância.
- Quando o semi-variograma é aplicado aos resíduos transformados, r_{ij}^*, ele simplifica a

$$\gamma(h_{ijk}) = \frac{1}{2}\text{Var}\,(r_{ij}^*) + \frac{1}{2}\text{Var}\,(r_{ik}^*) - \text{Cov}\,(r_{ij}^*, r_{ik}^*) = \frac{1}{2}(1) + \frac{1}{2}(1) - 0 = 1.$$

Semi-variograma

- Assim, em um modelo especificado corretamente para a covariância, o gráfico do semi-variograma para os resíduos transformados *versus* o tempo decorrido entre as observações correspondentes deve flutuar aleatoriamente em torno de uma linha horizontal centralizada em 1.

Exemplo

Exemplo: Influência da menarca nas mudanças do percentual de gordura corporal

- Estudo prospectivo do aumento de gordura corporal em uma coorte de 162 garotas.
- Sabe-se que o percentual de gordura nas garotas tem um aumento considerável no período em torno da menarca (primeira menstruação).
- Parece que este aumento continua significativo por aproximadamente quatro anos depois da menarca, mas este comportamento ainda não foi devidamente estudado.
- As meninas foram acompanhadas até quatro anos depois da menarca.

Exemplo

- Há um total de 1049 medidas, com uma média de 6,4 medidas por menina.
- Variáveis do estudo:
  - Resposta: Percentual de gordura corporal;
  - Covariáveis: Tempo em relação à menarca (idade da menina no instante observado menos idade quando teve a menarca) - pode ser positivo ou negativo.

Exemplo

```
# ---------------------------------------------------
# Carregando pacotes do R
library(here)
library(haven)
library(tidyr)
library(ggplot2)
library(dplyr)
# ---------------------------------------------------
# Carregando o arquivo de dados
fat <- read_dta(
  file = here::here("data", "fat.dta"))
fat
```

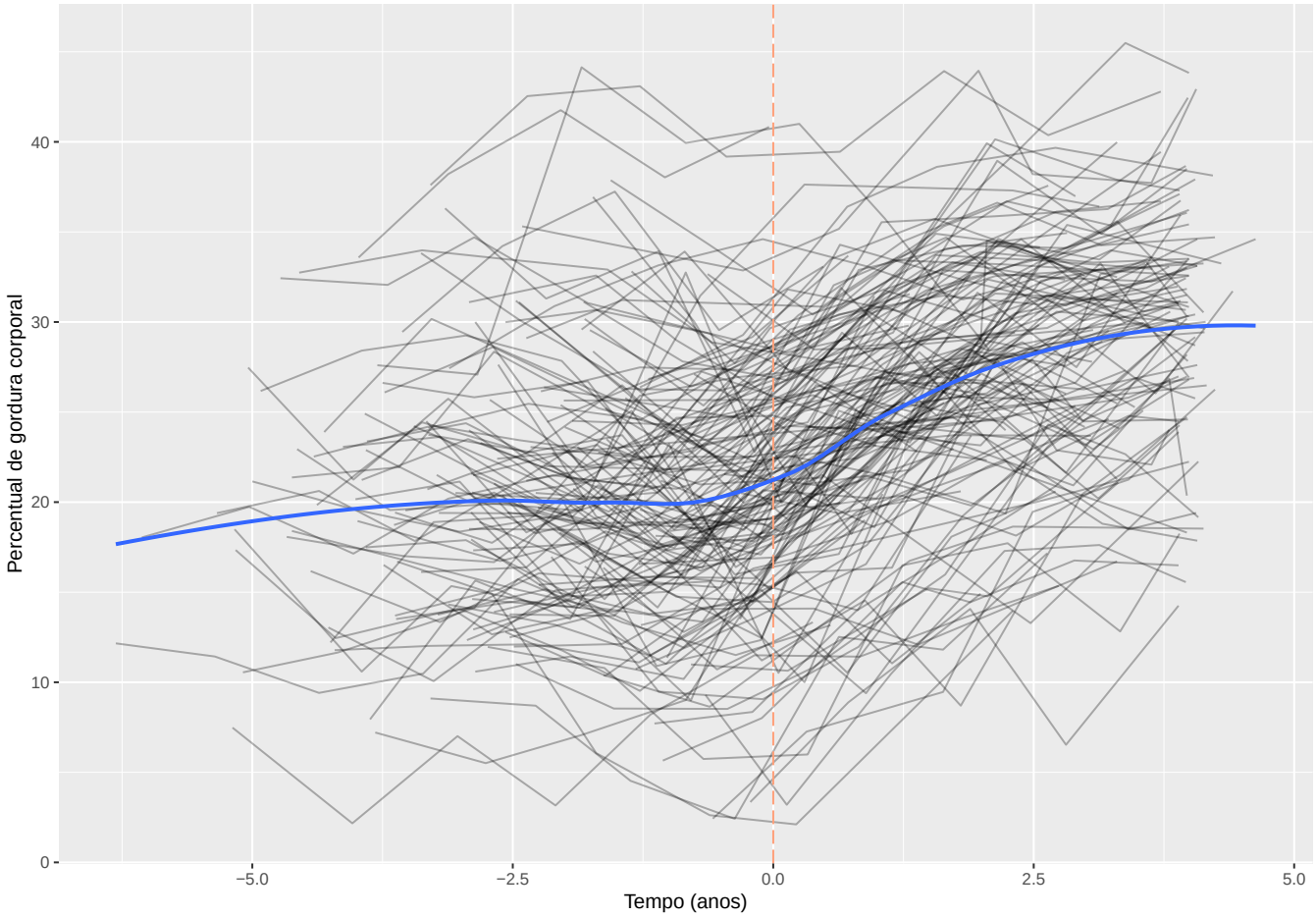
Exemplo

```
## # A tibble: 1,049 x 5
##       id   age agemen    time   pbf
##    <dbl> <dbl>  <dbl>   <dbl> <dbl>
##  1     1  9.32   13.2 -3.87    7.94
##  2     1 10.3    13.2 -2.86   15.6
##  3     1 11.2    13.2 -1.95   13.5
##  4     1 12.2    13.2 -1      23.2
##  5     1 13.2    13.2  0.0500 10.5
##  6     1 14.2    13.2  1.05   20.5
##  7     2  8.84   13.3 -4.44   16.2
##  8     2 10.1    13.3 -3.20   13.3
##  9     2 11.0    13.3 -2.25   16.0
## 10     2 12.8    13.3 -0.510  15.3
## # ... with 1,039 more rows
```

Exemplo

```
p <- ggplot(data = fat,
            mapping = aes(x = time, y = pbf,
                          group = id)) +
  geom_line(alpha = 0.3) +
  geom_vline(xintercept = 0,
             colour = "lightsalmon",
             linetype = "longdash") +
  labs(x = "Tempo (anos)",
       y = "Percentual de gordura corporal")
p + geom_smooth(data = fat,
                mapping = aes(x = time, y = pbf,
                              group = NULL),
                method = "loess", se = FALSE)
```

Exemplo


Percentual de gordura corporal
40
30
20
10
0
−5.0
−2.5
0.0
2.5
5.0
Tempo (anos)

Exemplo

- O modelo inicialmente proposto considera que cada garota tem uma curva de crescimento *spline* linear com um *knot* no tempo da menarca.
- Ajustou-se o seguinte modelo linear de efeitos mistos:

$$\mathrm{E}\left(Y_{ij}|b_i\right) = \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + b_{1i} + b_{2i} t_{ij} + b_{3i} (t_{ij})_+,$$

em que

$$(t_{ij})_+ = \begin{cases} t_{ij} & \text{se} \quad t_{ij} > 0, \\ 0 & \text{se} \quad t_{ij} \leq 0. \end{cases}$$

Exemplo

- A função `lspline` do pacote de mesmo nome facilita o ajuste do modelo proposto.

```
fat <- as.data.frame(fat)
library(nlme)
library(lspline)
mod1 <- lme(pbf ~ lspline(x = time,
                          knots = 0,
                          marginal = TRUE),
            random = ~ lspline(x = time,
                          knots = 0,
                          marginal = TRUE) | id,
            data = fat)
```

Exemplo

Table 1: Coeficientes de regressão estimados (efeitos fixos) e erros padrões para os o modelo linear por partes para dos dados de percentual de gordura.

	Estimativa	EP	Z
(Intercepto)	21.3614	0.5645	37.84
tempo	0.4171	0.1572	2.65
$(\text{tempo})_+$	2.0471	0.2280	8.98

Exemplo

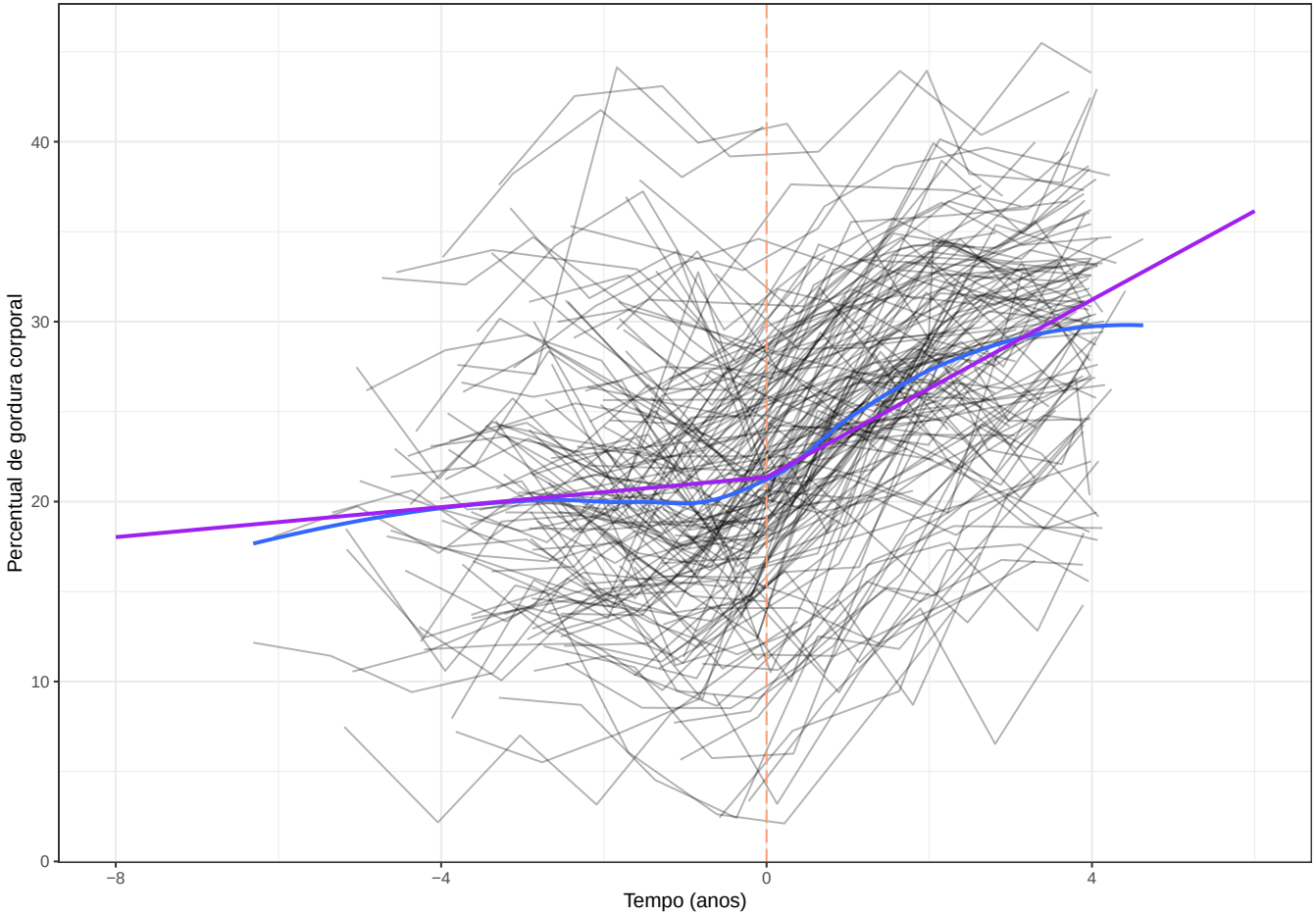
Com o pacote ggeffects é possível visualizar os efeitos marginais do modelo

```
library(ggeffects)

margeff <- ggpredict(model = mod1, terms = "time")

p + geom_smooth(data = fat,
                mapping = aes(x = time, y = pbf,
                              group = NULL),
                method = "loess", se = FALSE) +
  geom_line(data = margeff,
            mapping = aes(x = x, y = predicted,
                          group = NULL),
            colour = "purple", size = 1) +
  theme_bw()
```

Exemplo



Exemplo

Matriz G de covariância de b_i

```
G <- getVarCov(mod1, type = "random.effects")
matrix(G, ncol = 3, byrow = T)
```

```
##             [,1]       [,2]       [,3]
## [1,] 45.939154  2.526047 -6.109079
## [2,]  2.526047  1.630971 -1.750363
## [3,] -6.109079 -1.750363  2.749384
```

```
mod1$sigma^2 # sigma^2 ("R_i")
```

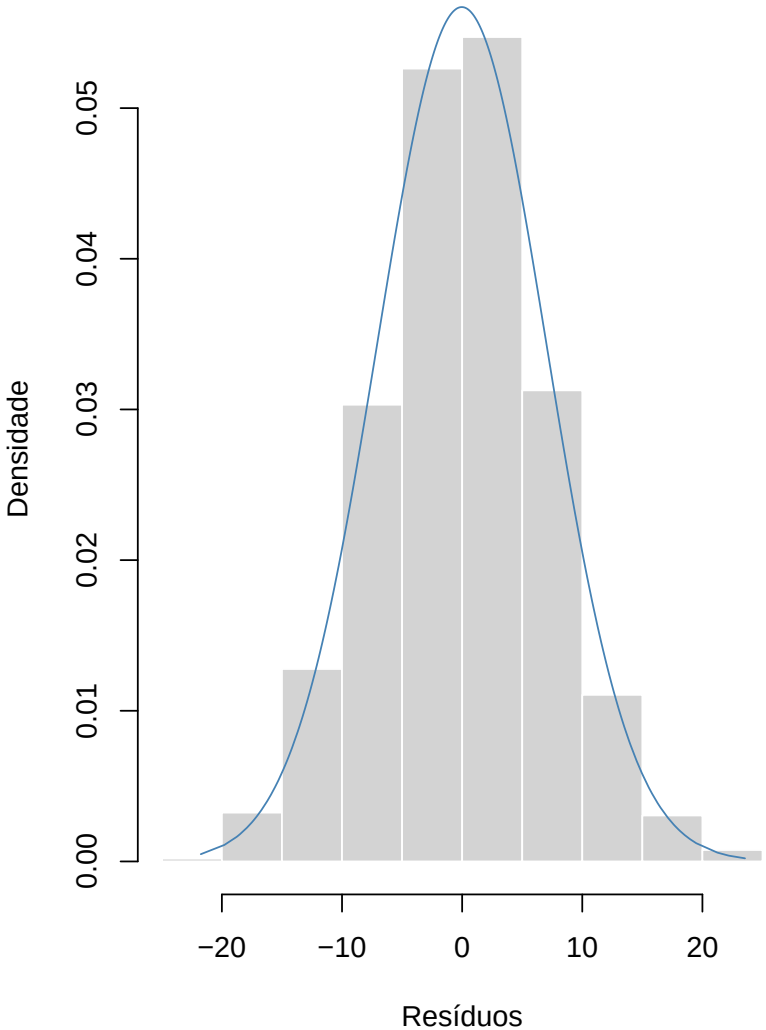
```
## [1] 9.473401
```

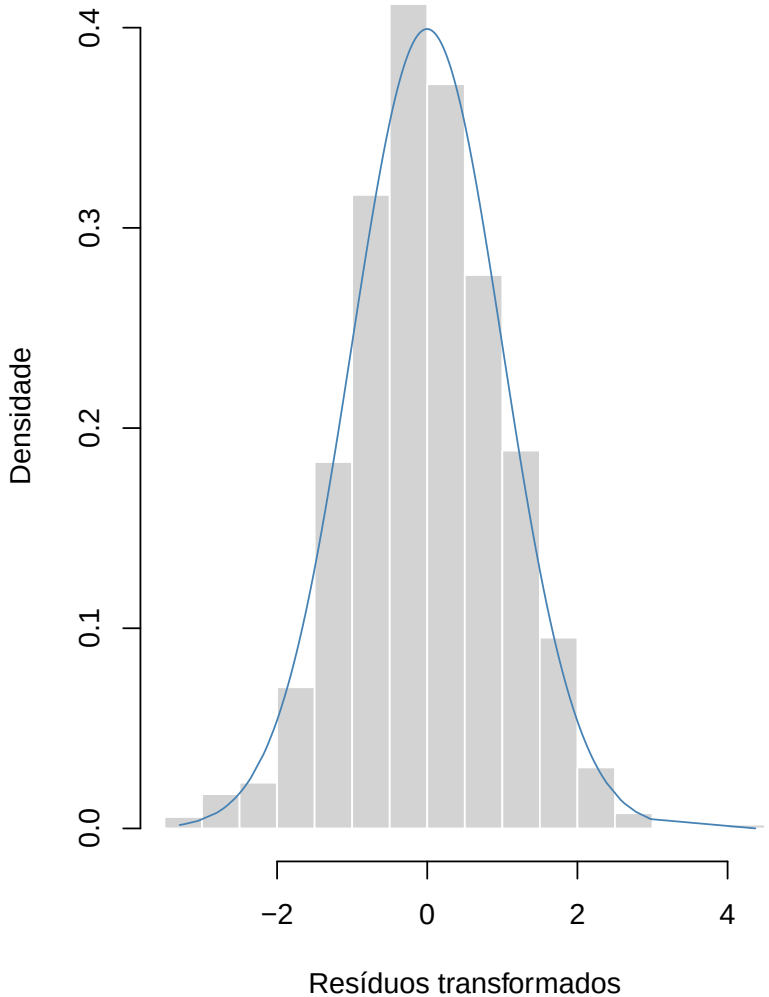
Exemplo

Calculando e transformando os resíduos e valores preditos

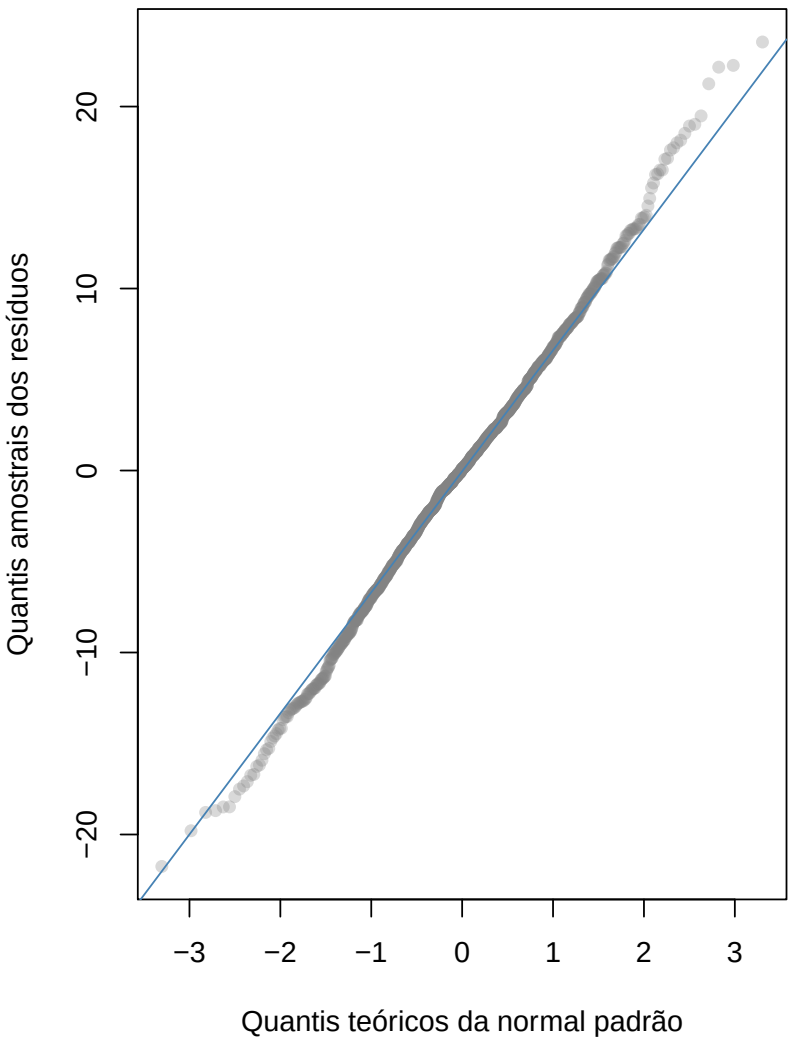
```
library(mgcv)
# Calculate the residuals
res <- residuals(mod1, level = 0)
pred <- fitted(mod1, level = 0)
# Cholesky residuals
est.cov <- extract.lme.cov(mod1, fat) # We extract the blocked
# covariance function of the residuals
Li <- t(chol(est.cov)) # We find the Cholesky transformation
# of the residuals. (The transform is to get the lower
# triangular matrix.)
rest <- solve(Li) %*% res # We then calculate the
# transformed residuals.
predt <- solve(Li) %*% pred
```

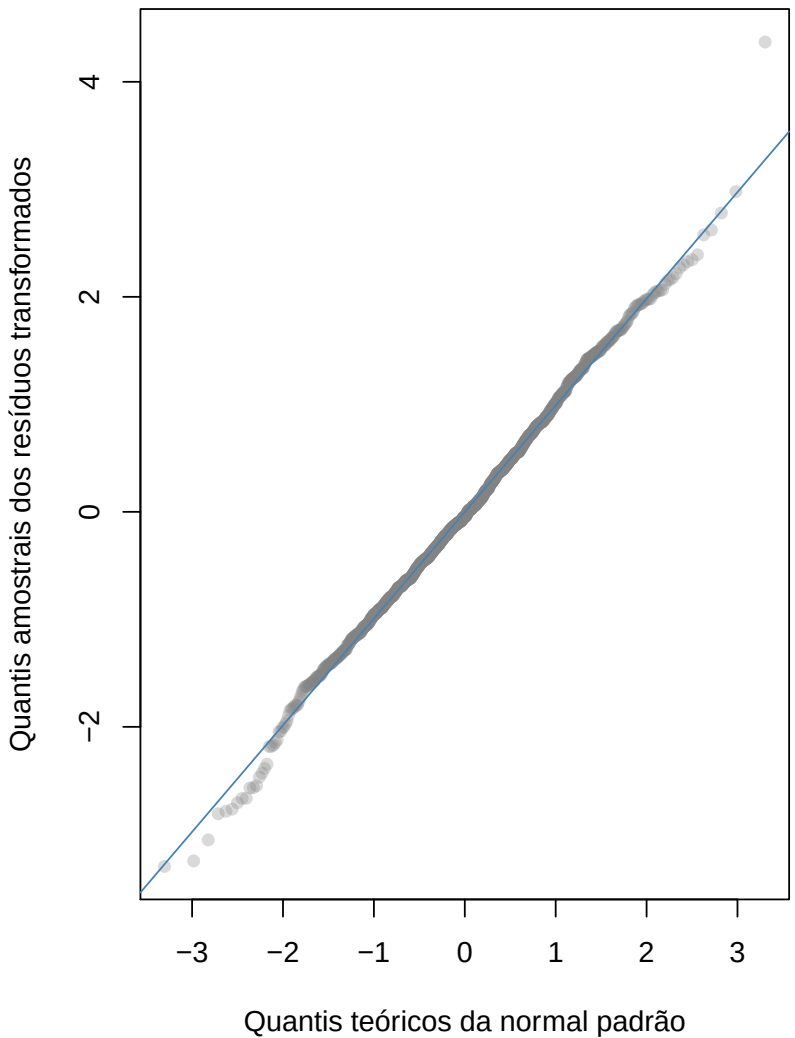
Exemplo


Densidade
0.00
0.01
0.02
0.03
0.04
0.05
−20
−10
0
10
20
Resíduos


Densidade
0.0
0.1
0.2
0.3
0.4
−2
0
2
4
Resíduos transformados

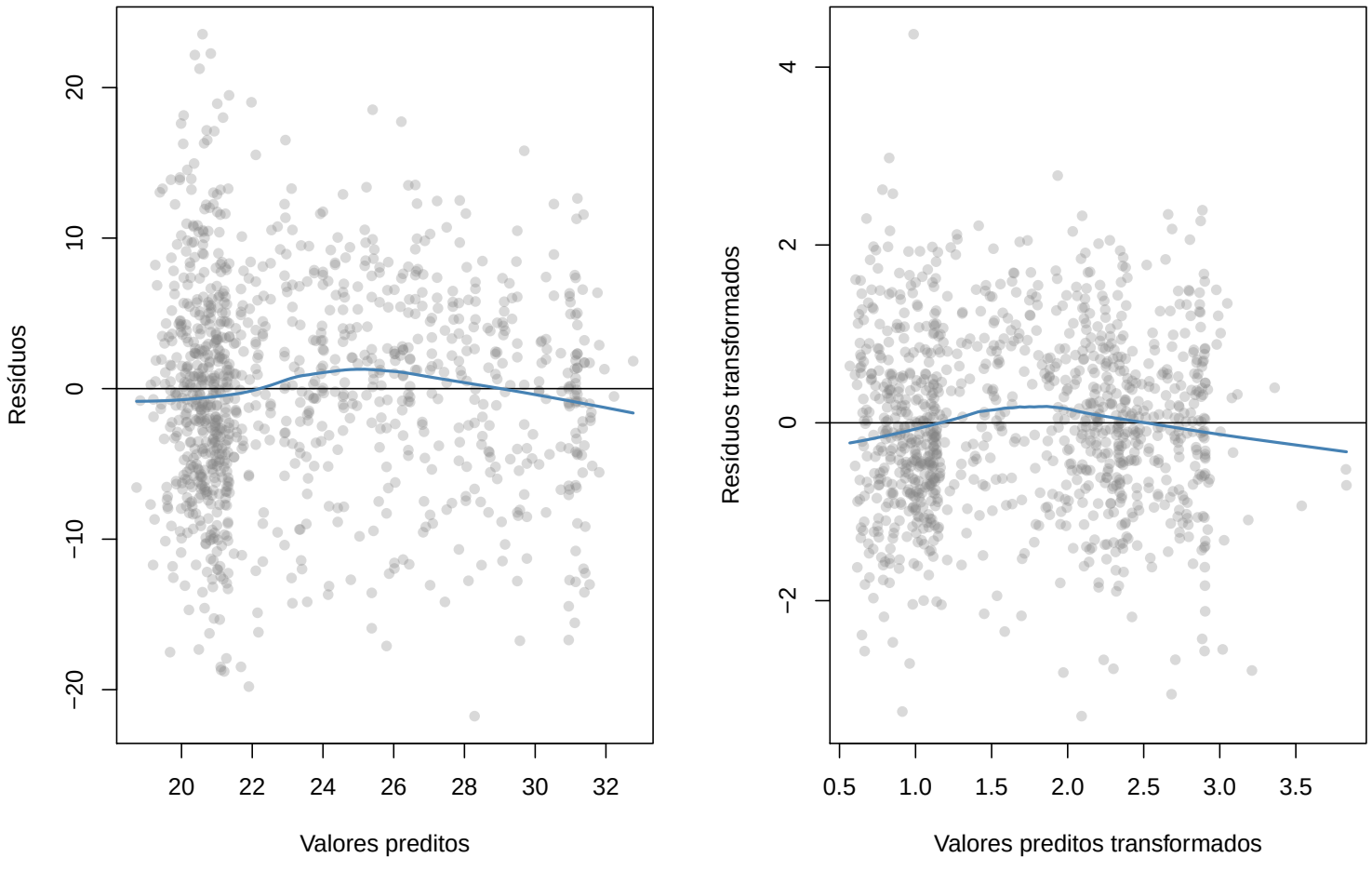
Exemplo


Quantis amostrais dos resíduos
20
10
0
−10
−20
−3
−2
−1
0
1
2
3
Quantis teóricos da normal padrão


Quantis amostrais dos resíduos transformados
4
2
0
−2
−3
−2
−1
0
1
2
3
Quantis teóricos da normal padrão

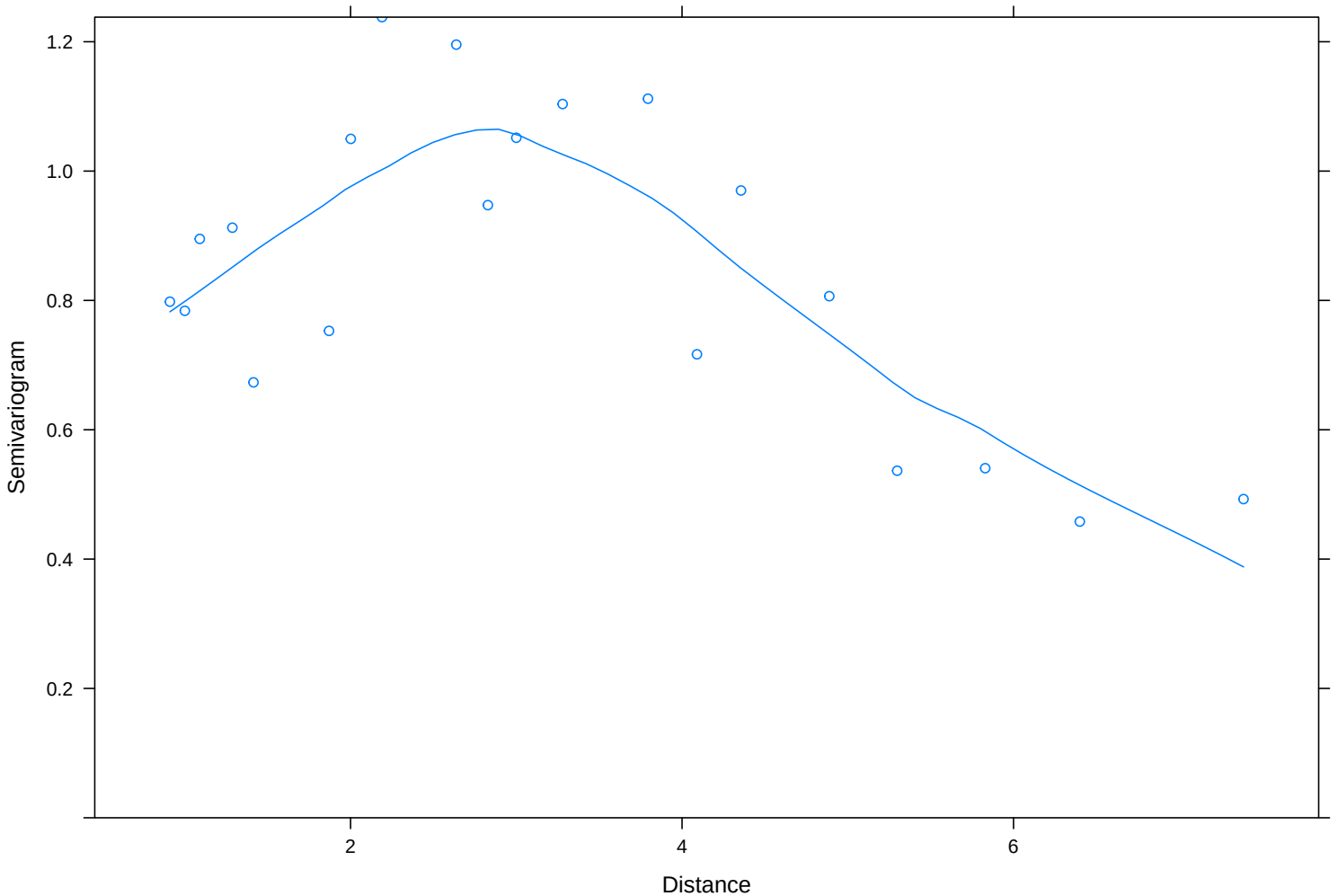
Exemplo

Gráfico de dispersão dos Resíduos versus Valores ajustados



Exemplo

Semi-variograma? (Este gráfico ainda precisa ser melhor estudado)



O que fazer quando as suposições do modelo são violadas?

- Verificar a estrutura da média.
- Transformar a resposta.
- Propor outra estrutura de variância-covariância para os erros (Modelo Marginal).
- Modelar a estrutura variância-covariância do erro intra-indivíduo (erro de medida, modelo de efeitos aleatórios).

Verificando a estrutura da média

- Existe alguma proposta teórica da área de pesquisa?
- Os perfis suavizados (*LOESS*) são as principais ferramentas de análise exploratória para estrutura da média em função do tempo.
- Propostas empíricas: funções *splines* (com um ou dois *knots*), modelos lineares ou com termos quadráticos.

Verificando a estrutura da média

- Outra possibilidade é a **transformação** da variável resposta.
- **Desvantagem:** a interpretação dos resultados se torna mais complexa.

Estrutura de covariância

- Em **modelos marginais** utilize o padrão de covariância a não-estruturada em delineamentos balanceados quando o número de tempos de medição não for excessivo.
  - Incluir heterocedasticidade quando possível.
- Em **modelos de efeitos aleatórios** utilizar uma outra estrutura para $R_i = \text{Cov}(\epsilon_i)$.

Exercício

Exercício

- Considere o exemplo anterior e ajuste um modelo linear de efeitos mistos com um termo *spline* quadrático.
- Realize a análise de resíduos para este modelo.

Avisos

Avisos

- **Próxima aula:** Modelos lineares generalizados para dados longitudinais.
  - Revisão de modelos lineares generalizados.
- **Para casa:** ler os Capítulos 9 e 10 do livro "**Applied Longitudinal Analysis**".
  - Caso ainda não tenha lido, leia também os Caps. 1, 2, 3, 4, 5, 6, 7 e 8.

Bons estudos!