Análise de Dados Longitudinais
Análise de Resíduos e Diagnóstico

Enrico A. Colosimo/UFMG
http://www.est.ufmg.br/~enricoc/

Introdução

Pontos Principais:

- A análise de dados longitudinais não fica completa sem a 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.

Suposições dos Modelos

- Estrutura da média: forma analítica, linearidade dos β's.
- Normalidade (resposta e efeitos aleatórios).
- Estrutura de Variância-Covariância: Homocedasticidade e correlação das medidas do mesmo indivíduo.

Resíduos

- Defina o vetor de resíduos para cada indivíduo

$$r_i = Y_i - X_i\hat{\beta}, \quad i = 1, \ldots, N,$$

que é um estimador para o vetor de erros

$$\epsilon_i = Y_i - X_i\beta, \quad i = 1, \ldots, N.$$

- Tratando-se de dados longitudinais, sabemos que os componentes do vetor de resíduos r_i são correlacionados e não necessariamente têm variância constante.

Utilidade dos Resíduos r_i

Gráficos:

- Gráfico de r_{ij} vs $\widehat{Y}_{ij}$: é útil para identificar alguma tendência sistemática (por exemplo, presença de curvatura) e presença de pontos extremos (”outliers”). O modelo corretamente especificado não deve apresentar nenhuma tendência neste gráfico.

  Limitação: este gráfico não tem necessariamente uma largura constante. Ou seja, cuidado ao interpretar este gráfico com relação a homocedasticidade.

- Gráfico de r_{ij} vs t_{ij}: é também útil para identificar alguma tendência sistemática da média no tempo.

Solução: Examinar resíduos transformados

- Há muitas possibilidades para transformar os resíduos.
- A tranformação 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:

$$r_i^* = L_i^{-1} r_i,$$

em que L_i é a matriz triangular superior resultante da decomposição de Cholesky da matriz de covariâncias estimada $\widehat{Var}(Y_i)$, ou seja, $\widehat{Var}(Y_i) = L_i L_i'$.

Resíduos transformados

- Podemos aplicar a mesma transformação ao vetor de valores preditos $\hat{Y}_i$, ao vetor da variável resposta Y_i e à matriz de covariáveis $\mathbf{X}_i$:

$$\begin{aligned} \hat{Y}_i^* &= L_i^{-1} \hat{Y}_i \\ Y_i^* &= L_i^{-1} Y_i \\ \mathbf{X}_i^* &= \hat{L}_i^{-1} \mathbf{X}_i \end{aligned}$$

e então todos os diagnósticos de resíduos usuais para a regressão linear padrão podem ser aplicados para r_i^*.

Gráficos de Adequação

- Gráfico de dispersão dos resíduos transformados r_{ij}^* versus os valores preditos transformados $\hat{Y}_{ij}^*$: não deve apresentar nenhum padrão sistemático para um modelo corretamente especificado. Ou seja, deve apresentar um padrão aleatório em torno de uma média zero. Útil para verificar homocedasticidade.

- Gráfico de dispersão dos resíduos transformados r_{ij}^* versus covariáveis transformadas X_{ij}^* (em especial, idade ou tempo): verificar padrões de mudança na resposta média ao longo do tempo;

- QQ-plot de r_i^*: verificar normalidade e identificar outliers.

Semi-variograma

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

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

em que $h_{ijk} = t_{ij} - t_{ik}$.

- O semi-variograma pode ser utilizado como uma ferramenta para verificar a adequação do modelo selecionado para a estrutura de covariância dos dados.

Semi-variograma

- Como os resíduos têm média zero, o semi-variograma pode ser reescrito como:

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

- Quando o semivariograma é aplicado aos resíduos transformados, r_{ij}^*, a seguinte simplificação é obtida:

$$\gamma(h_{ijk}) = \frac{1}{2}(1) + \frac{1}{2}(1) - 0 = 1.$$

Semi-variograma

- Logo, se o modelo é corretamente especificado para a matriz de covariâncias, o gráfico do semi-variograma amostral $\hat{\gamma}(h_{ijk})$ dos resíduos transformados versus h_{ijk} deveria flutuar aleatoriamente em torno de uma linha horizontal centrada em 1.

- O semi-variograma é muito sensível a outliers.

Estudo de caso: 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.

Estudo de Caso

- Há um total de 1049 medidas, com uma média de 6,4 medidas por menina.

- Variáveis de interesse:
  - 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.

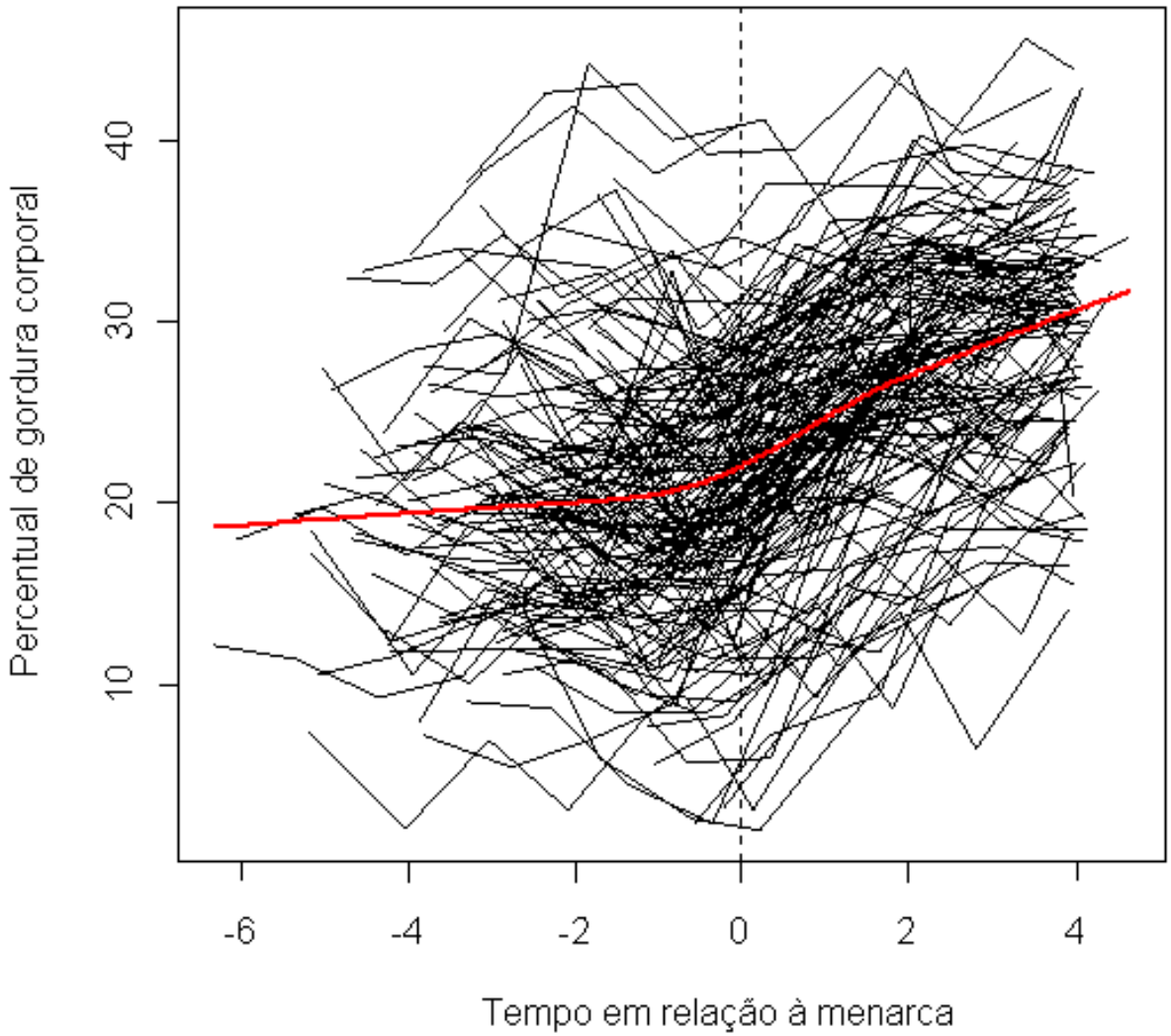

Figura: Gráfico de perfis com curva alisada

- 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:

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

em que

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

- Lembremos que no modelo linear de efeitos mistos, a matriz de variância-covariância de Y_i é dada por:

$$Var(Y_i) = Z_i \Sigma Z_i' + \sigma^2 I_{n_i},$$

em que Z_i é a matriz de covariáveis relacionadas aos efeitos aleatórios, Σ é a matriz de covariância dos efeitos aleatórios e n_i é o número de observações do $i-$ésimo indivíduo, $i = 1, \ldots, N$.

- Logo, os resíduos transformados neste caso podem ser obtidos a partir da decomposição de Cholesky da matriz estimada $\widehat{Var}(Y_i) = Z_i \hat{\Sigma} Z_i' + \hat{\sigma}^2 I_{n_i}$.

- Resultados do ajuste:

Tabela : Coeficientes de regressão estimados (efeitos fixos) e erros padrões

Variável	Estimativa	EP	t	p-valor
Intercepto	21,3614	0,5646	37,8400	0,0000
Tempo	0,4171	0,1572	2,6500	0,0081
$(\text{Tempo})_+$	2,0471	0,2280	8,9800	0,0000

Tabela : Covariâncias estimadas para os efeitos aleatórios ($\hat{G}$) e variância estimada para os erros ($\hat{\sigma}^2$)

Parâmetro	Estimativa	Parâmetro	Estimativa
$Var(b_{1i}) = g_{11}$	45,9407	$Cov(b_{1i}, b_{2i}) = g_{12}$	2,5275
$Var(b_{2i}) = g_{22}$	1,6309	$Cov(b_{1i}, b_{3i}) = g_{13}$	-6,1141
$Var(b_{3i}) = g_{33}$	2,7496	$Cov(b_{2i}, b_{3i}) = g_{23}$	-1,7513
$Var(e_i) = \sigma^2$	9,4734		

- Análise de resíduos:

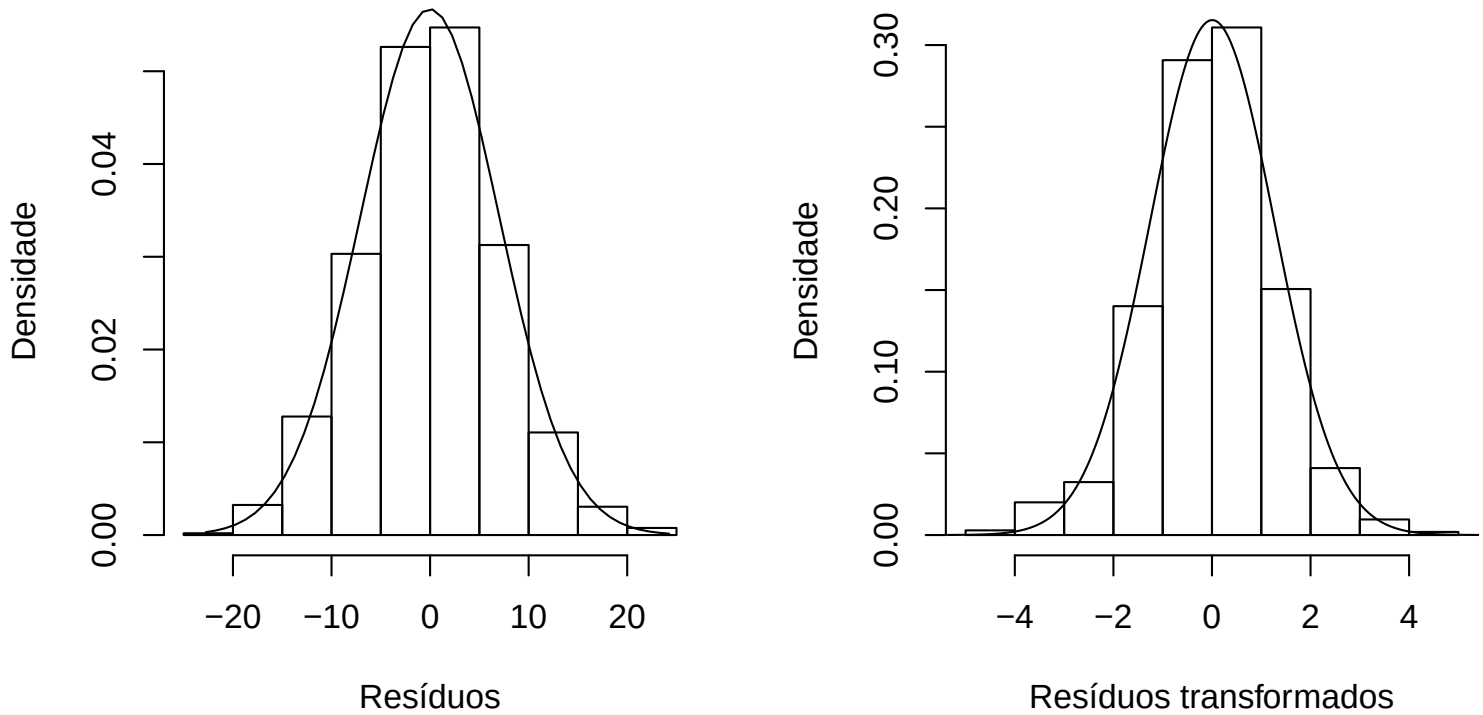

Figura: Histograma dos resíduos e resíduos transformados, com curva Normal

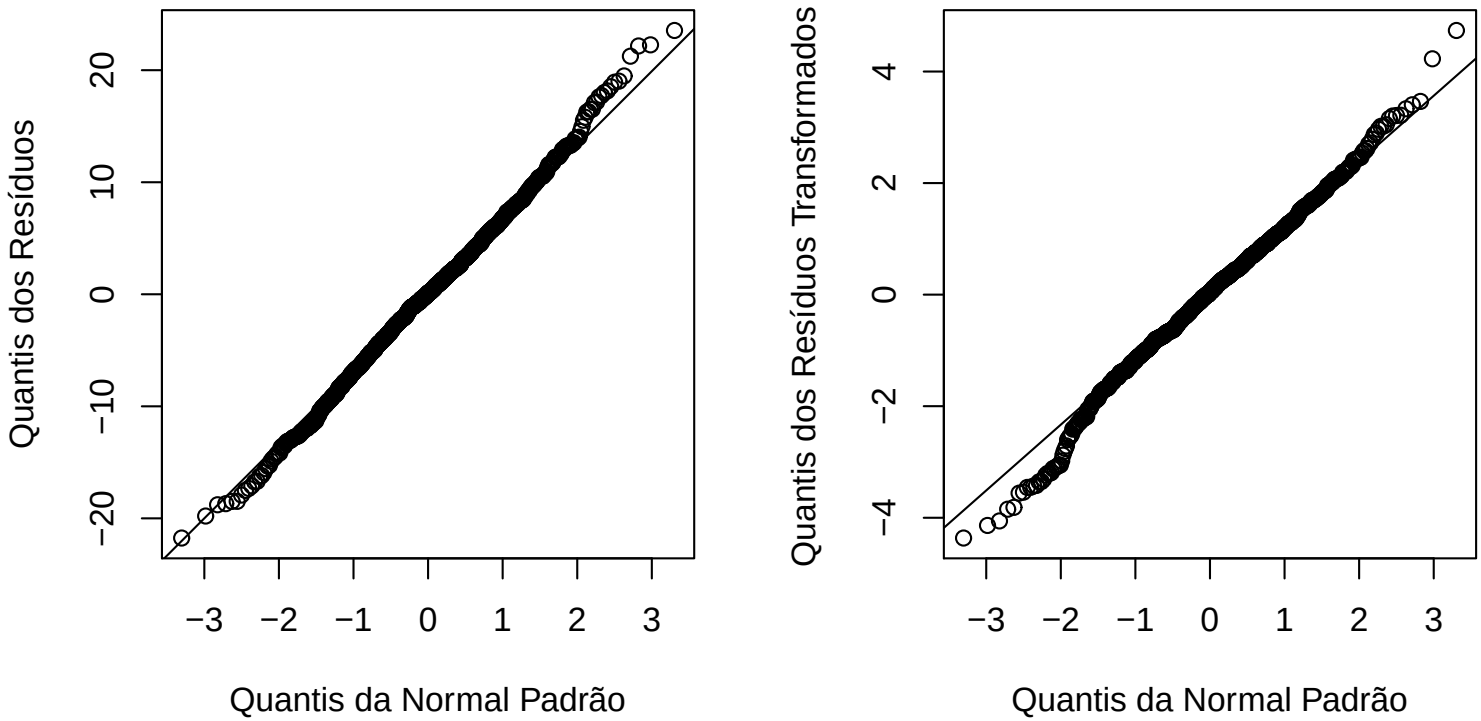

Figura: QQ-plot dos resíduos e resíduos transformados

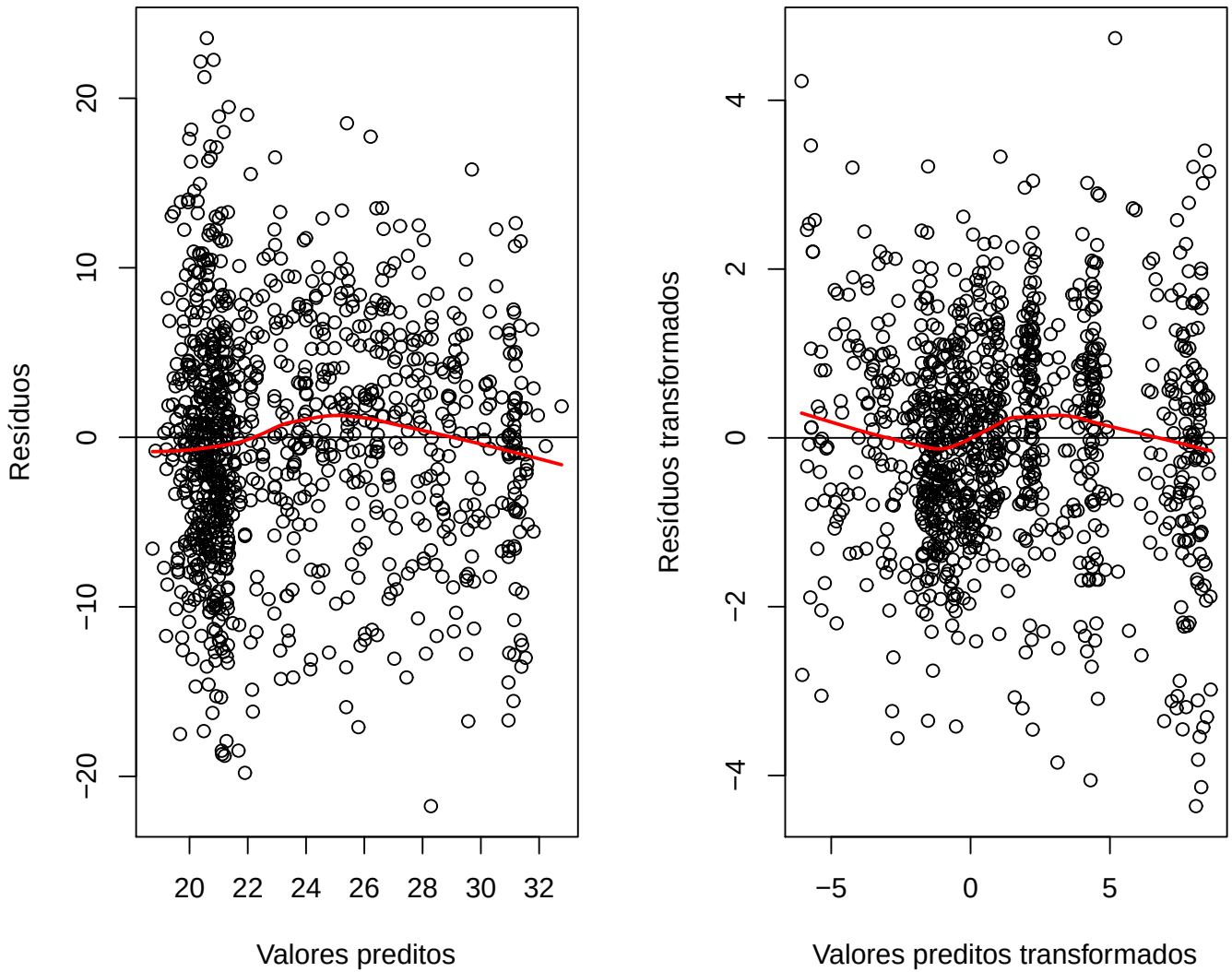

Figura: Resíduos vs Preditos e Resíduos transformados vs Preditos transformados

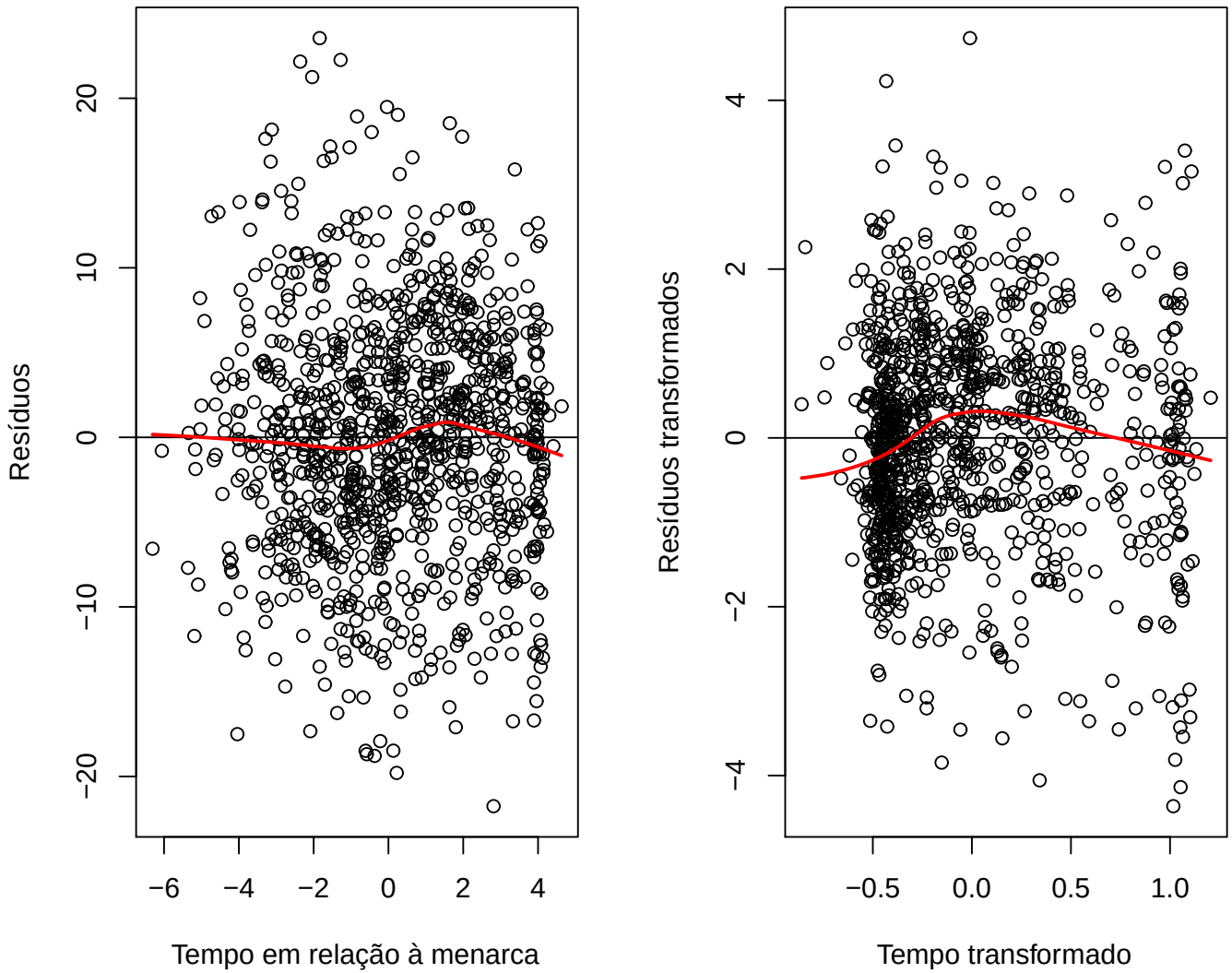

Figura: Resíduos vs Tempo e Resíduos transformados vs Tempo transformado

- Da figura anterior (Resíduos vs Tempo), observa-se uma tendência quadrática no período após a menarca.

- Refinando o modelo anterior, consideraremos agora que cada garota tem uma curva de crescimento spline linear-quadrática com um knot no tempo da menarca.

- Ajustou-se o seguinte modelo linear de efeitos mistos:

$$\begin{aligned} E(Y_{ij}|b_i) &= \beta_1 + \beta_2 t_{ij} + \beta_3 (t_{ij})_+ + \beta_4 (t_{ij})_+^2 + b_{1i} + b_{2i} t_{ij} + \\ &+ b_{3i} (t_{ij})_+ + b_{4i} (t_{ij})_+^2, \end{aligned}$$

em que

$$(t_{ij})_+^2 = \begin{cases} t_{ij}^2 & \text{se} \quad t_{ij} > 0 \\ 0 & \text{se} \quad t_{ij} \leq 0. \end{cases}$$

- Resultados do ajuste:

Tabela : Coeficientes de regressão estimados (efeitos fixos) e erros padrões

Variável	Estimativa	EP	t	p-valor
Intercepto	20,4201	0,5817	35,1032	0,0000
Tempo	-0,0155	0,1612	-0,0962	0,9234
$(\text{Tempo})_+$	4,8439	0,4055	11,9446	0,0000
$(\text{Tempo})_+^2$	-0,6469	0,0772	-8,3842	0,0000

Tabela : Covariâncias estimadas para os efeitos aleatórios ($\hat{G}$) e variância estimada para os erros ($\hat{\sigma}^2$)

Parâmetro	Estimativa	Parâmetro	Estimativa
$Var(b_{1i}) = g_{11}$	48,0586	$Cov(b_{1i}, b_{3i}) = g_{13}$	-9,5900
$Var(b_{2i}) = g_{22}$	1,7326	$Cov(b_{1i}, b_{4i}) = g_{14}$	0,6479
$Var(b_{3i}) = g_{33}$	5,3693	$Cov(b_{2i}, b_{3i}) = g_{23}$	-1,5342
$Var(b_{4i}) = g_{44}$	0,1172	$Cov(b_{2i}, b_{4i}) = g_{24}$	-0,1735
$Cov(b_{1i}, b_{2i}) = g_{12}$	3,0295	$Cov(b_{3i}, b_{4i}) = g_{34}$	-0,4395
$Var(e_i) = \sigma^2$	8,0274		

- Análise de resíduos:

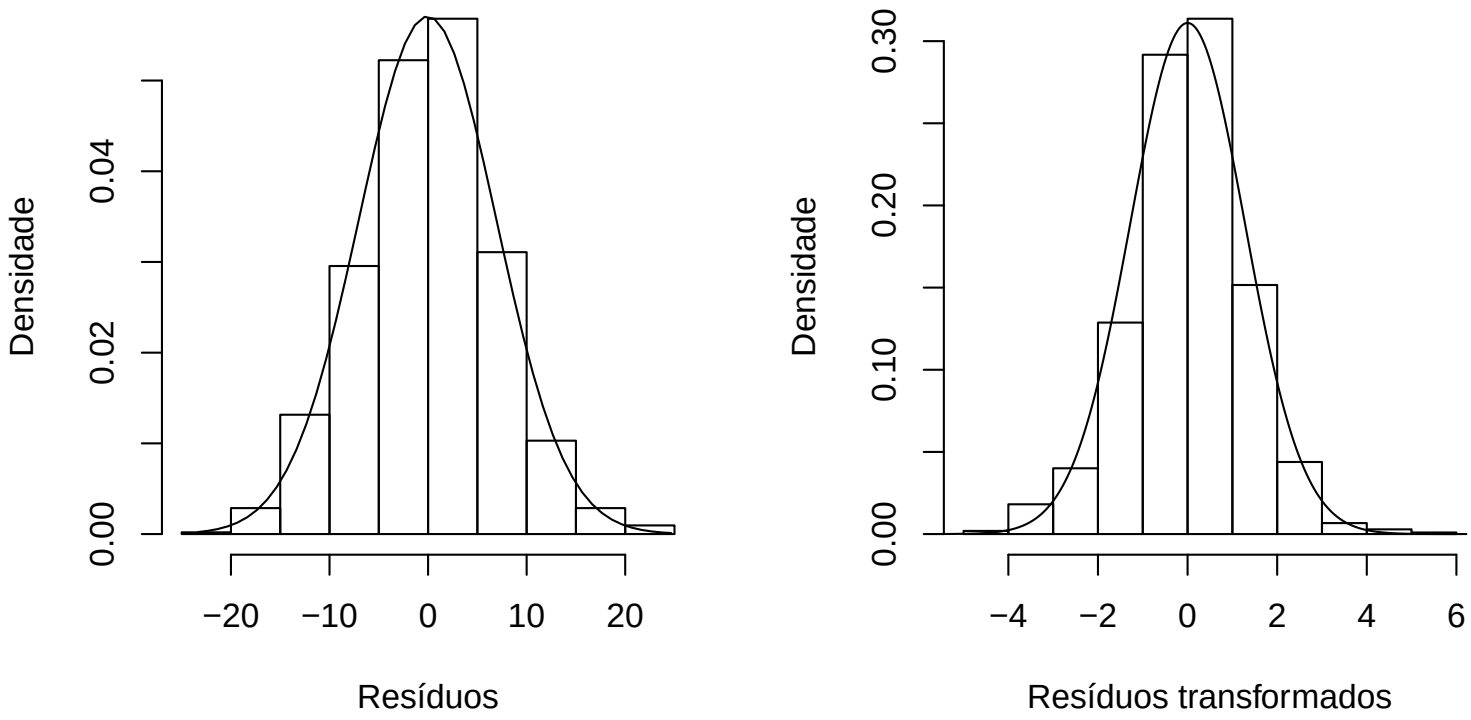

Figura: Histograma dos resíduos e resíduos transformados, com curva Normal

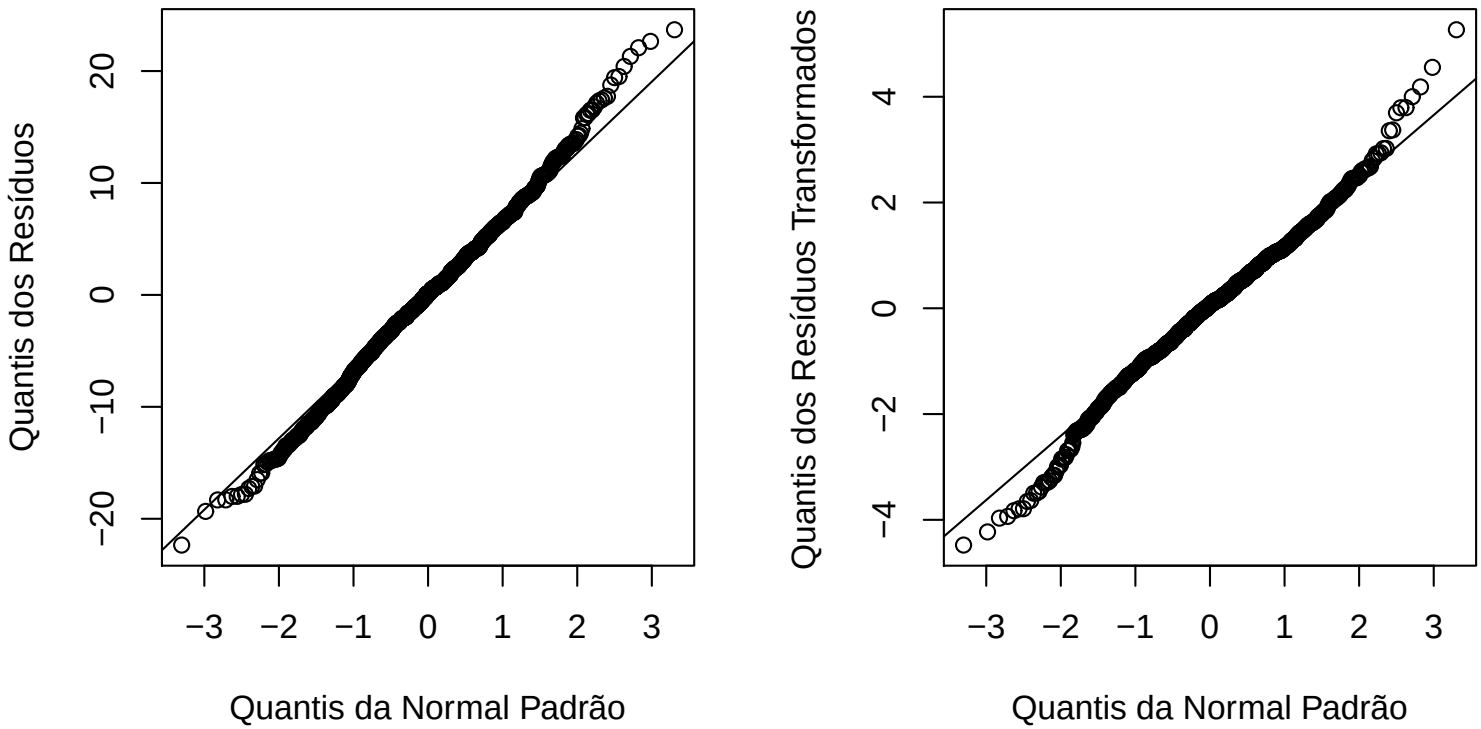

Figura: QQ-plot dos resíduos e resíduos transformados

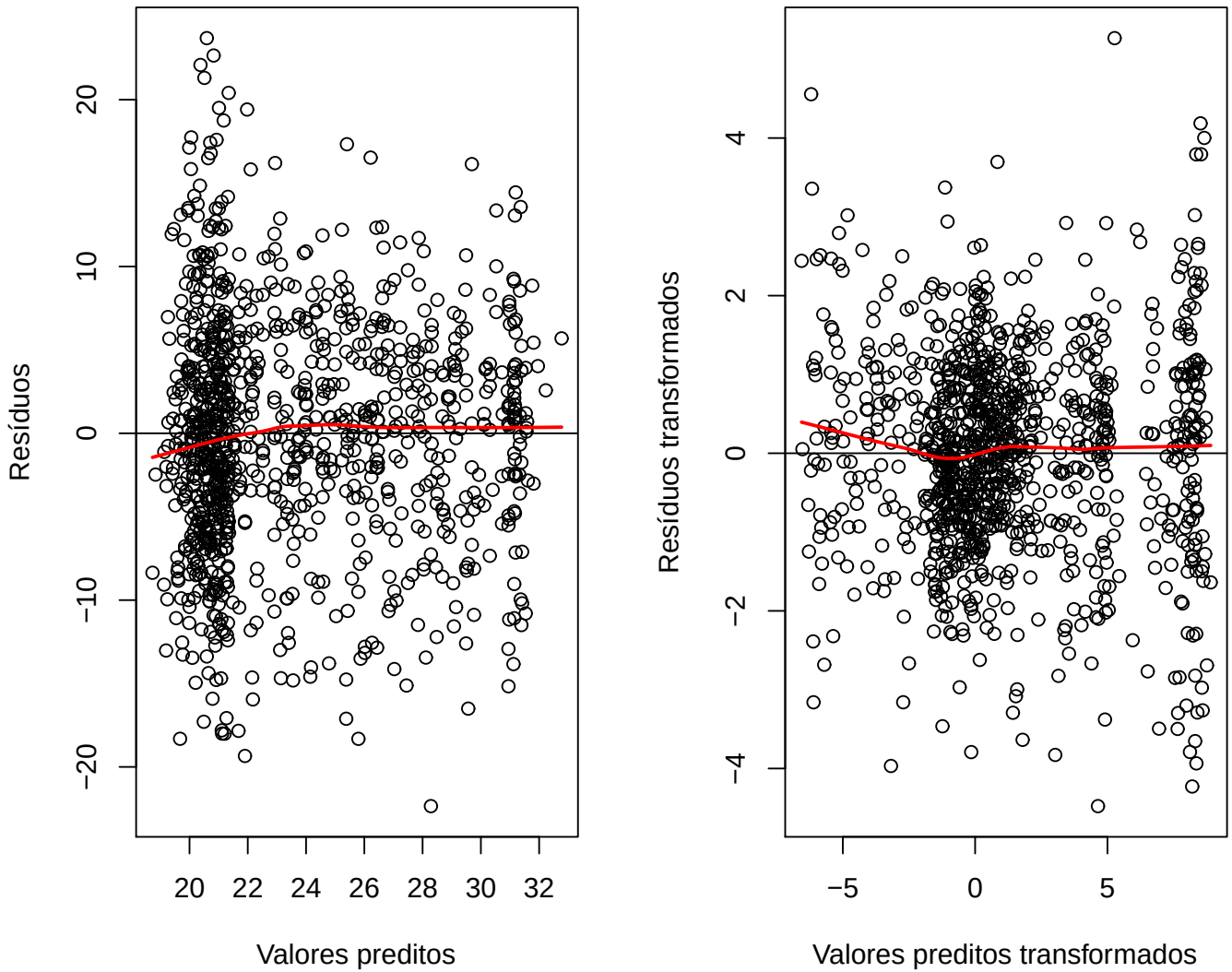

Figura: Resíduos vs Preditos e Resíduos transformados vs Preditos transformados

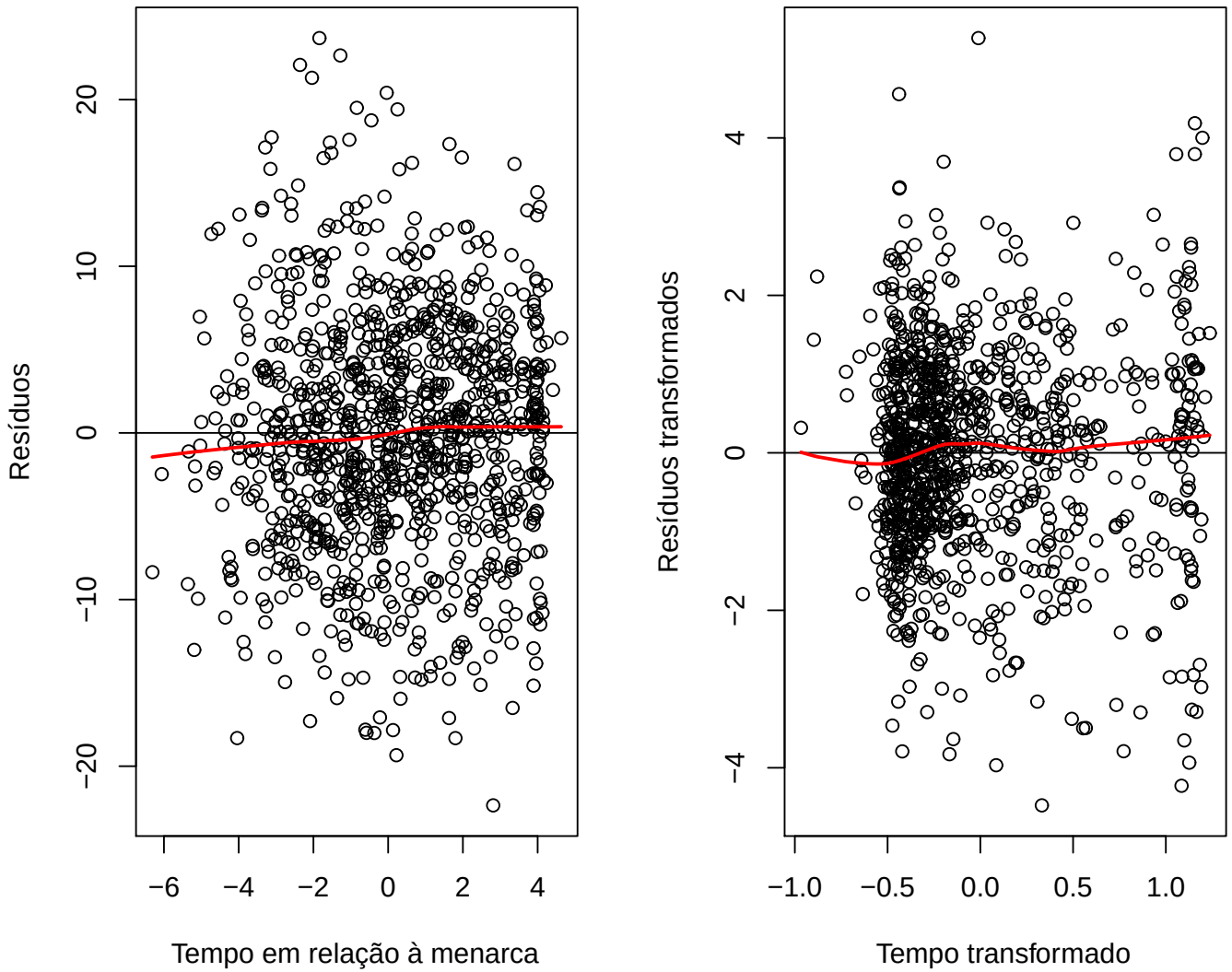

Figura: Resíduos vs Tempo e Resíduos transformados vs Tempo transformado

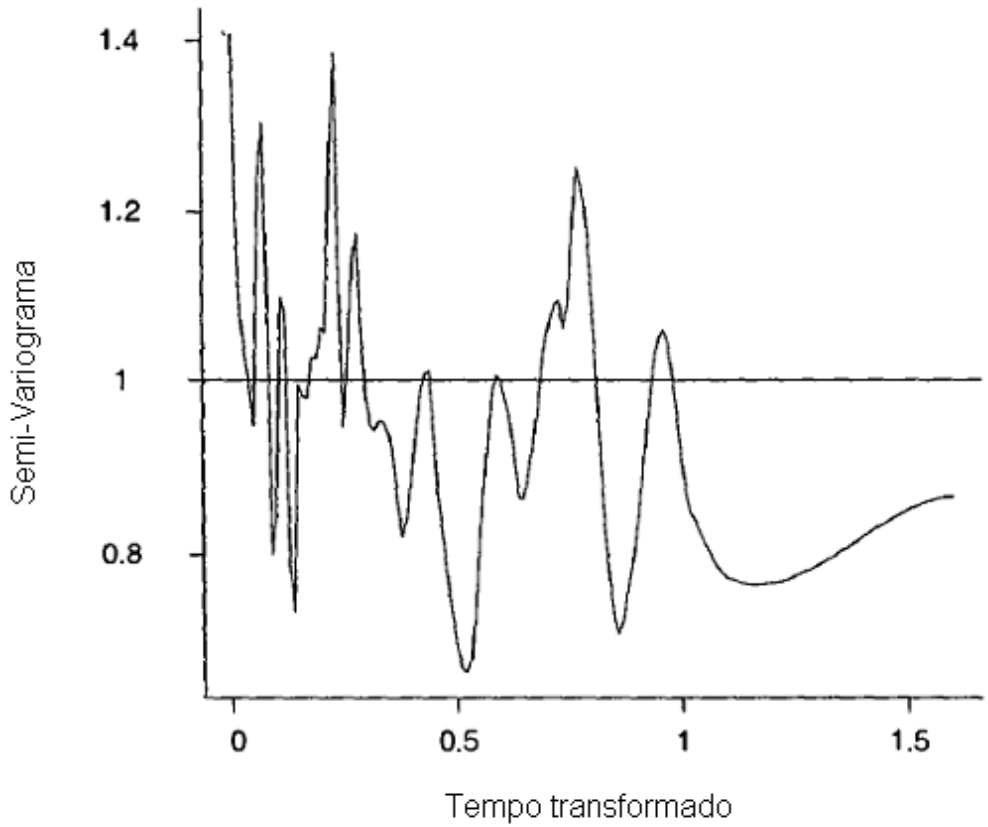

Figura: Semi-variograma empírico para os resíduos transformados

- Gráficos de dispersão não apresentam mais nenhuma tendência acentuada.
- Semi-variograma está oscilando aleatoriamente em torno da linha horizontal 1.
- Pela análise de resíduos, confirmamos a adequação do segundo modelo proposto.

O que fazer frente a violação de suposições?

- 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 Efeito Aleatórios).

Verificar a Estrutura da Média

- Existe alguma proposta teórica da área?
- Perfis, especialmente os alisados, são as principais ferramentas.
- Propostas Empíricas: splines (com um ou no máximo dois knots), modelos lineares ou quadráticos. Possivelmente algo como decaimento exponencial.

Transformar a resposta

- Vantagens quando temos distribuição assimétrica para a resposta. Por exemplo: custo. Utilizar transformação logarítmica.
- Desvantagem: interpretação dos resultados.

Propor outra estrutura de Variância-Covariância para os erros (Modelo Marginal)

- Utilizar a não-estruturada em delineamentos balanceados quando o número de tempos medidos não for excessivo.
- Incluir heterocedasticidade quando possível.

Modelar a variância-covariância do erro Intra Indivíduos (Modelo de Efeito Aleatórios)

- Suposição: $Var(\varepsilon_i) = \sigma^2 I$.
- Podemos estruturar a

$$Var(\varepsilon_i).$$

  Isso pode ser feito inclusive em termos de covariáveis.
- O R ajusta alguns tipos de estrutura.