Análise de Dados Longitudinais
Modelos Lineares Mistos

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

Modelo Linear Misto

1. Modelo de Efeitos Fixos: apresenta somente fatores fixos, exceto o termo do erro experimental (erro de medida).

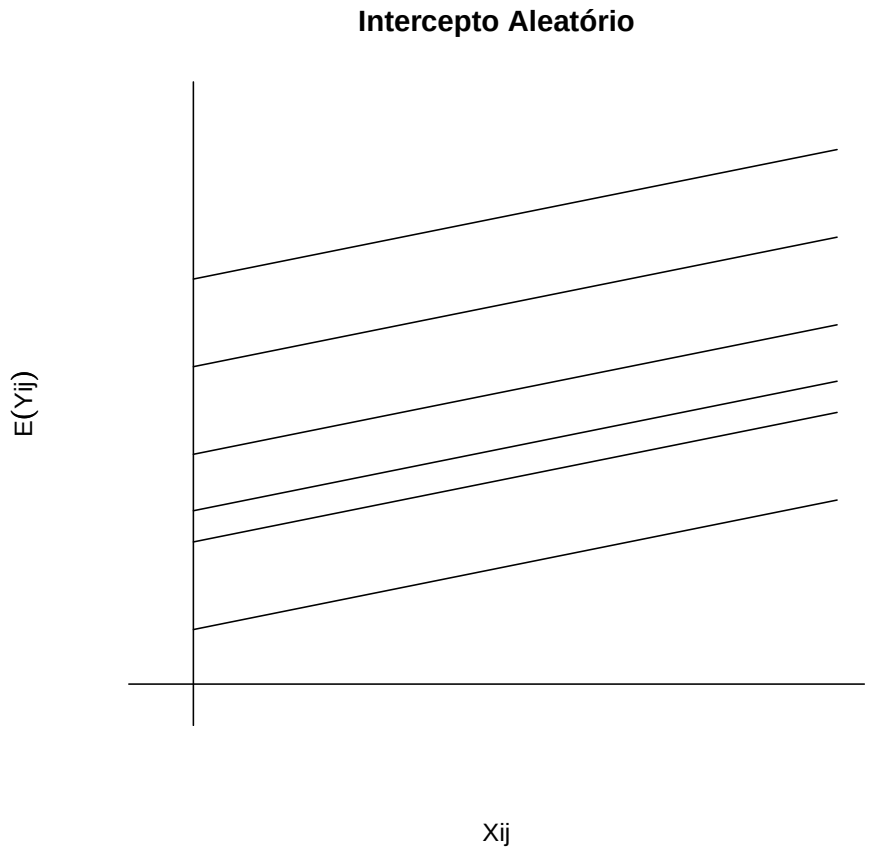
2. Modelo Misto: apresenta tanto fatores fixos quanto aleatórios, além do erro experimental.

Ideia:

- Os parâmetros da regressão variam de indivíduo para indivíduo explicando as fontes de heterogeneidade da população.
- Cada indivíduo tem a sua própria trajetória média e um subconjunto dos parâmetros de regressão são tomados como aleatórios.
- Efeitos fixos são compatilhados por todos os indivíduos e os aleatórios são especificos de cada um.

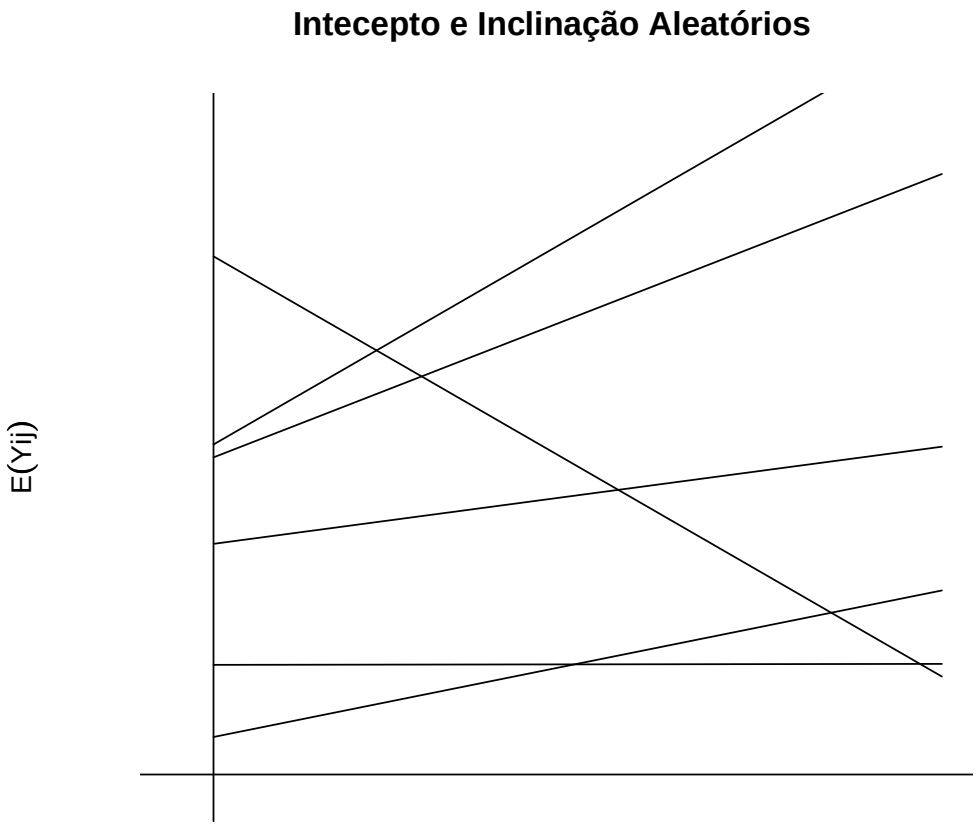
Exemplo: Intercepto aleatório

$$Y_{ij} = \beta_{0i} + \beta_1 t_{ij} + \varepsilon_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + \varepsilon_{ij}$$



Exemplo: Intercepto e Inclinação aleatórios

$$Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij} = \beta_0 + b_{0i} + \beta_1 t_{ij} + b_{1i} t_{ij} + \varepsilon_{ij}$$



Modelo Linear Misto

- **Características:**
  1. Características populacionais β (fixos);
  2. Características individuais β_i ou b_i (aleatórios).

- **Efeito:**
  1. Média: $E(Y_i) = X_i\beta$
  2. Estrutura de Covariância: Efeito aleatório induz $Var(Y_i)$.

     Separa a variação entre individuos daquela intra individuos.
  3. Permite obter estimativa de trajetórias individuais no tempo.

Modelo Linear Misto - Simetria Composta

Exemplo: $Y_{ij} = \beta_{0i} + \beta t_{ij} + \varepsilon_{ij}$ (Intercepto aleatório).

- $\beta_{0i} \sim N(\beta_0, \sigma_0^2)$.
- $\varepsilon_{ij} \sim N(0, \sigma^2)$.
- β_{0i} e ε_{ij} são independentes.

1. $Var(Y_{ij}) = \sigma^2 + \sigma_0^2$.
2. $Cov(Y_{ij}, Y_{ij'}) = \sigma_0^2$

Modelo Linear Misto - Inclinação aleatória

Exemplo: $Y_{ij} = \beta_{0i} + \beta_{1i} t_{ij} + \varepsilon_{ij}$ (Intercepto e inclinação aleatórios).

- $\beta_{0i} \sim N(\beta_0, \sigma_0^2)$, $\beta_{1i} \sim N(\beta_1, \sigma_1^2)$, $Cov(\beta_{0i}, \beta_{1i}) = \sigma_{01}$.
- $\varepsilon_{ij} \sim N(0, \sigma^2)$.
- $\beta' = (\beta_{0i}, \beta_{1i})$ e ε_{ij} são independentes.

1. $Var(Y_{ij}) = \sigma_0^2 + \sigma_1^2 t_{ij}^2 + 2 t_{ij} \sigma_{01} + \sigma^2$.
2. $Cov(Y_{ij}, Y_{ij'}) = \sigma_0^2 + t_{ij} t_{ij'} \sigma_1^2 + (t_{ij} + t_{ij'}) \sigma_{01}$.

Vantagens

1. Predizer trajetórias individuais

$$Y_{ij} = X_{ij}\beta + b_i + \varepsilon_{ij}$$

Resposta Média populacional:

$$E(Y_{ij}) = X_{ij}\beta$$

Resposta média para o i-ésimo indivíduo (trajetória):

$$E(Y_{ij}/b_i) = X_{ij}\beta + b_i.$$

2. Flexibilidade em acomodar estruturas não balanceadas

Forma Geral do Modelo Misto

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i$$

em que:

$(\beta)_{p\times 1}$: efeitos fixos;
$(b_i)_{q\times 1}$: efeitos aletaórios.

e,

$$b_i \sim N_q(0, \Sigma) \text{ e } \varepsilon_{ij} \sim N(0, \sigma^2)$$

Sendo b_i e ε_{ij} independentes.

$$q \leq p \Rightarrow Z_i \text{ é um subconjunto de } X_i$$

Incluimos efeitos aleatórios somente para as covariáveis que variam com o tempo.

Característica do Modelo

1. Média Populacional ou Marginal

$$E(Y_i) = X_i\beta.$$

2. Média condicional ou específica por indivíduo

$$E(Y_i|b_i) = X_i\beta + Z_i b_i.$$

3. Covariância Marginal

$$Var(Y_i) = Z_i Var(b_i) Z_i' + \sigma^2 I_n.$$

4. Podemos assumir que $\varepsilon_i \sim N(0, R_i)$ mas o usual é tomar $R_i = \sigma^2 I_{n_i}$ e interpreta-lo como covariância condicional. Ou seja,

$$Var(Y_i/b_i) = R_i = \sigma^2 I_n.$$

Inferência para o Modelo Misto

$$Y_i = X_i\beta + Z_i b_i + \varepsilon_i,$$

em que,

$$b_i \sim N_q(0, \Sigma(\alpha)) \text{ e } \varepsilon_{ij} \sim N(0, \sigma^2),$$

b_i e ε_{ij} independentes.

Desta forma tem-se:

p efeitos fixos e $\frac{q(q+1)}{2} + 1$ efeitos aletaórios.

Inferência Estatística para $\theta = (\beta, \alpha, \sigma^2)$;

1. Máxima Verossimilhança.
2. Máxima Verossimilhança Restrita.

Função de Verossimilhança

$$
\begin{aligned}
L(\theta/y) &= \prod_{i=1}^{N} p(y_i/\theta) \\
&= \prod_{i=1}^{N} \int p(y_i, b_i/\theta) db_i \\
&= \prod_{i=1}^{N} \int p(y_i/b_i, \theta) p(b_i/\theta) db_i
\end{aligned}
$$

em que,

$$p(y_i/b_i, \theta) \sim N_n(X_i\beta + Z_i b_i, \sigma^2 I)$$

e

$$p(b_i/\theta) \sim N_q(0, \Sigma)$$

Observações

1. EMV É obtido usando verossimilhança perfilada e iterações via algoritmo EM ou/e Newton-Raphson. Detalhes numéricos podem ser encontrados em Pinheiro e Bates (2000), Cap. 2.

2. O EMVR também pode ser obtido através de

$$l^*(\theta) = l(\theta) + \textit{termo}.$$

3. A função lme do R fornece EMVR e EMV usando um enfoque híbrido (EM + Newton-Raphson). Esta função é de autoria de Pinheiro e Bates.

4. EMV e EMVR têm assintoticamente as propriedades usuais de um estimador de máxima verossimilhança (consistência e normalidade).

Avaliação dos Componentes de Variância

1. Número de componentes é igual a $\frac{q(q+1)}{2} + 1$ em que q é o número de efeitos aleatórios no modelo.

2. Muitas situações envolvem $q = 2$ (intercepto e inclinação aleatórios) e portanto:

$$\frac{2(2+1)}{2} + 1 = 4,$$

que permite termos heterogeneidade de variâncias e covariâncias pois ficam em função do tempo.

3. A escolha da "melhor"estrutura de variância-covariância pode ser realizada utilizando o teste da RMVR. Estes testes, usualmente, são na fronteira do espaço de parâmetros. Neste caso, a estatística da RMVR não tem, sob H_0 uma distribuição qui-quadrado.

Dist. da Estatística da RMVR sob H_0

1. Para comparar dois modelos encaixados, respectivamente com $q+1$ e q efeitos aleatórios correlacionados, a distribuição, sob H_0 é uma mistura (50:50) de dist. qui-quadrados. Ou seja,

$$RMVR \sim 0,5\chi_q + 0,5\chi_{q+1}$$

2. Exemplo ($H_0 : \sigma_1^2 = \sigma_{01} = 0$)
Modelo completo: q=2 (intercepto e inclinação aleatórios)
Modelo restrito: q=1 (somente intercepto aleatório)

   Teste usual (errado): nível de significância (5%) : 5,99
   Teste correto:

$$RMVR \sim 0,5\chi_1 + 0,5\chi_2$$

   nível é 5,14 (Tabela, Apend. C, Fitzmaurice et al, 2004).

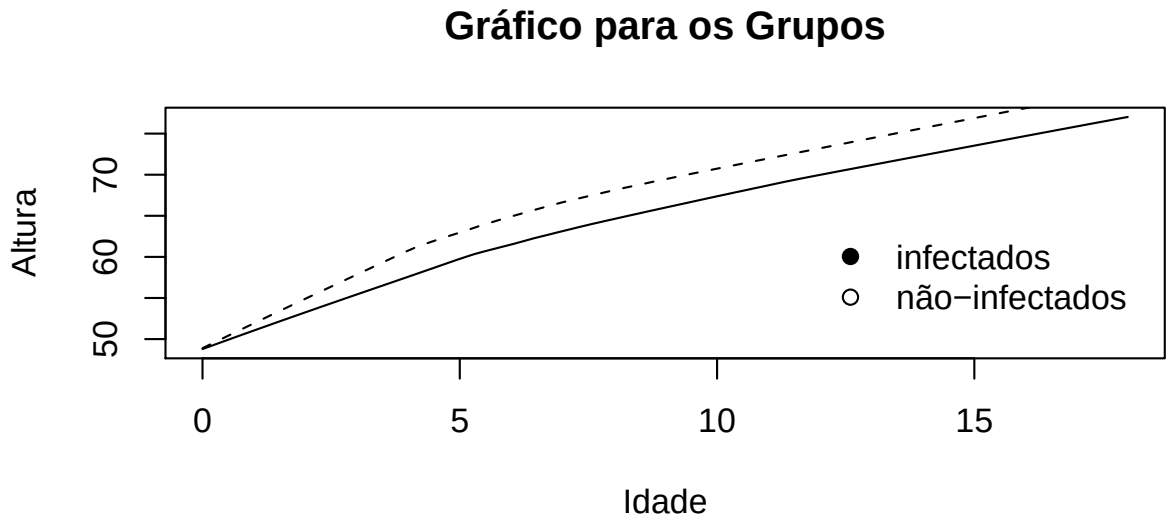
3. Proposta ad hoc: para testar a $0,05$, use o nível de $0,10$, neste caso o nível é 4,61.

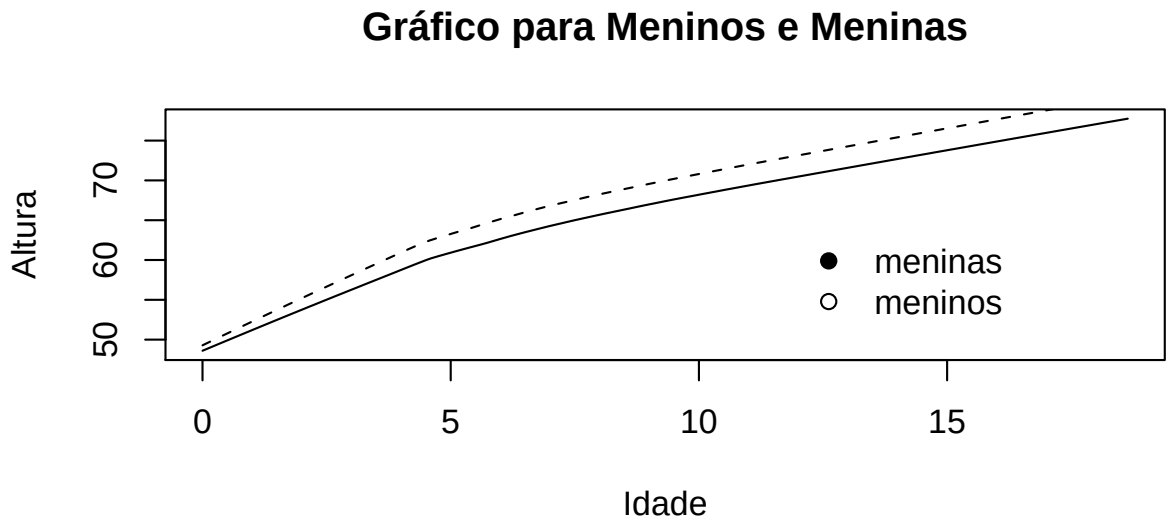
Transmissão Vertical - HIV

Estudo Longitudinal Desbalanceado: Avaliação longitudinal do crescimento de lactentes nascidos de mães infectadas com o HIV-1.

- Comparar longitudinalmente altura de lactentes infectados e não-infectados nascidos de mães infectadas pelo HIV.
- Uma coorte aberta acompanhada no ambulatório de AIDS pediátrica do Hospital das Clínicas da Universidade Federal de Minas Gerais.
- Período: 1995 a 2003.
- Inclusão: primeiros três meses de vida.
- Grupos: (1) não-infectados: 97; (2) infectados: 42.
- Controlado por sexo.

Perfis médio por grupo





Propostas de Modelos para a Média

1. Modelo Polinomial:

$$\begin{aligned} E(Y_{ij}) &= \beta_0 + \beta_{1i} t_{ij} + \beta_2 t_{ij}^2 + \beta_3 sexo_i + \beta_4 grupo_i + \beta_5 (t_{ij} * sexo_i) \\ &+ \beta_6 (t_{ij} * grupo_i) + \beta_7 (t_{ij}^2 * sexo_i) + \beta_8 (t_{ij}^2 * grupo_i) \end{aligned}$$

2. Modelo Segmentado (com knot em t=5 meses):

$$\begin{aligned} E(Y_{ij}) &= \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 5)_+ \beta_3 sexo_i + \beta_4 grupo_i + \beta_5 t_{ij} * sexo_i \\ &+ \beta_6 t_{ij} * grupo_i + \beta_7 (t_{ij} - 5)_+ * sexo_i + \beta_8 (t_{ij} - 5)_+ * grupo_i \end{aligned}$$

No modelo marginal, houve uma preferência pelo segmentado.

Crianças - Transmissão Vertical - Modelos Misto

- Modelo com duas inclinações para a média com termos de interação.
- Testados todos os termos aleatórios com os respectivos TRVR.
- Modelo Final: intercepto + duas inclinações aleatórias.
- Modelo Segmentado (com knot em t=5 meses):

$$\begin{aligned} Y_{ij} &= \beta_{0i} + \beta_{1i} t_{ij} + \beta_{2i}(t_{ij} - 5)_+ + \beta_3 sexo_i + \beta_4 grupo_i + \beta_5 t_{ij} * sexo_i \\ &+ \beta_6 t_{ij} * grupo_i + \beta_7 (t_{ij} - 5)_+ * sexo_i + \beta_8 (t_{ij} - 5)_+ * grupo_i + \epsilon_{ij} \end{aligned}$$

Crianças - Transmissão Vertical - Modelos Misto

```
> print(s$tTable, digits=3)
              Value Std.Error   DF t-value   p-value
(Intercept)  48.519    0.3191 1173  152.03  0.00e+00
Idade         2.998    0.0865 1173   34.68 6.15e-182
Age          -1.800    0.1107 1173  -16.26  8.87e-54
sexo          0.542    0.3985  136    1.36  1.76e-01
status       -0.888    0.4444  136   -2.00  4.76e-02
Idade:status -0.351    0.1198 1173   -2.93  3.42e-03
Age:status    0.297    0.1523 1173    1.95  5.18e-02
Idade:sexo    0.261    0.1076 1173    2.42  1.55e-02
Age:sexo     -0.190    0.1377 1173   -1.38  1.68e-01
>
   AIC  BIC logLik
  5526 5609  -2747

Random effects:
 Formula: ~Idade + Age | ident
 Structure: General positive-definite, Log-Cholesky parametrization
            StdDev Corr
(Intercept) 2.037  (Intr) Idade
Idade       0.531  -0.084
Age         0.651  -0.059 -0.908
Residual    1.422
```

Crianças - Transmissão Vertical - Modelos Misto

- Modelo quadrático para a média com termos de interação.
- Termos aleatórios: intercepto + linear, intercepto + linear + quadrático.
- Comparação dos dois modelos: RMVR (combinação de qui-quadrado)

```
> anova(out,out1)
     Model df      AIC      BIC    logLik   Test  L.Ratio p-value
out      1 16 5708.332 5791.164 -2838.166
out1     2 13 5817.677 5884.978 -2895.838 1 vs 2 115.3451  <.0001
```

- Resultados para os quatro termos da interação.

Coeficiente	Interc.		Inter + linear		três termos	
	Est.	EP	Est.	EP	Est.	EP
Idade:grupo	−0,142	0,027	−0,161	0.049	−0,148	0,050
Idade2:grupo	0,018	0,005	0,022	0,004	0,026	0,007
Idade:sexo	0,100	0,025	0,131	0,045	0,114	0,046
Idade2:sexo	−0,015	0,005	−0,009	0,004	−0,013	0,007

Crianças - Transmissão Vertical - Modelo Marginal (gls)

- Modelo quadrático para a média com termos de interação.
- Algumas formas para a $Var(Y_i)$: exponencial, simetria composta.
- Resultados para os quatro termos da interação.

Coeficiente	Indep. (Incorreto)		Exponencial		Simetria Composta	
	Est.	EP	Est.	EP	Est.	EP
Idade:grupo	−0,164	0,041	−0,160	0.057	−0,142	0,027
Idade2:grupo	0,020	0,008	0,017	0,008	0,018	0,005
Idade:sexo	0,046	0,037	0,165	0,052	0,100	0,025
Idade2:sexo	−0,014	0,007	−0,020	0,008	−0,015	0,005

Crianças - Transmissão Vertical - GEE

- Modelo quadrático para a média com termos de interação.
- Algumas formas para a $Var(Y_i)$: exponencial, simetria composta.
- Modelo para média com 9 termos (interceptos diferentes)
- Resultados para os quatro termos da interação.

Coeficiente	Independente		Simetria Composta	
	Est.	EP	Est.	EP
Idade:grupo	−0,164	0,059	−0,142	0,057
Idade2:grupo	0,020	0,011	0,018	0,008
Idade:sexo	0,046	0,050	0,100	0,047
Idade2:sexo	−0,014	0,009	−0,015	0,007

Obs. As estimativas são as mesmas do gls-normal e o erro-padrão fica inflacionado (ambos simetria composta).

Crianças - Transmissão Vertical - GEE

Resumo dos Modelos

Modelo	Est. Média	Est. Var/Cov	No. Param.	AIC/BIC
GLS	Duas inclinações	Expon. Cont.	9+2	5615/5672
GEE	Duas Inclinações	AR1	9+2	–
Misto 1	Quadrático.	Int+Idade+Idade^2	9+7	5708/5791
MIsto 2	Duas Inclinações	Int + Idade + Age	9+7	**5526/5609**

Análise de Resíduos e Diagnóstico

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:

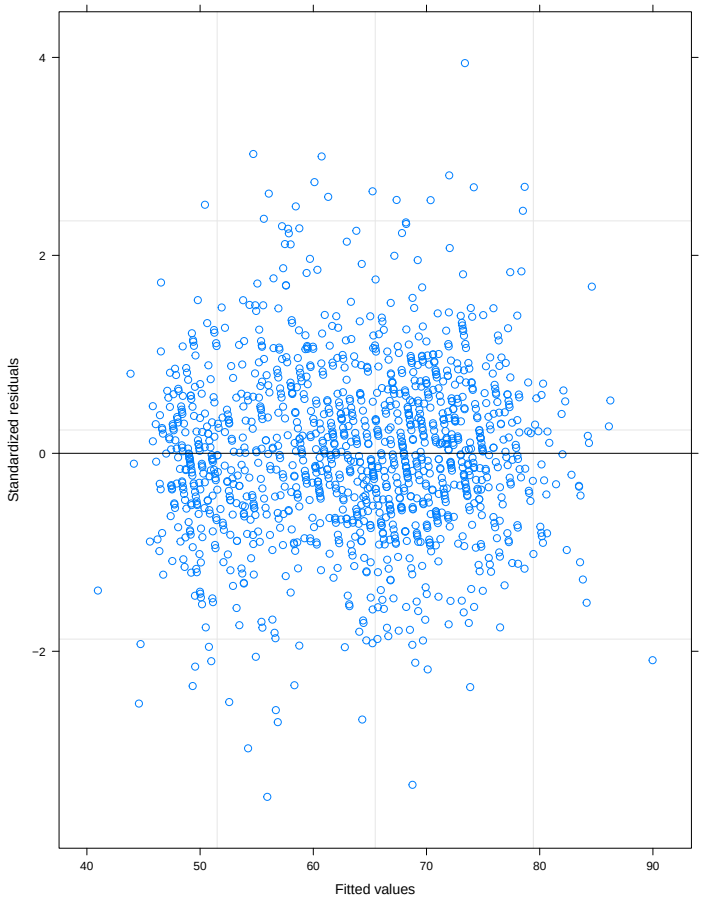
$$\gamma(h_{ijk}) = \frac{1}{2}(1) + \frac{1}{2}(1) - 0 = 1.$$

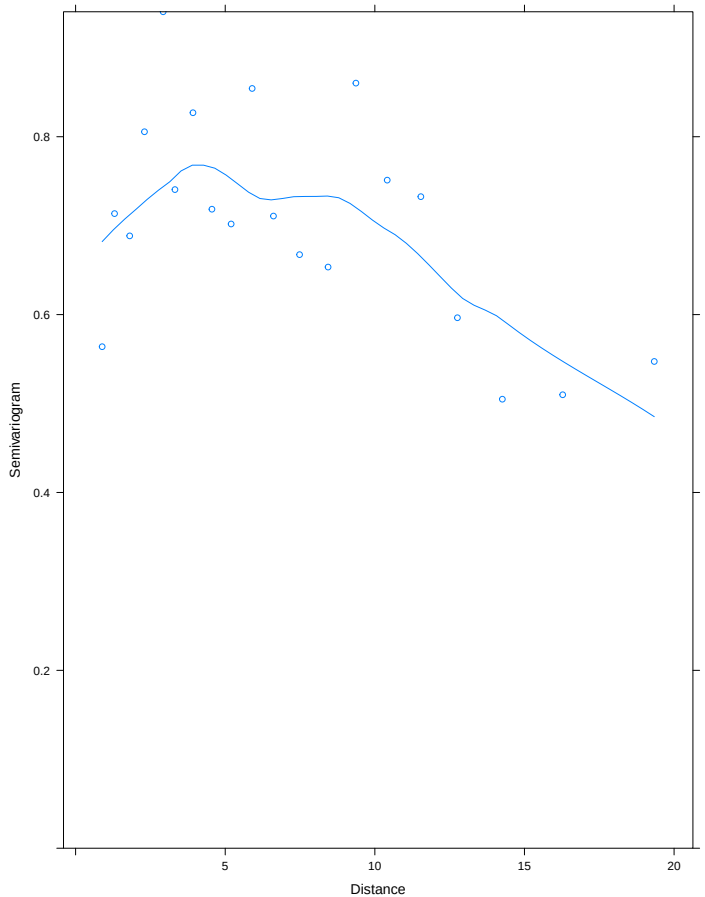
Semi-variograma

- 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.

Crianças - Transmissão Vertical: Análise de Resíduos





Crianças - Transmissão Vertical - Interpretação dos Resultados

Modelo misto com duas inclinações.

$$\begin{aligned} E(Y_{ij}) &= \beta_0 + \beta_1 t_{ij} + \beta_2 (t_{ij} - 5)_+ \beta_3 sexo_i + \beta_4 grupo_i + \beta_5 t_{ij} * sexo_i \\ &+ \beta_6 t_{ij} * grupo_i + \beta_7 (t_{ij} - 5)_+ * sexo_i + \beta_8 (t_{ij} - 5)_+ * grupo_i \end{aligned}$$

- (sexo=0 (menina) e status=0 (soronegativa))
  - 0 a 5 meses: $3 \pm 1,96 * 0,085$ - crescimento de 3 cm/mês (2,8; 3,2)
  - 5 a 18 meses: veloc. cresc.: $\hat{\beta}_1 + \hat{\beta}_2 = 3 - 1,8 = 1,2$ cm/mês.
    $Var(\hat{\beta}_1 + \hat{\beta}_6) = 0,00748 + 0,0123 - 2 * 0,00876 = 0,048^2$ IC
    $(1,2 \pm 1,96 * 0,048)$ = (1,1; 1,3).
- (sexo=0 (menina) e status=1 (soropositiva)
  - 0 a 5 meses: veloc. cresc.: $\hat{\beta}_1 + \hat{\beta}_6 = 3 - 0,35 = 2,6$ cm/mês.
    $Var(\hat{\beta}_1 + \hat{\beta}_6) = 0,00748 + 0,0144 - 2 * -0,00411 = 0,12^2$
    IC $(2,6 \pm 1,96 * 0,12)$ = (2,4; 2,8).
  - 5 a 18 meses: $\hat{\beta}_1 + \hat{\beta}_6 + \hat{\beta}_2 + \hat{\beta}_8 = 1,1$ cm/mês.
    $Var(\hat{\beta}_1 + \hat{\beta}_6 + \hat{\beta}_2 + \hat{\beta}_8) = 0.17^2$
    IC $(1,1 \pm 1,96 * 0,17)$ = (0,8; 1,4).

Crianças - Transmissão Vertical - Interpretação dos Resultados

Modelo misto com duas inclinações.

Resumo dos Resultados (cm/mês)

Termo	Período (meses)	
	0-5	5-18
Menina - Soronegativa	3,0 (2,8; 3,2)	1,2 (1,1; 1,3)
Menina - Soropositiva	2,6 (2,4; 2,8)	1,1 (0,8; 1,4)
Menino - Soronegativo	3,3 (3,2; 3,4)	1,3 (1,1; 1,5)
Menino - Soropositivo	2,9 (2,7; 3,1)	1,2 (1,1; 1,4)

Formulação em dois Estágios do Modelo Linear Misto

1. Estágio 1

   Medidas Longitudinais no i-ésimo indivíduo são modeladas como:

   $$Y_i = Z_i\beta_i + \varepsilon_i$$

   em que Z_i covariáveis intra-indivíduo (tempo dependente) e

   $\varepsilon_i \sim N(0, \sigma^2 I_n)$.

2. Estágio 2

   β_i: aleatório (variando de indivíduo para indivíduo) tal que:

   $$E(\beta_i) = A_i\beta$$

   em que A_i contém somente covariáveis que variam entre indivíduos (não dependente do tempo) e

Formulação em dois Estágios do Modelo Linear Misto

$$Var(\beta_i) = \Sigma.$$

Desta forma,

$$\begin{aligned} Y_i &= Z_i(A_i\beta + b_i) + \varepsilon_i \\ &= X_i\beta + Z_i b_i + \varepsilon_i \end{aligned}$$

Em que,

$$X_i = Z_i A_i$$

obtém-se o modelo de efeitos aleatórios.

Predição dos Efeitos Aleatórios

Objetivo: predizer perfis individuais ou identificar indivíduos acima ou abaixo do perfil médio.

Obs.: não dizemos estimar os efeitos pois os mesmos são aleatórios. Dizemos predizer os efeitos aleatórios.

Deseja-se:

$$\widehat{Y}_i = \widehat{E}(Y_i/b_i) = X_i\widehat{\beta} + Z_i\widehat{b}_i$$

e para tal necessita-se de $\widehat{b}_i$, o chamado Estimador BLUP ("Best Linear Unbiased Predictor") de b_i.

Predição dos Efeitos Aleatórios

No modelo linear misto,

- Y_i e b_i tem uma distribuição conjunta normal multivariada.
- Usando conhecidas propriedades da normal multivariada, temos que

$$E(b_i / Y_i, \widehat{\beta}) = \Sigma Z_i' Var(Y_i)^{-1}(Y_i - X_i\widehat{\beta})$$

- Usando os estimadores MVR dos componentes de variância,

$$\widehat{b_i} = \widehat{\Sigma} Z_i' \widehat{Var}(Y_i)^{-1}(Y_i - X_i\widehat{\beta})$$

o BLUP de b_i.

Predição dos Efeitos Aleatórios

Desta forma obtemos:

$$\begin{aligned}
\widehat{Y}_i &= X_i\widehat{\beta} + Z_i\widehat{b}_i \\
&= X_i\widehat{\beta} + Z_i\widehat{\Sigma}Z_i'\widehat{Var}(Y_i)^{-1}(Y_i - X_i\widehat{\beta}) \\
&= X_i\widehat{\beta} + (Z_i\widehat{\Sigma}Z_i' + \widehat{R}_i - \widehat{R}_i)\widehat{Var}(Y_i)^{-1}(Y_i - X_i\widehat{\beta}) \\
&= (\widehat{R}_i\widehat{Var}(Y_i)^{-1})X_i\widehat{\beta} + (I_{n_i} - \widehat{R}_i\widehat{Var}(Y_i)^{-1})Y_i
\end{aligned}$$

em que $Var(\varepsilon_i) = R_i$.

Interpretação: média ponderada entre a média populacional $X_i\widehat{\beta}$ e o i-ésimo perfil observado. Isto signfica que o perfil predito é encolhido na direção da média populacional.

Interpretação dos Efeitos Aleatórios Preditos

Ou seja,

- R_i: variação intra-indivíduo:
- $Var(Y_i)$: variação total.

- R_i grande, mais peso em $X_i\widehat{\beta}$;
- $Var(b_i)$ grande, mais peso em Y_i;
- n_i pequeno mais peso em $X_i\widehat{\beta}$.

Estudo de caso: Influência da menarca nas mudanças do percentual de gordura corporal - (FLW, 2011, p-220-228, 273-285)

- 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.
- **Objetivo**: avaliar o comportamento do percentual de gordura, antes e após a 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 (bioimpedância);
  - Covariáveis: Tempo em relação à menarca (Idade da menina no instante observado menos Idade quando teve a menarca) - pode ser positivo ou negativo.

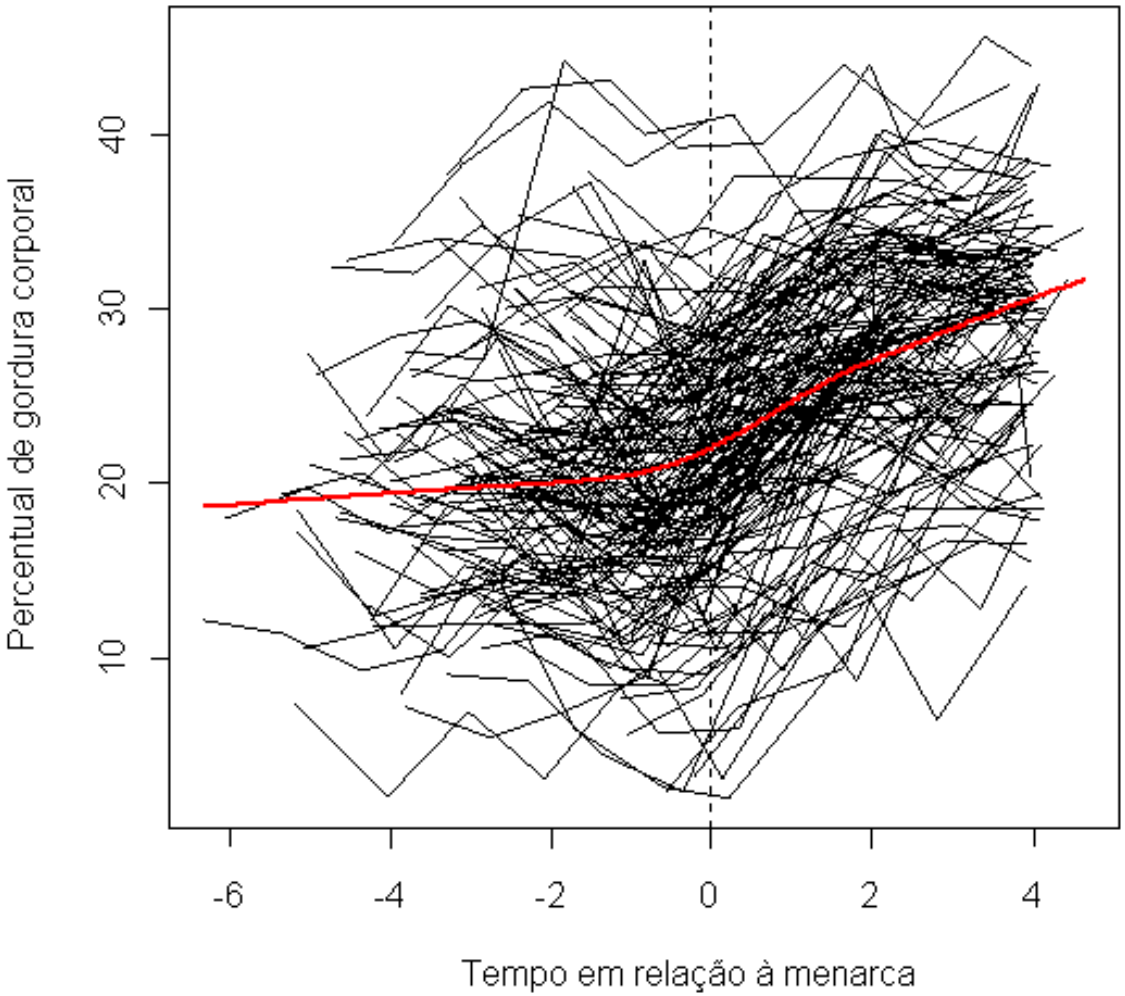


Figura: Gráfico de perfis com curva alisada

- 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 da $i-$ésima garota, $i = 1, \ldots, N$.

- 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:

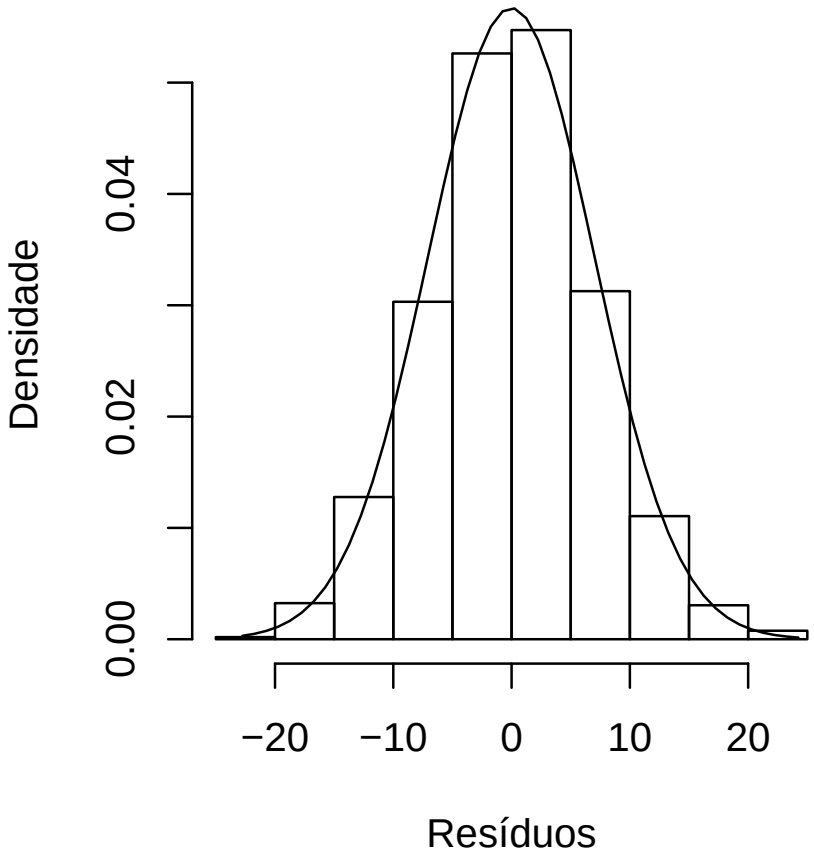
Tabela: Coeficientes de regressão estimados (efeitos fixos) e erros padrões

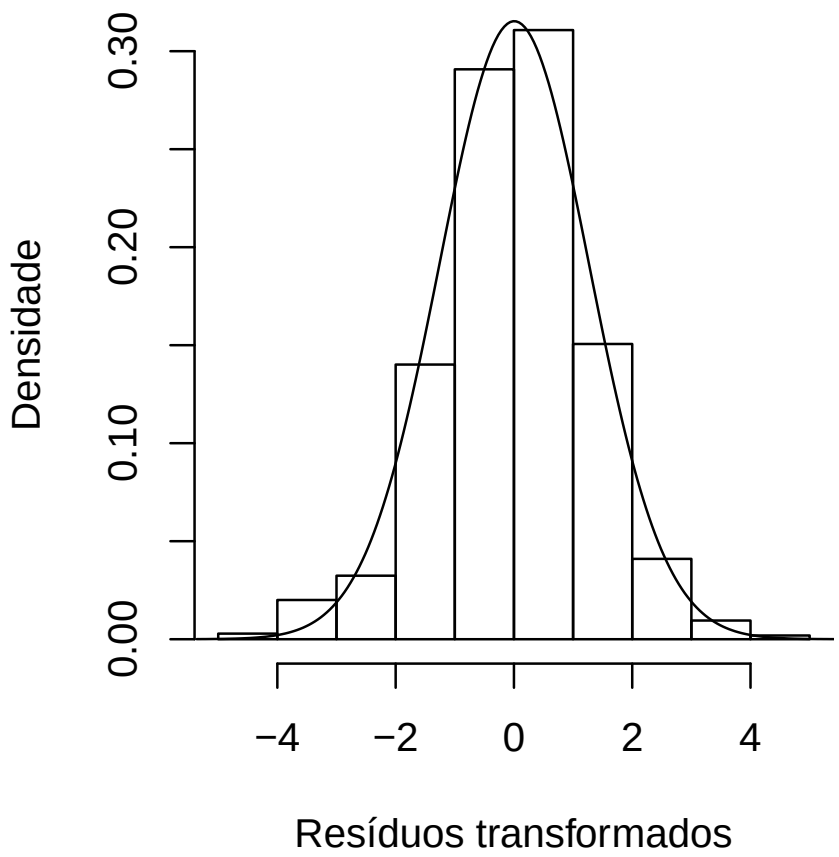
Variável	Estimativa	EP	t	p-valor
Intercepto	21,36	0,56	37,84	$<$0,001
Tempo	0,42	0,16	2,65	0,008
$(\text{Tempo})_+$	2,05	0,23	8,98	$<$0,001

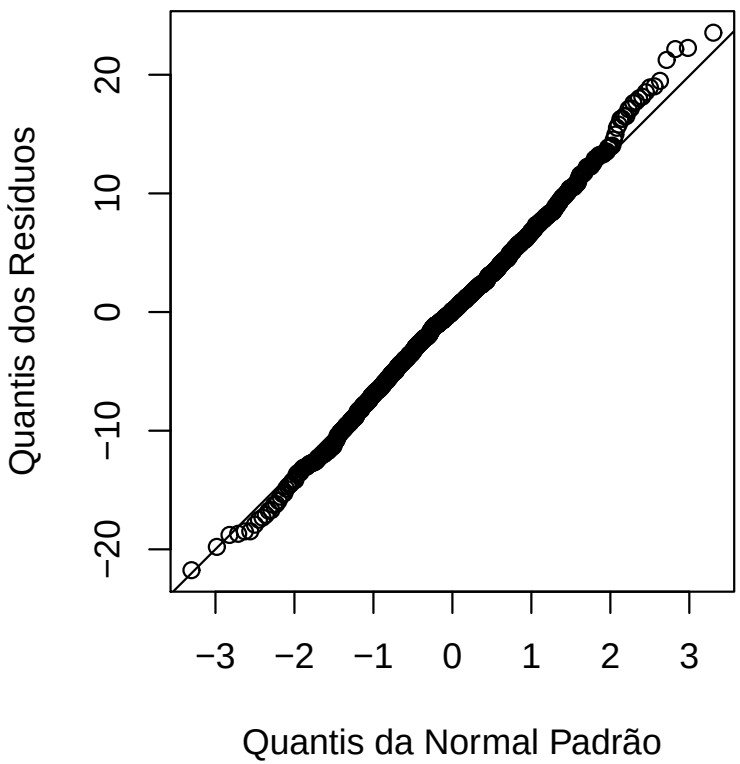
Tabela: Covariâncias estimadas para os efeitos aleatórios ($\hat{G}$) e variância estimada para os erros ($\hat{\sigma}^2$)

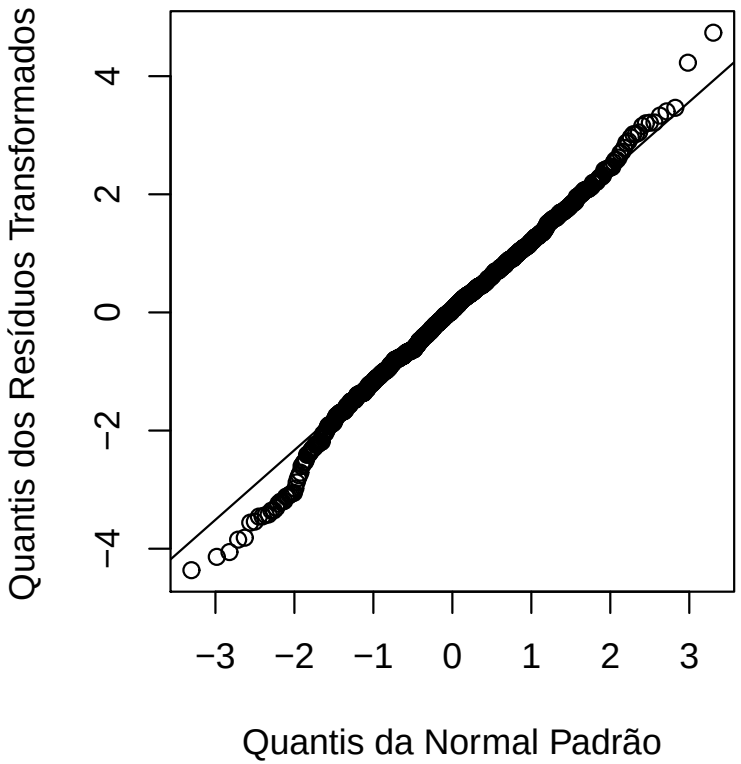
Parâmetro	Estimativa	Parâmetro	Estimativa
$Var(b_{1i}) = g_{11}$	45,94	$Cov(b_{1i}, b_{2i}) = g_{12}$	2,53
$Var(b_{2i}) = g_{22}$	1,63	$Cov(b_{1i}, b_{3i}) = g_{13}$	-6,11
$Var(b_{3i}) = g_{33}$	2,75	$Cov(b_{2i}, b_{3i}) = g_{23}$	-1,75
$Var(e_i) = \sigma^2$	9,47		

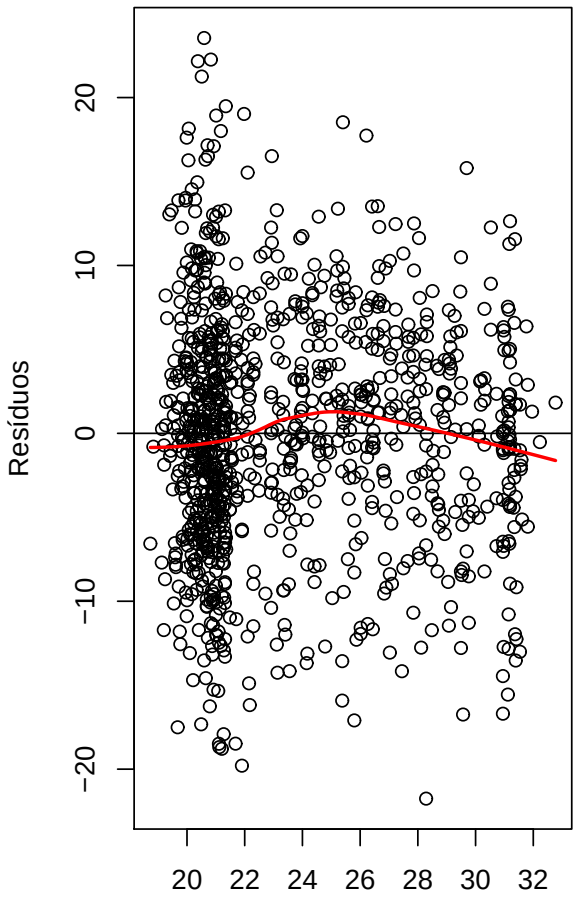
- Análise de resíduos:

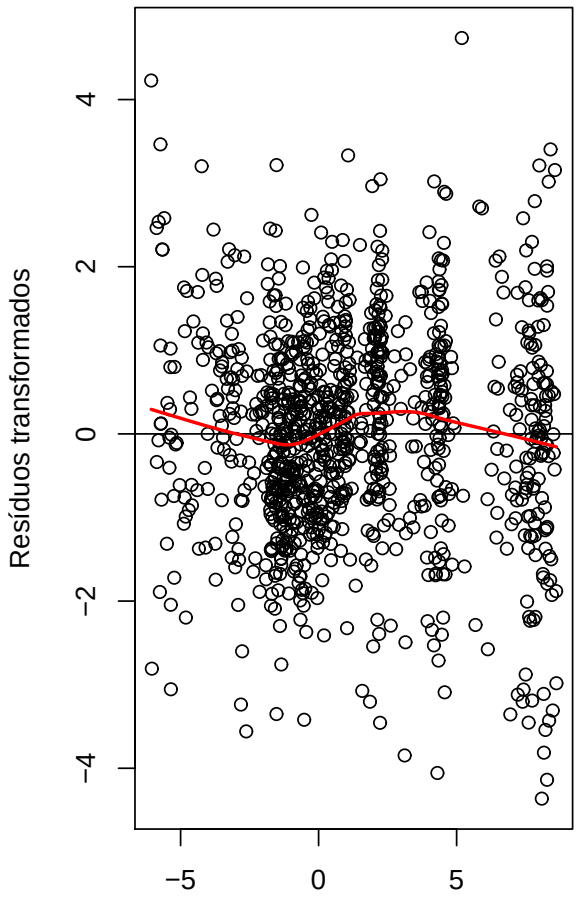


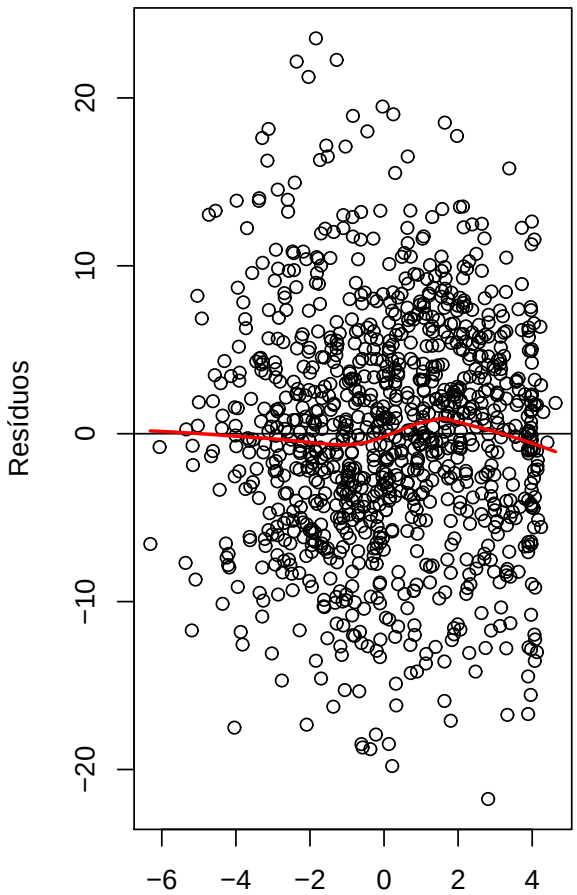


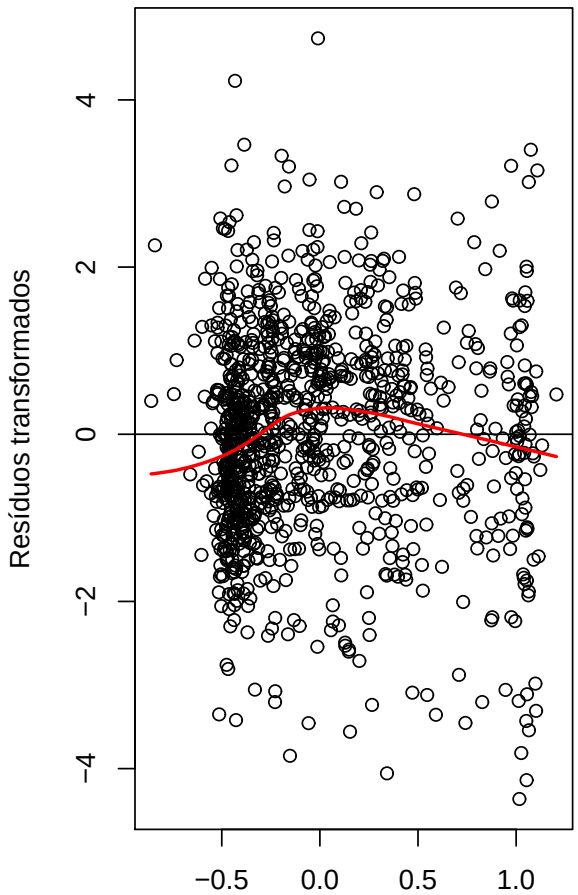

Quantis dos Resíduos
20
10
0
−10
−20
−3
−2
−1
0
1
2
3
Quantis da Normal Padrão


Quantis dos Resíduos Transformados
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
Quantis da Normal Padrão


Resíduos
20
10
0
−10
−20
20
22
24
26
28
30
32
Valores preditos


Resíduos transformados
4
2
0
−2
−4
−5
0
5
Valores preditos transformados


Resíduos
20
10
0
−10
−20
−6
−4
−2
0
2
4
Tempo em relação à menarca


Resíduos transformados
4
2
0
−2
−4
−0.5
0.0
0.5
1.0
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:

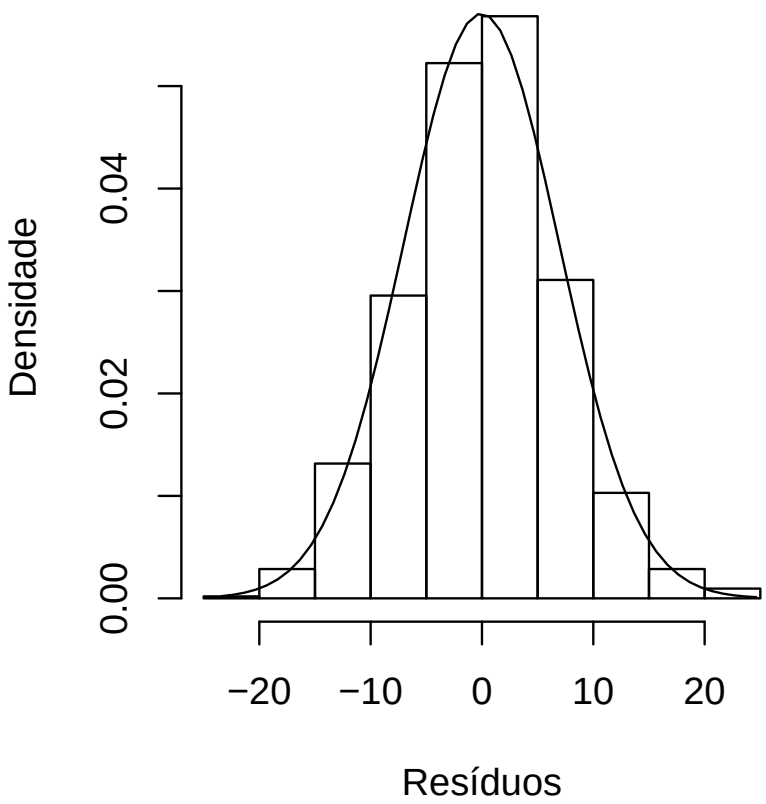
Tabela: Coeficientes de regressão estimados (efeitos fixos) e erros padrões

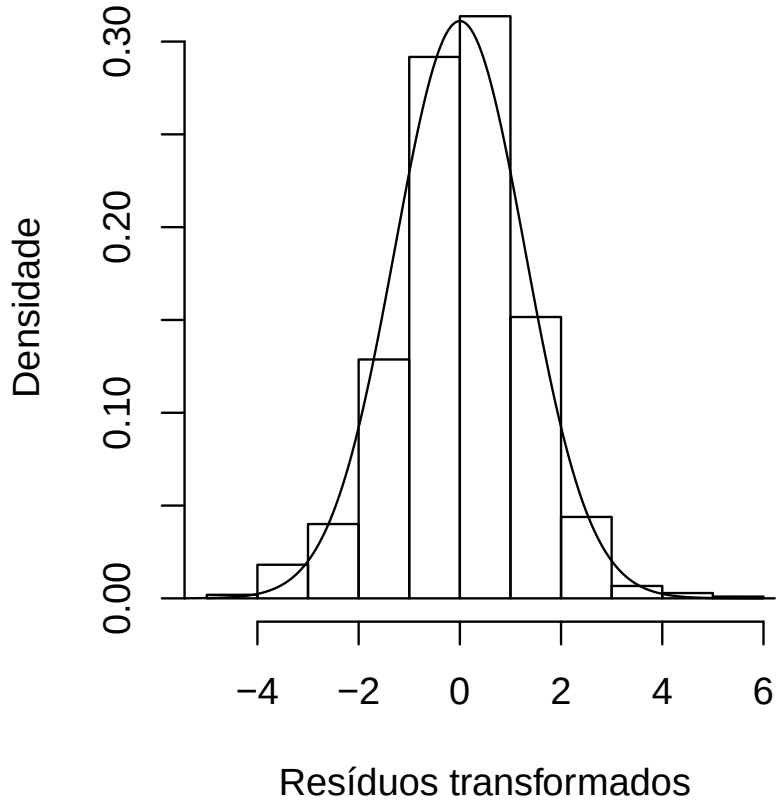
Variável	Estimativa	EP	t	p-valor
Intercepto	20,42	0,58	35,10	$<$0,001
Tempo	-0,02	0,16	-0,10	0,92
$(\text{Tempo})_+$	4,84	0,41	11,94	$<$0,001
$(\text{Tempo})_+^2$	-0,65	0,08	-8,38	$<$0,001

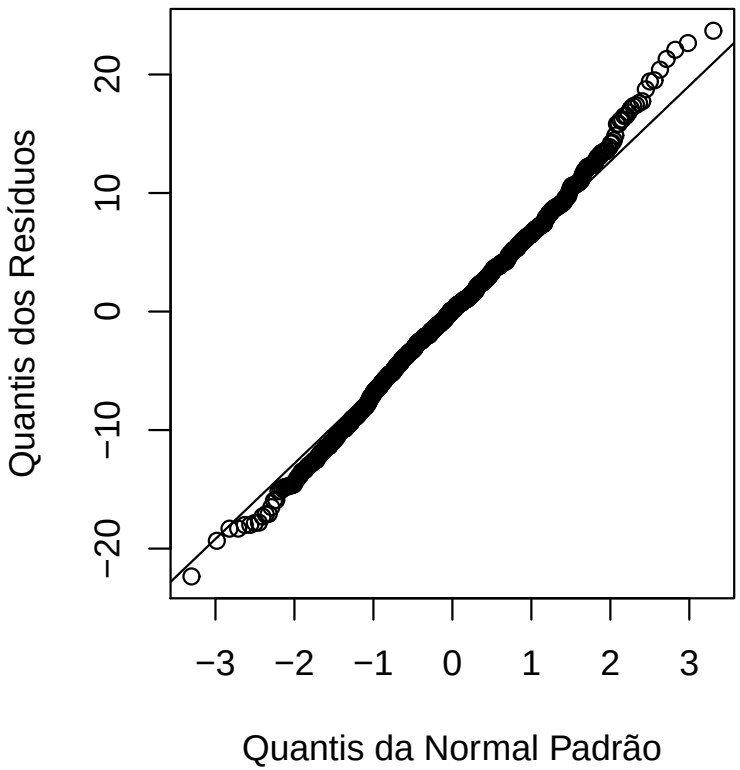
Tabela: Covariâncias estimadas para os efeitos aleatórios ($\hat{G}$) e variância estimada para os erros ($\hat{\sigma}^2$)

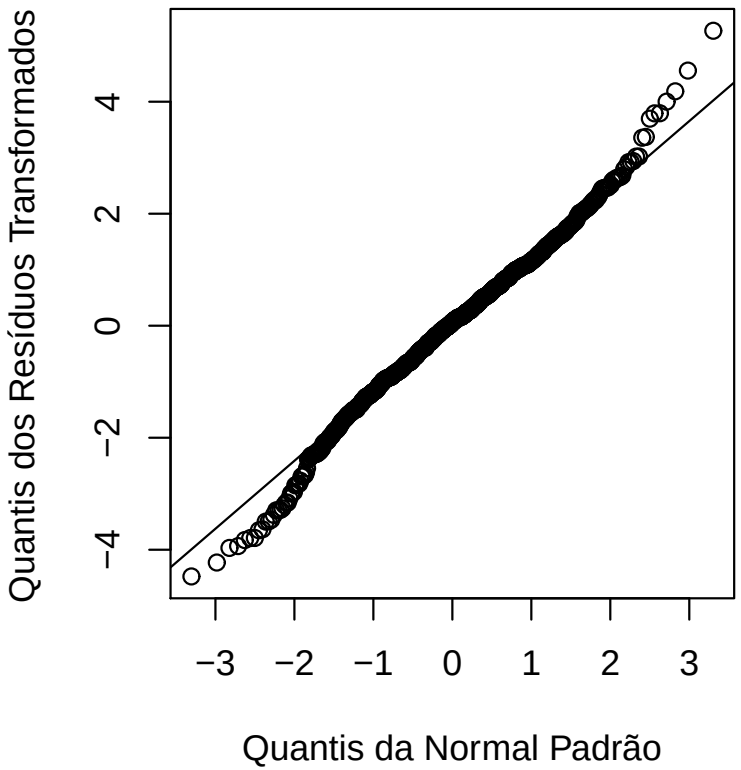
Parâmetro	Estimativa	Parâmetro	Estimativa
$Var(b_{1i}) = g_{11}$	48,06	$Cov(b_{1i}, b_{3i}) = g_{13}$	-9,59
$Var(b_{2i}) = g_{22}$	1,73	$Cov(b_{1i}, b_{4i}) = g_{14}$	0,65
$Var(b_{3i}) = g_{33}$	5,37	$Cov(b_{2i}, b_{3i}) = g_{23}$	-1,53
$Var(b_{4i}) = g_{44}$	0,12	$Cov(b_{2i}, b_{4i}) = g_{24}$	-0,17
$Cov(b_{1i}, b_{2i}) = g_{12}$	3,03	$Cov(b_{3i}, b_{4i}) = g_{34}$	-0,44
$Var(e_i) = \sigma^2$	8,03		

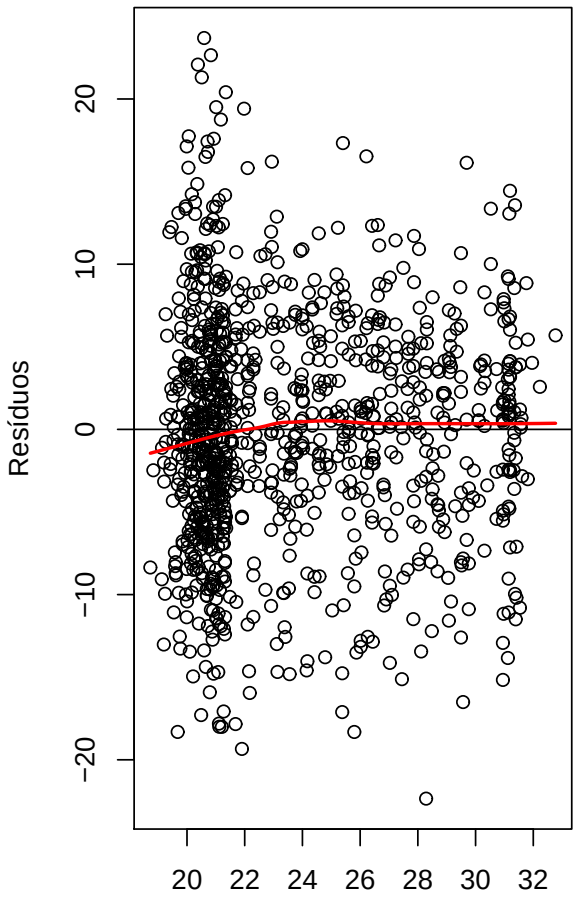
- Análise de resíduos:

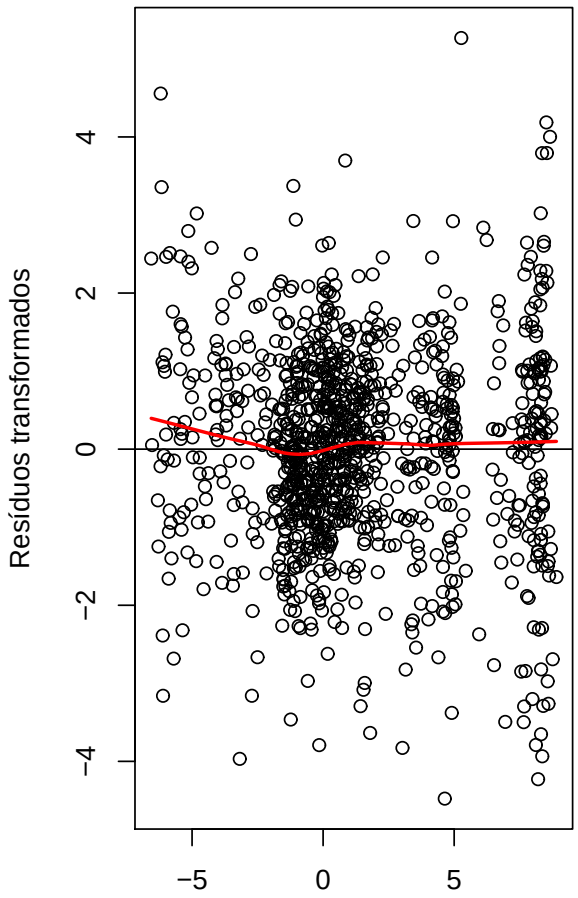


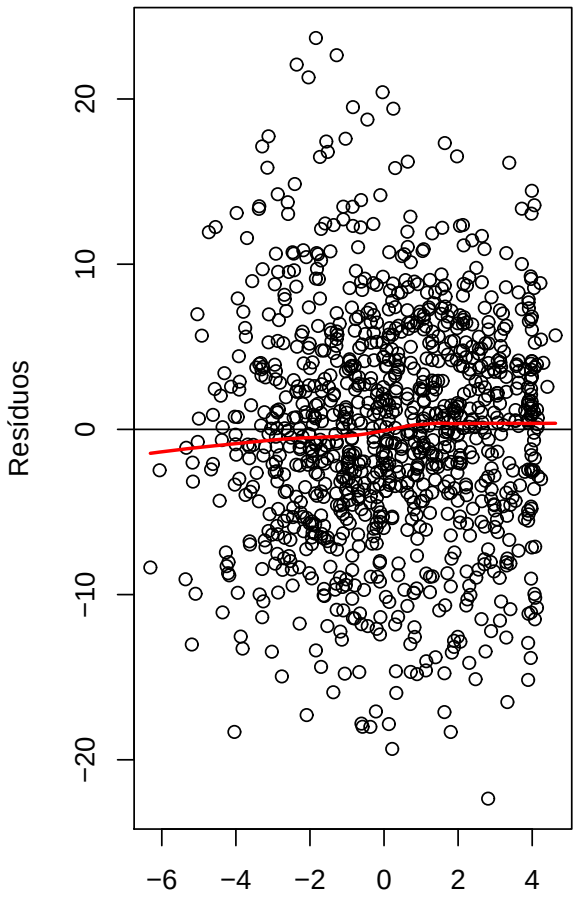


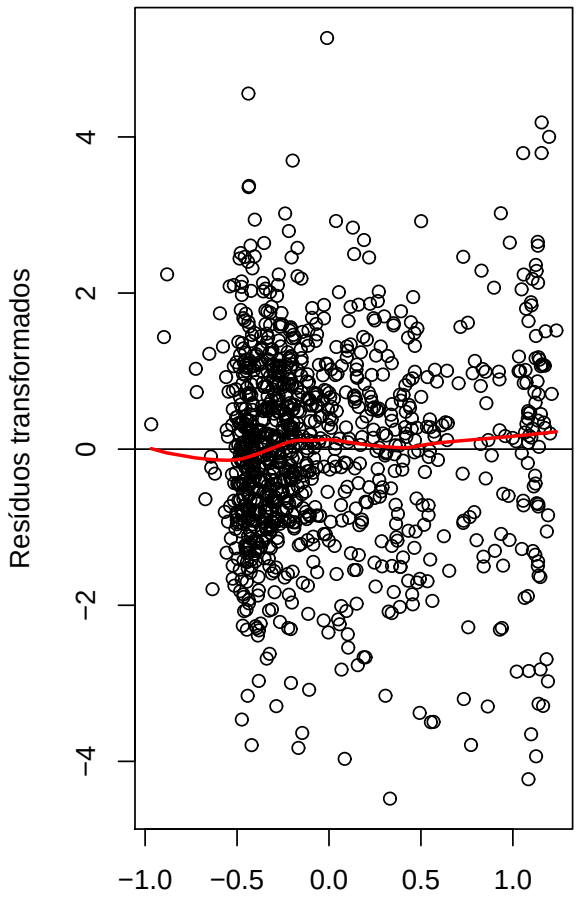

Quantis dos Resíduos
20
10
0
−10
−20
−3
−2
−1
0
1
2
3
Quantis da Normal Padrão


Quantis dos Resíduos Transformados
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
Quantis da Normal Padrão


Resíduos
20
10
0
−10
−20
20
22
24
26
28
30
32
Valores preditos


Resíduos transformados
4
2
0
−2
−4
−5
0
5
Valores preditos transformados


Resíduos
20
10
0
−10
−20
−6
−4
−2
0
2
4
Tempo em relação à menarca


Resíduos transformados
4
2
0
−2
−4
−1.0
−0.5
0.0
0.5
1.0
Tempo transformado

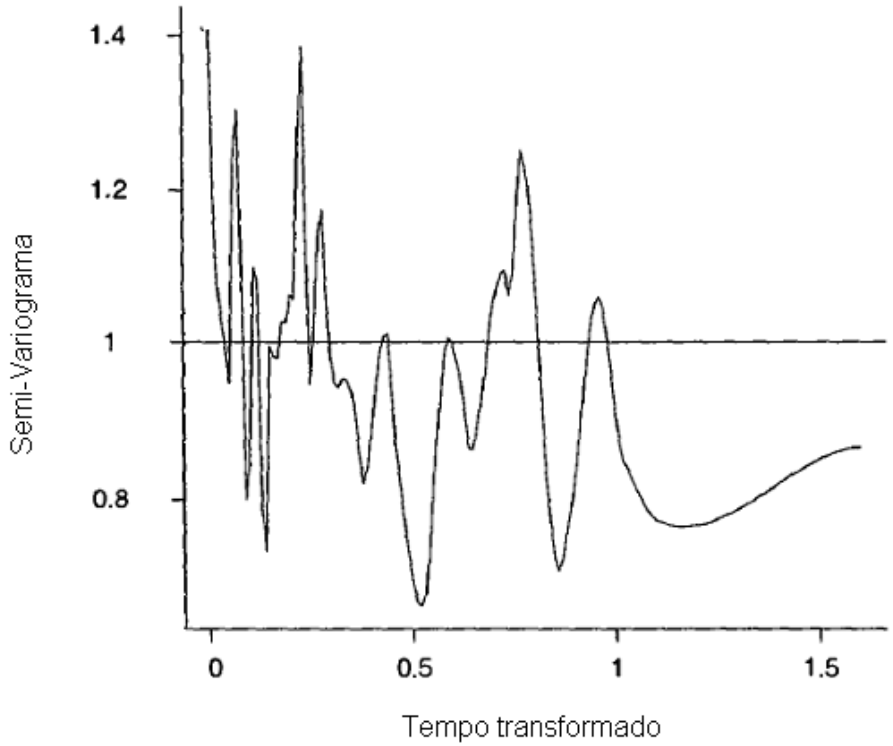


Figura: Semi-variograma empírico para os resíduos transformados

- 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.