

See discussions, stats, and author profiles for this publication at: <https://www.researchgate.net/publication/394448739>

# Análise de Dados Longitudinais com Modelos Lineares com Efeitos Mistas: Aplicações Práticas no R para Ciências da Saúde

Method · August 2025

DOI: 10.13140/RG.2.2.18814.50243

---

CITATIONS

0

READS

50

1 author:



André Moreno Morcillo

State University of Campinas (UNICAMP)

204 PUBLICATIONS 2,003 CITATIONS

SEE PROFILE

# Análise de Dados Longitudinais com Modelos Lineares com Efeitos Mistos: Aplicações Práticas no R para Ciências da Saúde

André Moreno Morcillo<sup>1</sup>

## Resumo

*Este texto apresenta uma introdução aplicada à análise de dados longitudinais por meio de Modelos Lineares com Efeitos Mistos, utilizando o pacote lme4 do R. São explorados três exemplos práticos que ilustram os principais comandos, os critérios para comparação entre modelos e a avaliação dos pressupostos estatísticos.*

A análise de estudos longitudinais, nos quais os sujeitos são avaliados duas ou mais vezes ao longo do tempo, continua a representar um grande desafio. Em ensaios clínicos prospectivos, as condições de pesquisa são rigorosamente controladas, com protocolos que especificam tanto o período de observação quanto a frequência das avaliações. Mesmo assim, há risco de perda de informações, seja devido à desistência dos participantes ou por descumprimento do protocolo do projeto.

Em algumas pesquisas os pacientes são incluídos em momentos distintos ou avaliados em períodos diferentes, resultando em desigualdade no número de medidas entre os participantes. Por exemplo, em estudos que abrangem coortes prospectivas ou retrospectivas de crianças e adolescentes atendidos em especialidades médicas, a frequência das avaliações pode variar de acordo com a gravidade da doença, a saturação do serviço, entre outros fatores, de modo que nem todos apresentam o mesmo número de medidas.

Uma das principais preocupações em estudos longitudinais é a perda de dados, seja devido à desistência ou à ausência do participante em uma ou mais avaliações. Quanto maior o período de observação, maior será o risco de perda de informação. Projetos com curtos períodos de observação e realizados em ambiente hospitalar (enfermaria, UTI etc.) têm menor chance de perda de dados. Projetos de longa duração realizados em comunidades ou ambientes ambulatoriais apresentam maior risco de perda de informação.

---

<sup>1</sup> André Moreno Morcillo, MD, PhD pela Universidade Estadual de Campinas, São Paulo, Brasil.

[morcillo@unicamp.br](mailto:morcillo@unicamp.br)

ID Lattes: 7261459499191412

ResearchGate: <https://www.researchgate.net/profile/Andre-Morcillo/publications>

Um exemplo de projeto curto em ambiente hospitalar foi conduzido por Santos (2009) na unidade de emergência do Hospital de Clínicas da Unicamp. O objetivo era avaliar o efeito da fisioterapia respiratória sobre a frequência cardíaca, frequência respiratória e saturação de oxigênio em pacientes com pneumonia comunitária. Neste estudo, os pacientes foram recrutados durante a rotina de atendimento na sala de emergência quando apresentavam sintomas de pneumonia e estavam sob observação no setor. Foram realizadas três avaliações: (T0) antes do início da fisioterapia, (T1) ao término da fisioterapia e (T2) quinze minutos após a sessão de fisioterapia. Dessa forma, cada paciente tinha três medidas de frequência cardíaca, frequência respiratória e saturação de oxigênio, sem qualquer perda de informação.

T0	T1	T2
Antes da Fisioterapia	Ao término da Fisioterapia	15' após o término da Fisioterapia
FC/FR/SatO <sub>2</sub>	FC/FR/SatO <sub>2</sub>	FC/FR/SatO <sub>2</sub>

Um projeto com longo período de seguimento foi conduzido por Carniel (2008) no Centro de Investigação em Pediatria da Unicamp, em colaboração com o Instituto Butantan, para avaliar o efeito de uma nova vacina contra hepatite B. Neste estudo, as crianças foram recrutadas na maternidade, onde receberam a primeira dose da vacina (T0). Um mês depois (T1), uma equipe do grupo de pesquisa realizou a primeira visita domiciliar para administrar a segunda dose da vacina. Nova visita domiciliar ocorreu quando as crianças completaram seis meses de idade e receberam a terceira dose da vacina (T6). Foram coletadas duas amostras de sangue para sorologia, sendo a primeira no momento da aplicação da terceira dose da vacina e a segunda um mês após (T7). O recrutamento diário foi realizado em duas maternidades, durante a internação para o parto. Nessa pesquisa, a área de cobertura era extensa, abrangendo praticamente toda a cidade de Campinas. Houve perda de dados devido à desistência de algumas famílias e ao óbito de um participante. No entanto, a principal dificuldade foi a perda de contato com as famílias por mudança de endereço ou número de telefone, o que impossibilitava o agendamento das visitas domiciliares.

T0	T1	T6	T7
Ao Nascer	1º mês	6º Mês	7º Mês
1ª Dose	2ª Dose	3ª Dose	
		1ª Coleta	2ª Coleta

Uma das principais dificuldades na análise estatística de dados longitudinais reside na **correlação** existente entre as medidas de um mesmo sujeito. Para contornar esse problema, foram desenvolvidos métodos específicos, destacando-se: a análise de variância para medidas repetidas, a análise de variância multivariada, equações estruturais, modelos lineares com efeitos mistos, regressão quantílica com efeitos mistos e regressão de dados em painel, entre outros.

#### MODELOS LINEARES COM EFEITOS MISTOS

Os Modelos Lineares com Efeitos Mistos<sup>2</sup>, extensão da Regressão Linear Múltipla<sup>3</sup>, foram desenvolvidos para lidar com dados hierárquicos, longitudinais ou com variabilidade entre grupos. Eles permitem controlar **efeitos fixos** (variáveis preditoras de interesse direto da pesquisa) e **efeitos aleatórios** (variabilidade entre grupos), tornando-os mais flexíveis e adequados para análises complexas.

Nos Modelos Lineares com Efeitos Mistos a variável desfecho deve ser quantitativa e as preditoras podem ser quantitativas ou qualitativas. As preditoras podem ser incluídas como **efeitos fixos** ou **aleatórios**.

Os **efeitos fixos** são variáveis preditoras relevantes e planejadas para o objetivo da pesquisa. Por outro lado, os **efeitos aleatórios** são variáveis que podem ter efeito sobre a variável desfecho. Ao incluir esses efeitos aleatórios, o modelo consegue ajustar melhor a variabilidade presente nos dados, resultando em estimativas mais precisas dos coeficientes de regressão dos efeitos fixos. Os efeitos aleatórios geralmente, são utilizados para modelar variações entre grupos, como diferenças entre indivíduos, hospitais ou centros de saúde, escolas etc.

---

<sup>2</sup> O método de estimação mais comum é a Máxima Verossimilhança (Maximum Likelihood - ML) ou sua variante, a Máxima Verossimilhança Restrita (Restricted Maximum Likelihood - REML).

<sup>3</sup> O método de estimação é o dos Mínimos Quadrados Ordinários (MQO).

Considere o seguinte exemplo: deseja-se avaliar o efeito de um novo método de ensino a ser aplicado a alunos do primeiro ano do ensino médio. Para tal, são selecionadas metade das escolas de uma cidade para trabalhar com o novo método de ensino e a outra metade com o tradicional. Os pesquisadores aplicam uma avaliação inicial, que é repetida após um determinado período, sendo que a **nota** será a variável desfecho. Para avaliar os resultados, optam por modelar os dados considerando como preditoras de efeito fixo **time** (momento da avaliação) o **tipo de ensino** (novo/tradicional) e a **idade** dos estudantes. Entretanto, eles sabem que não há homogeneidade no ensino entre as escolas e que isso pode influenciar a nota dos estudantes. Portanto, para controlar o efeito da falta de homogeneidade das escolas, a variável **escola** será introduzida como efeito aleatório. O modelo ficaria da seguinte forma:

$$\text{nota} = \text{time}, \text{tipo de ensino}, \text{idade} | \text{escola}$$

A forma mais adequada de expressar este modelo é:

$$\text{nota}_{ij} = \beta_0 + \beta_1 \cdot \text{time} + \beta_2 \cdot \text{tipo de ensino}_{ij} + \beta_3 \cdot \text{idade}_{ij} + u_j + \epsilon_{ij}$$

Nota<sub>ij</sub>: Nota do aluno i na escola j.

$\beta_0$ : Intercepto (média geral das notas).

$\beta_1$ : momento da avaliação (inicial / final)

$\beta_2$ : Efeito do tipo de ensino (novo vs. tradicional).

$\beta_3$ : Efeito da idade do estudante.

$u_j$ : Efeito aleatório da escola j (variação entre escolas).

$\epsilon_{ij}$ : Erro residual (variação dentro das escolas).

Pode-se dizer que os efeitos fixos são geralmente incluídos no modelo para testar hipóteses específicas previstas no projeto, enquanto os efeitos aleatórios são incluídos para modelar a variação entre grupos (pessoas, escolas, hospitais, bairros etc.).

Nos Modelos Lineares com Efeitos Mistos, não há estimativa direta dos coeficientes das variáveis de efeito aleatório. Em vez disso, o software utiliza a informação contida nelas para calcular de forma mais precisa os coeficientes de regressão das variáveis de efeitos fixos, o que ajuda a reduzir os resíduos do modelo.

Embora os Modelos Lineares com Efeitos Mistos permitam a análise de painéis não balanceados, não são robustos à falta de informação (missing) nas variáveis preditoras, podendo ser necessária a imputação de dados. Isso envolve substituir os valores ausentes por estimativas baseadas em outras partes dos dados, o que pode ser feito usando técnicas como imputação por média, imputação múltipla etc.

Por se tratar de um modelo linear, é fundamental realizar uma análise exploratória inicial da variável resposta e das preditoras quantitativas, visando identificar outliers. Após a finalização do modelo, deve-se analisar os **resíduos**, que devem ter distribuição normal com média zero. A confirmação de que os pressupostos do modelo foram atendidos garante a confiabilidade dos coeficientes estimados.

Os pressupostos dos Modelos Lineares com Efeitos Mistos são:

1. Deve haver **correlação linear** entre a variável resposta (quantitativa) e as preditoras (quantitativas).
2. Não deve haver **colinearidade** entre as variáveis preditoras quantitativas.
3. Os resíduos devem ter **distribuição normal** com média zero.
4. Deve haver **homocedasticidade** e **independência** dos resíduos.

#### ESTRUTURA DO BANCO DE DADOS

Um banco de dados adequadamente estruturado é condição fundamental para se obter um bom resultado.

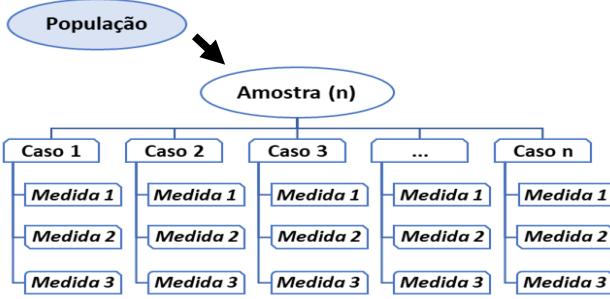
Uma situação confortável ocorre em projetos nos quais os sujeitos são avaliados com a mesma periodicidade e têm o mesmo número de medidas. Essas bases de dados são denominadas **painéis balanceados**. Quando os sujeitos são avaliados em períodos diferentes ou não possuem o mesmo número de medidas, são denominados **painéis desbalanceados**. A estrutura do banco de dados deve ser adequada à técnica de análise e ao software que será utilizado.

Os dados de medidas repetidas podem ser organizados de duas maneiras: bancos de dados **largos** (wide format) e **longos** (long format). Por exemplo, a Anova para medidas repetidas<sup>4</sup> ou a Manova no SPSS requerem banco de dados largos. A análise de dados em painel, os Modelos Lineares com Efeitos Mistos e a Regressão Quantílica com Efeitos Mistos exigem banco de dados longos.

Para demonstrar os dois tipos de banco de dados, vamos considerar uma pesquisa em que uma amostra de **n** crianças foi selecionada e a altura foi avaliada em três ocasiões ao longo do tempo. As variáveis são: identificação da criança (id), data das medições (data\_1, data\_2, data\_3) e altura (altura\_1, altura\_2, altura3).

---

<sup>4</sup> Para a Anova com dados repetidos o SPSS exige um banco de dados largo, porém, no Stata e no R o banco de dados deverá ser do tipo longo.



O banco de dados do tipo largo terá sete colunas e cada caso ocupará uma única linha da planilha. As medidas de altura e respectivas datas são colocadas em **colunas** diferentes (***id, altura\_1, altura\_2, altura\_3, data\_1, data\_2, data\_3***).

No banco de dados tipo longo, os dados de cada avaliação são colocados em **linhas** diferentes da planilha. Portanto, cada criança terá no banco de dados longo o número de linhas equivalente ao número de medidas que foram realizadas. Teremos somente quatro colunas: ***id, data, altura, medida***. Os três valores da altura de cada criança estarão digitados na coluna ***altura***. Neste caso as colunas ***id*** e ***medida*** são fundamentais para o software identificar o sujeito e a ordem da medida de cada criança.

Banco de dados tipo largo

<b><i>id</i></b>	<b><i>Data_1</i></b>	<b><i>Altura_1</i></b>	<b><i>Data_2</i></b>	<b><i>Altura_2</i></b>	<b><i>Data_3</i></b>	<b><i>Altura_3</i></b>
1	10/05/2003	101	10/12/2003	106	10/06/2004	110
2	10/05/2003	104	10/12/2003	105	10/06/2004	108
3	10/05/2003	102	10/12/2003	107	10/06/2004	109
...	...	...	...	...	...	...
n	10/05/2003	103	10/12/2003	104	10/06/2004	108

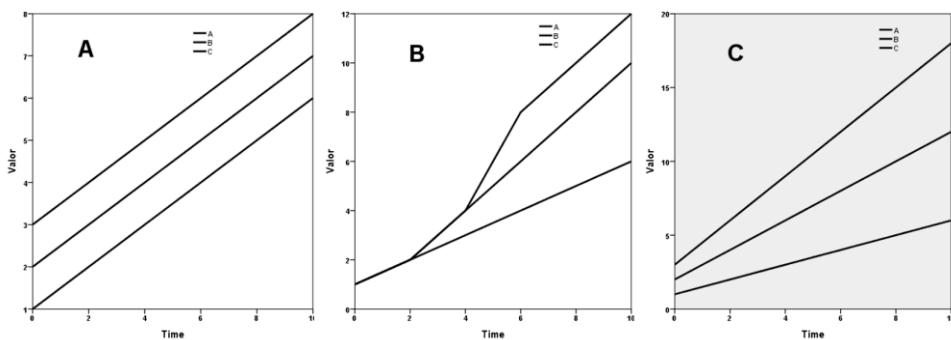
Banco de dados tipo longo

<b><i>id</i></b>	<b><i>Medida</i></b>	<b><i>Datas</i></b>	<b><i>Altura</i></b>
1	1	10/05/2003	101
1	2	10/12/2003	106
1	3	10/06/2004	110
2	1	10/05/2003	104
2	2	10/12/2003	105
2	3	10/06/2004	108
3	1	10/05/2003	102
...		...	...
n	1	10/05/2003	103
n	2	10/12/2003	104
n	3	10/06/2004	108

## MODELOS LINEARES COM EFEITOS MISTOS NA ANÁLISE DE DADOS LONGITUDINAIS

Os Modelos Lineares com Efeitos Mistos são recomendados para a análise de dados longitudinais. São particularmente úteis quando as medições são desbalanceadas, ou seja, quando os sujeitos têm número diferente de observações ou são feitas em intervalos de tempo irregulares.

É fundamental o planejamento das variáveis preditoras de efeitos fixos e quais são os efeitos aleatórios que devem ser incorporados ao modelo. Dois aspectos devem ser considerados na definição dos efeitos aleatórios na análise de dados longitudinais: o **intercepto aleatório**<sup>5</sup> e o **slope aleatório**<sup>6</sup>. O intercepto e o slope sempre são introduzidos nos modelos como efeitos aleatórios. Isto torna a análise de dados longitudinais muito mais simples que os Modelos Lineares com Efeitos Mistos para análise de dados hierárquicos, muito comuns na epidemiologia. O **intercepto aleatório** modela a heterogeneidade no início da avaliação e o **slope aleatório** modela a heterogeneidade na inclinação das curvas individuais. No gráfico abaixo, em **A** temos heterogeneidade somente no intercepto; em **B** temos heterogeneidade somente no slope e em **C** temos heterogeneidade no intercepto e no slope.

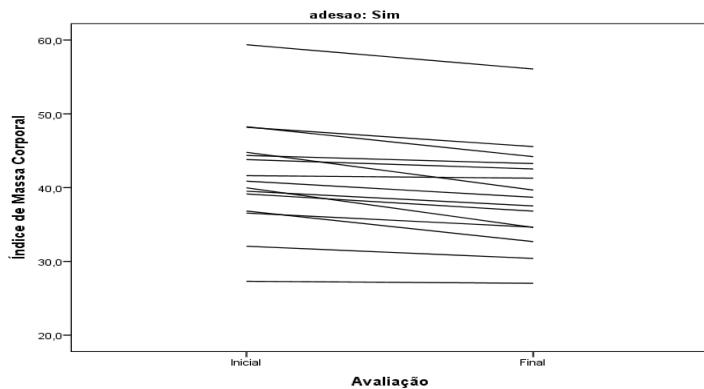


Na figura abaixo, são apresentados os dados do índice de massa corporal (IMC) de 15 crianças acompanhadas em um ambulatório de obesidade. Elas foram avaliadas em duas ocasiões com o objetivo de estudar o efeito de uma intervenção sobre o IMC. Observa-se grande heterogeneidade no IMC inicial (intercepto aleatório), evidenciando que o IMC no início da pesquisa não era homogêneo entre as crianças. Nesse contexto, devemos considerar modelar o **intercepto aleatório**. Além disso, a inclinação das curvas das crianças também é heterogênea. Portanto, também é recomendável modelar o

<sup>5</sup> Representa a heterogeneidade da variável dependente entre os sujeitos no ponto inicial da observação.

<sup>6</sup> Representa a heterogeneidade na inclinação das curvas dos sujeitos durante todo o período de observação.

**slope aleatório.** Um gráfico apresentando as curvas individuais é útil para o planejamento inicial dos efeitos aleatórios, pois permite avaliar visualmente a heterogeneidade no intercepto e no slope. Posteriormente, usando métodos estatísticos, os efeitos aleatórios mais adequados serão confirmados.



O objetivo em um estudo longitudinal é avaliar o comportamento da variável desfecho em relação ao tempo (**time**) considerando o efeito de **k** preditoras ( $x_1, x_2, x_3, \dots, x_k$ ) de efeito fixo e os efeitos aleatórios. Então, as perguntas fundamentais são:

Houve variação da variável desfecho em relação ao tempo (**time**) decorrido entre as avaliações, após o ajuste pelas preditoras de efeito fixo? Qual é o sentido e a intensidade da variação? As respostas a estas perguntas serão encontradas no **coeficiente de regressão** da variável tempo (**time**).

Além da variável desfecho e das preditoras, duas outras são fundamentais para análise de dados longitudinais. Uma variável que identifica cada sujeito da pesquisa, neste texto apelidada de “**id**”<sup>7</sup> e outra, que representa o momento em que a medida foi realizada, aqui designada por “**tempo**” ou “**time**”. Qualquer outro nome pode ser utilizado.

Um ótimo exemplo deste tipo de análise foi publicado no Brasil por Spyrides, Struchiner, Barbosa e Kac (2005), com o objetivo foi modelar o ganho de peso e comprimento de 479 crianças recém-nascidas de um centro de saúde do Rio de Janeiro, que foram pesadas e medidas aos 15 dias de vida e aos dois, seis e nove meses. O peso e o comprimento foram as variáveis desfecho. Foram consideradas variáveis preditoras relativas à gestação, parto, aleitamento materno e nível socioeconômico das famílias.

---

<sup>7</sup> Id – abreviação de “identificação” ou “Identification”

Os autores discutem com detalhes a questão da seleção das variáveis de efeito fixo e os efeitos aleatórios.

#### MODELOS LINEARES COM EFEITOS MISTOS COM O COMANDO LMER() DO PACOTE LME4 DO R

A seguir, apresentamos uma síntese dos comandos mais empregados na análise.

Inicialmente instalamos as bibliotecas **lme4** (Bates et al. 2015, 2017), **haven**, **lmerTest** e **ggplot2** com o comando `install.packages()`. O carregamento das bibliotecas é feito com o comando `library()`.

```
install.packages("lme4")
install.packages("haven")
install.packages("lmerTest")
install.packages("ggplot2")
library(lme4)
library(lmerTest)
library(haven)
library(ggplot2)
```

O comando que executa o modelo linear de efeitos mistos é `lmer()` do pacote **lme4** cuja sintaxe é:

```
modelo <- lmer ( y ~ time + x1 + x2 + ... + xk + (1 | id) , data = bd, REML = FALSE)
```

Onde **y** é a variável desfecho; **id** é a identificação dos sujeitos, **time** é a preditora que representa o tempo ou a ordem das medidas; **x1, x2, ..., xk** é o conjunto de preditoras de efeito fixo e **(1 | id)** é o termo aleatório para ajustar o intercepto aleatório e **bd** é o nome do banco de dados.

Os quatro possíveis efeitos aleatórios são:

**(1 | id)** – somente efeito aleatório no intercepto

**(0 + time | id)** – somente efeito aleatório no slope

**( time | id )** – efeitos aleatórios no intercepto e no slope correlacionados

**( time || id )** - efeitos aleatórios no intercepto e no slope não correlacionados

Para avaliar estatisticamente os efeitos fixos usamos o comando `anova(modelo)` e para os efeitos aleatórios o comando `ranova(modelo)` ambos do pacote **lmerTest**.

```
anova(modelo)
ranova(modelo)
```

Na comparação de dois ou mais modelos usamos os critérios **AIC**<sup>8</sup> ou **BIC**<sup>9</sup> ou fazemos o teste **anova()**.

```
AIC(modelo1, modelo2, ..., modelok)  
BIC(modelo1, modelo2, ..., modelok)  
anova(modelo1, modelo2, ..., modelok)
```

### Análise dos resíduos

Chamamos de **resíduo bruto** à diferença que há entre o valor predito pelo modelo e o valor real [ $\text{Resíduo} = \hat{y} - y$ ]. Os resíduos brutos, podem ser transformados em resíduos padronizados (**studentizados**)<sup>10</sup>. Os resíduos brutos são obtidos pelo comando **residuals(modelo)** e os resíduos studentizados pelo comando **rstudent(modelo)**.

```
res_bruto <- residuals(modelo)  
res_stud <- rstudent (modelo)
```

Avaliamos a normalidade dos resíduos pelo teste de Shapiro-Wilk, sendo que p-valor menor ou igual a 0,05 rejeita a normalidade dos resíduos.

```
shapiro.test(res_bruto)  
shapiro.test(res_stud)
```

O gráfico **qqnorm()** também é eficiente para avaliar a normalidade dos resíduos:

```
qqnorm(res_bruto) ; qqline(res_bruto)  
qqnorm(res_stud) ; qqline(res_stud)
```

Os valores preditos pelo modelo são obtidos pelo comando **predict(modelo)**.

```
pred <- predict(modelo)
```

O comando **plot(pred, res\_stud)**<sup>11</sup> gera um gráfico apresentando a distribuição dos resíduos padronizados em relação aos valores preditos.

```
plot(pred, res_stud, ylim = c(-4,4), ylab = "Resíduos Studentizados", xlab = "Valores preditos pelo modelo",  
      pch=21, col="blue", bg="blue", cex=1, cex.lab= 0.8, cex.axis = 0.75)  
abline(h = 0, v = 0, col = "red", lwd=3)
```

---

<sup>8</sup> O AIC (Critério de Informação de Akaike) é uma medida usada para comparar modelos ajustados com lmer() e outros modelos estatísticos.

<sup>9</sup> O BIC (Critério de Informação Bayesiano) é um critério utilizado para comparar modelos ajustados, como os modelos lmer(), levando em conta o ajuste aos dados e a complexidade do modelo.

<sup>10</sup>  $rs_i = \frac{\varepsilon_i}{s_\varepsilon}$  e  $s_\varepsilon = \sqrt{\frac{\sum(\varepsilon_i)^2}{n-p-1}}$ , onde p é o número de parâmetros do modelo incluindo a constante e n é o número de casos.

<sup>11</sup> O comando **plot(modelo4)** gera um gráfico mostrando a distribuição dos resíduos brutos em relação aos valores preditos pelo modelo.

Podemos avaliar a eficiência do modelo de forma gráfica com os comandos:

```
plot(y, pred) ; abline(0, 1, col="red")
```

No exemplo que vamos analisar a seguir, um grupo de 43 crianças diabéticas foi avaliado em duas oportunidades, com o objetivo de estimar o impacto da doença sobre o crescimento e composição corporal (Paulino, Lemos-Marini, Guerra-Junior, Morcillo, 2013).

O banco de dados (**bd**) do tipo longo contém a variável desfecho que é o z-score do IMC (**zimc**) sendo as preditoras de efeito fixo **time**, **sexo**, **renda**, **escolaridade** e **t\_pesquisa**. A variável **time** foi codificada como **1** (primeira avaliação) ou **2** (segunda avaliação). A variável **id** recebe a identificação dos pacientes.

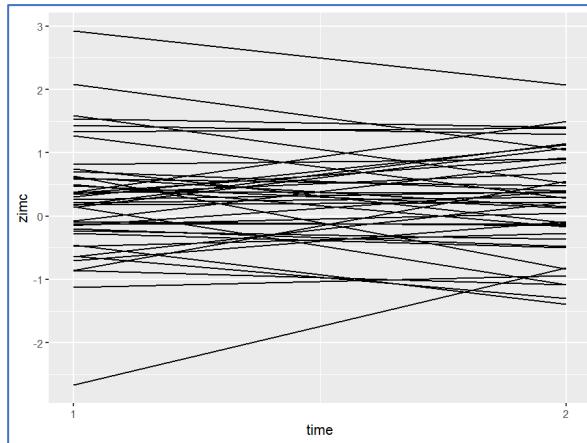
<b>zimc</b>	- z-score do índice de massa corporal
<b>id</b>	- identificação do paciente
<b>time</b>	- (1) – 1 <sup>a</sup> avaliação e (2) – 2 <sup>a</sup> avaliação
<b>sexo</b>	- (1) – masculino e (0) – feminino
<b>escolaridade da mãe</b>	- (1) - ≤ 8 anos e (2) – 9   –   11anos e 3 ≥ 12 anos
<b>renda per capita</b>	- renda familiar per capita (salário-mínimo) :
<b>t_pesquisa</b>	- tempo (anos) entre as duas avaliações

Sabemos que além da doença crônica (diabetes), os fatores sociais também são determinantes no crescimento e composição social. Assim, não basta realizar um teste t de Student para amostras pareadas comparando as médias da primeira e segunda avaliação e encerrar a análise. Temos que comparar as médias inicial e final levando em consideração as preditoras sexo, escolaridade da mãe, renda familiar e o tempo de acompanhamento na pesquisa. Esta análise multivariada não pode ser realizada por Regressão Linear Múltipla. Os Modelos Lineares com Efeitos Mistos permitem a adequada análise destes dados.

Inicialmente avaliamos graficamente o comportamento das curvas individuais de **zimc**. Para obter o gráfico com as curvas individuais de todos os casos usamos o comando **ggplot()** do pacote **ggplot2**.

```
ggplot(data = bd, aes(x = time, y = zimc, group = id)) + geom_line() + scale_x_continuous(breaks=seq(0, 2, 1))
```

Verifica-se no gráfico abaixo que há heterogeneidade nos valores iniciais (**intercepto**) e no **slope** do z-score do IMC, ou seja, há grande variação do **zimc** na primeira avaliação e muita variabilidade na inclinação das curvas individuais. Estes dados são indicativos de que é recomendável modelar os efeitos aleatórios no intercepto e no slope.



Vamos inicialmente avaliar quatro modelos simples, incluindo somente o desfecho **zimc**, a preditora **time** como fator<sup>12</sup> e os quatro possíveis efeitos aleatórios (**1 | id**)<sup>13</sup>, (**0 + time | id**)<sup>14</sup>, (**time | id**)<sup>15</sup> e (**time || id**)<sup>16</sup>. O objetivo desta etapa é identificar qual **efeito aleatório** é o mais adequado aos nossos dados.

```
m0 <- lmer(zimc ~ as.factor(time) + (1 | id), data = bd, REML = TRUE)
m1 <- lmer(zimc ~ as.factor(time) + (0 + time | id), data = bd, REML = TRUE)
m2 <- lmer(zimc ~ as.factor(time) + (time | id), data = bd, REML = TRUE)
m3 <- lmer(zimc ~ as.factor(time) + (time || id), data = bd, REML = TRUE)
```

O modelo **m2** com efeitos aleatórios somente no slope não convergiu. Entre os modelos **m0**, **m1** e **m3** o primeiro apresenta o menor valor de **AIC**, portanto, será o nosso modo de avaliação inicial. Os valores de AIC são apresentados abaixo.

---

<sup>12</sup> Neste caso a variável time não expressa o intervalo de tempo (meses, semanas, dias), é somente uma indicação do momento da avaliação, portanto deve ser inserida no modelo na forma de variável do tipo qualitativa ou fator. Quando a variável preditora é qualitativa e introduzida no modelo com a opção `as.factor()` ela é automaticamente transformada em variáveis indicadoras (dummy).

<sup>13</sup> Somente Intercepto aleatório

<sup>14</sup> Somente o slope aleatório

<sup>15</sup> intercepto aleatório por “id” e slope por “time” correlacionados

<sup>16</sup> Intercepto e slope aleatórios não correlacionados

> AIC(m0, m1, m3)		
df	AIC	
m0	3	204.7010
m1	3	219.2234
m3	4	206.7010

Agora, executaremos o modelo **m0** incluindo as variáveis preditoras time, sexo, escolaridade, renda e tempo de pesquisa, com intercepto aleatório por **id** ( $1 | id$ ) e, a cada etapa, eliminaremos a preditora com maior p-valor. As preditoras qualitativas são automaticamente transformadas em variáveis **dummy** pela opção **as.factor()**.

```
m0 <- lmer(zimc ~ as.factor(time) + as.factor(sexo) + as.factor(escolaridade) + renda + t_pesquisa + (1 | id), data = bd, REML = TRUE)
```

```
> summary(m0)
Linear mixed model fit by REML. t-tests use Satterthwaite's method [lmerModLmerTest]
Formula:
zimc ~ as.factor(time) + as.factor(sexo) + as.factor(escolaridade) +
    renda + t_pesquisa + (1 | id)
Data: bd

REML criterion at convergence: 193.6

Scaled residuals:
    Min      Q1     Median      Q3      Max 
-2.88041 -0.46282 -0.02474  0.44859  1.63185 

Random effects:
Groups   Name        Variance Std.Dev.
id      (Intercept) 0.4415  0.6644
Residual          0.2716  0.5212
Number of obs: 86, groups: id, 43

Fixed effects:
            Estimate Std. Error       df t value Pr(>|t|)    
(Intercept) 0.912448  1.263297 37.146749  0.722  0.4746  
as.factor(time)2 -0.006977  0.112396 42.000001 -0.062  0.9508  
as.factor(sexo)1  0.450290  0.237345 36.999999  1.897  0.0656 .  
as.factor(escolaridade)2 -0.583091  0.278695 36.999999 -2.092  0.0433 *  
as.factor(escolaridade)3 -0.727374  0.944951 36.999999 -0.770  0.4463  
renda           -0.105714  0.342138 36.999999 -0.309  0.7591  
t_pesquisa       -0.164920  0.300166 37.000000 -0.549  0.5860  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Correlation of Fixed Effects:
              (Intr) as.fctr(t)2 as.(O)1 as.fctr(s)2 as.(O)3 renda
as.fctr(t)2 -0.044
as.fctr(s)1 -0.145  0.000
as.fctr(s)2 -0.004  0.000   -0.016
as.fctr(s)3  0.219  0.000   -0.070  0.233
renda         -0.255  0.000   -0.115 -0.285   -0.556
t_pesquisa    -0.978  0.000    0.076 -0.002   -0.146  0.111
> |
```

No output acima podemos observar que os maiores p-valores são das variáveis **time** e **renda**. Embora a variável **time** tenha o maior p-valor, ela não será removida porque é a nossa principal preditora. A variável **renda** tem  $p = 0,7591$  e será removida do modelo.

```
m0 <- lmer(zimc ~ as.factor(time) + as.factor(sexo) + as.factor(escolaridade) + t_pesquisa + (1 | id), data = bd, REML = TRUE)
```

```

> summary(m0)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: zimc ~ as.factor(time) + as.factor(sexo) + as.factor(escolaridade) +
   t_pesquisa + (1 | id)
Data: bd

REML criterion at convergence: 193.4

Scaled residuals:
    Min      Q1     Median      Q3      Max 
-2.88597 -0.45422 -0.00363  0.43608  1.65632 

Random effects:
Groups   Name        Variance Std.Dev. 
id      (Intercept) 0.4277  0.6540  
Residual          0.2716  0.5212  
Number of obs: 86, groups: id, 43

Fixed effects:
            Estimate Std. Error   df t value Pr(>|t|)    
(Intercept) 0.812999  1.207011 38.165127 0.674 0.5046  
as.factor(time)2 -0.006977 0.112396 42.000001 -0.062 0.9508  
as.factor(sexo)1 0.441844  0.232943 37.999999 1.897 0.0655 .  
as.factor(escolaridade)2 -0.607602 0.263967 37.999999 -2.302 0.0269 *  
as.factor(escolaridade)3 -0.889655 0.776141 37.999999 -1.146 0.2589  
t_pesquisa      -0.154642 0.294745 37.999999 -0.525 0.6029  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
              (Intr) as.fctr(t)2 as.(O1 as.fctr(s)2 as.(O3
as.fctr(t)2 -0.047
as.fctr(s)1 -0.181  0.000
as.fctr(s)2 -0.082  0.000   -0.051
as.fctr(s)3  0.096  0.000   -0.163  0.094
t_pesquisa   -0.988 0.000    0.089  0.031   -0.103

```

Agora, podemos eliminar a preditora **t\_pesquisa**, visto que tem o maior p-valor.

```
m0 <- lmer(zimc ~ as.factor(time) + as.factor(sexo) + as.factor(escolaridade) + (1 | id), data = bd, REML = TRUE)
```

```

> summary(m0)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: zimc ~ as.factor(time) + as.factor(sexo) + as.factor(escolaridade) + 
   (1 | id)
Data: bd

REML criterion at convergence: 193.1

Scaled residuals:
    Min      Q1     Median      Q3      Max 
-2.90361 -0.46611 -0.00005  0.42602  1.69642 

Random effects:
Groups   Name        Variance Std.Dev. 
id      (Intercept) 0.4173  0.6460  
Residual          0.2716  0.5212  
Number of obs: 86, groups: id, 43

Fixed effects:
            Estimate Std. Error   df t value Pr(>|t|)    
(Intercept) 0.187144  0.182668 47.102721 1.025 0.3108  
as.factor(time)2 -0.006977 0.112396 42.000001 -0.062 0.9508  
as.factor(sexo)1 0.452779  0.229843 38.999999 1.970 0.0560 .  
as.factor(escolaridade)2 -0.603353 0.261380 38.999999 -2.308 0.0264 *  
as.factor(escolaridade)3 -0.931434 0.764838 38.999999 -1.218 0.2306  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
              (Intr) as.fctr(t)2 as.(O1 as.fctr(s)2
as.fctr(t)2 -0.308
as.fctr(s)1 -0.609  0.000
as.fctr(s)2 -0.341  0.000   -0.054
as.fctr(s)3 -0.033  0.000   -0.155  0.098

```

Na próxima etapa, eliminamos a preditora **sexo** com p-valor maior que 0,05.

```
m0 <- lmer(zimc ~ as.factor(time) + as.factor(escolaridade) + (1 | id), data = bd, REML = TRUE)
```

```

> summary(m0)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: zimc ~ as.factor(time) + as.factor(escolaridade) + (1 | id)
Data: bd

REML criterion at convergence: 195.8

Scaled residuals:
    Min      Q1     Median      Q3      Max 
-2.72552 -0.48637 -0.05427  0.46989  1.74014 

Random effects:
Groups   Name        Variance Std.Dev. 
id      (Intercept) 0.4571  0.6761  
Residual          0.2716  0.5212  
Number of obs: 86, groups: id, 43

Fixed effects:
            Estimate Std. Error   df t value Pr(>|t|)    
(Intercept) 0.406230  0.149278 52.926738 2.721 0.00878 ** 
as.factor(time)2 -0.006977 0.112396 42.000000 -0.062 0.95080  
as.factor(escolaridade)2 -0.575469 0.270232 40.000000 -2.130 0.03941 *  
as.factor(escolaridade)3 -0.697742 0.782317 40.000000 -0.892 0.37779  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
              (Intr) as.fctr(t)2 as.fctr(s)2
as.fctr(t)2 -0.376
as.fctr(s)2 -0.474  0.000
as.fctr(s)3 -0.164  0.000   0.090

```

Chegamos ao modelo final:

$$y = 0,406230 - 0,575469 \cdot \text{Escolaridade2}$$

Observe que a variável preditora **time** não permaneceu no modelo após ajuste por escolaridade da mãe. Isto é evidência de que não ocorreu variação dos z-scores do IMC durante o período observado.

O nível de escolaridade 2 (9 a 11 anos) tem média 0,575469 unidades de z-score do IMC menor que os níveis 1 (< 8 anos de escolaridade) e 3 (12 ou mais anos). Não há diferença estatisticamente significante entre as médias de z-scores do IMC entre os níveis 1 e 3 de escolaridade.

Para obtermos os intervalos de confiança de 95% dos coeficientes estimados usamos o comando **confint()**<sup>17</sup>.

```
confint(m0, level = 0.95, parm = names(fixef(m0)))
```

```
> confint(m0, level = 0.95, parm = names(fixef(m0)))
Computing profile confidence intervals ...
              2.5 %    97.5 %
(Intercept)      0.1181968  0.69426385
as.factor(time)2 -0.2296471  0.21569360
as.factor(escolaridade)2 -1.0979271 -0.05301126
as.factor(escolaridade)3 -2.2102511  0.81476726
```

### Análise dos resíduos

Esta etapa é fundamental. Iniciamos calculando os valores previstos pelo modelo e os resíduos. Devemos avaliar se os resíduos têm distribuição normal, sendo que para isso podemos usar o teste de Shapiro-Wilk e o gráfico quantil-quantil (Q-Q).

```
res <- residuals(m0)
pred <- predict(m0)
shapiro.test(res)
```

No output abaixo observamos que os resíduos têm distribuição normal ( $p = 0,3504$ ) com média igual a zero. Portanto, este modelo é adequado.

```
> summary(res)
   Min. 1st Qu. Median Mean 3rd Qu. Max.
-1.42044 -0.25348 -0.02828 0.00000 0.24489 0.90690
> shapiro.test(res)

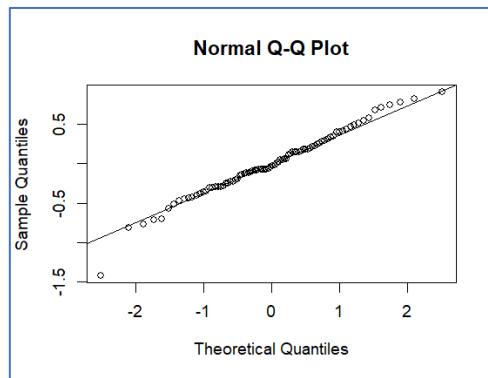
Shapiro-Wilk normality test

data: res
W = 0.98365, p-value = 0.3504
```

<sup>17</sup> confint(object, parm = NULL, level = 0.95, method = c("profile", "Wald", "boot"), nsim = 1000)

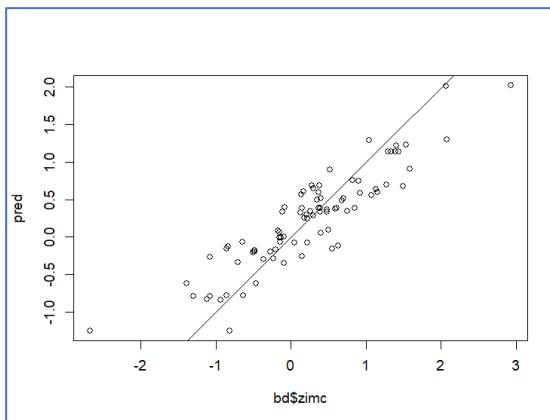
Avaliando a distribuição dos resíduos de forma gráfica.

```
qqnorm(res) ; qqline(res)
```



Por último, podemos plotar os valores previstos pelo modelo com os valores reais observados no banco de dados.

```
plot(bd$zimc,pred) ; abline(0,1)
```

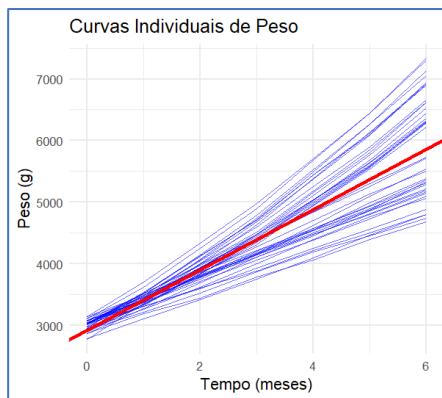


No exemplo apresentado a seguir, criado por simulação, vamos avaliar o ganho de peso de 40 crianças, supostamente expostas ao HIV durante a gestação, acompanhadas mensalmente até o sexto mês de vida. O nosso objetivo é modelar o ganho de peso destas crianças até o 6º mês de vida, ajustado por sexo e renda familiar. A variável desfecho é o peso (**id.peso**) e as preditoras **time** (mês de avaliação do peso), **sexo** (**id.sexo**) e renda familiar (**id.renda**). O banco de dados do tipo longo contém as seguintes variáveis: **id** (identificação da criança), **time** (0 – nascimento, 1 – 1º mês, 2 – 2º mês, ... 6 – 6º mês), **id.sexo** (0 – masculino e 1 – feminino) e **id.renda** (1 – muito baixa, 2 – baixa, 3 – média e 4 – renda alta). As preditoras **id.sexo** e **id.renda** são qualitativas.

Parte do banco de dados é apresentada na figura abaixo. O nome do banco de dados é “dados”.

	> dados	id	time	id.sex	id.renda	id.peso
1	1	0	0	0	4	3028.023
2	1	1	0	0	4	3517.370
3	1	2	0	0	4	3982.912
4	1	3	0	0	4	4409.634
5	1	4	0	0	4	4851.537
6	1	5	0	0	4	5306.579
7	1	6	0	0	4	5736.091
8	2	0	1	1	4	2901.090
9	2	1	1	1	4	3444.320
10	2	2	1	1	4	4076.557
11	2	3	1	1	4	4740.143
12	2	4	1	1	4	5493.931
13	2	5	1	1	4	6263.291
14	2	6	1	1	4	7056.516
15	3	0	1	1	4	3020.776
16	3	1	1	1	4	3529.940
17	3	2	1	1	4	4139.887
18	3	3	1	1	4	4784.472

No gráfico abaixo são apresentadas as curvas individuais de peso em azul. A reta de regressão em vermelho representa o ganho médio de peso do grupo de crianças, tendo como única preditora **time**<sup>18,19</sup>. Podemos observar que há heterogeneidade tanto no **intercepto** (nascimento) quanto no **slope** das curvas individuais.



Inicialmente vamos analisar quatro modelos, considerando como efeitos aleatórios somente o intercepto (modelo 1), somente o slope (modelo 2), o intercepto e o slope correlacionados (modelo 3) e o intercepto e o slope não correlacionados (modelo 4). Posteriormente, vamos comparar os quatro modelos e decidir qual é o efeito aleatório mais indicado para modelar os nossos dados longitudinais. Neste caso **time** foi introduzida no modelo como variável quantitativa.

---

<sup>18</sup> `ggplot(dados, aes(x = time, y = id.peso, group = id)) + geom_line(alpha = 0.6, color = "blue", size = 0.3) + geom_abline(intercept = coef(modelo_linear)[1], slope = coef(modelo_linear)[2], color = "red", size = 1.2) + labs(x = "Tempo (meses)", y = "Peso (g)", title = "Curvas Individuais de Peso") + theme_minimal()`

<sup>19</sup> O modelo linear geral é:  $\hat{y} = 2926,3 + 489,8 \cdot time$

```

modelo1 <- lmer(id.peso ~ time + as.factor(id.sex) + as.factor(id.renda) + (1 | id), data = dados)
modelo2 <- lmer(id.peso ~ time + as.factor(id.sex) + as.factor(id.renda) + (0 + time | id), data = dados)
modelo3 <- lmer(id.peso ~ time + as.factor(id.sex) + as.factor(id.renda) + (time | id), data = dados)
modelo4 <- lmer(id.peso ~ time + as.factor(id.sex) + as.factor(id.renda) + (time || id), data = dados)

```

Para comparar os resultados dos quatro modelos vamos usar inicialmente o critério **AIC()** e, estatisticamente, pelo comando **anova()**. Quanto menor o valor do AIC mais adequado é o modelo.

**AIC (modelo1, modelo2, modelo3, modelo4)**

```

> AIC(modelo1, modelo2, modelo3, modelo4)
      df      AIC
modelo1 8 3947.167
modelo2 8 3504.903
modelo3 10 3447.150
modelo4 9 3447.205

```

O modelo 4 apresenta o menor valor do critério **AIC**, portanto, parece ser o mais adequado para os nossos dados. Neste momento podemos comparar estatisticamente os quatro modelos usando o comando **anova()**.

**anova (modelo1, modelo2, modelo3, modelo4)**

```

> anova(modelo1, modelo2, modelo3, modelo4)
refitting model(s) with ML (instead of REML)
Data: dados
Models:
modelo1: id.peso ~ time + as.factor(id.sex) + as.factor(id.renda) + (1 | id)
modelo2: id.peso ~ time + as.factor(id.sex) + as.factor(id.renda) + (0 + time | id)
modelo4: id.peso ~ time + as.factor(id.sex) + as.factor(id.renda) + (time || id)
modelo3: id.peso ~ time + as.factor(id.sex) + as.factor(id.renda) + (time | id)
npar   AIC   BIC logLik deviance  Chisq Df Pr(>Chisq)
modelo1 8 3998.5 4027.6 -1991.3   3982.5
modelo2 8 3552.1 3581.2 -1768.1   3536.1 446.402 0
modelo4 9 3500.1 3532.8 -1741.0   3482.1 54.065 1  1.94e-13 ***
modelo3 10 3499.7 3536.1 -1739.8   3479.7 2.347 1  0.1255
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

Os modelos 1 e 2 tem igual performance. O modelo 4 é estatisticamente melhor que os modelos 1 e 2; não há diferença entre os modelos 3 e 4. Portanto, temos mais uma indicação de que o modelo 4 é o melhor para os nossos dados. Iniciaremos incluindo no modelo todas as preditoras, que serão avaliadas a cada etapa. Se o p-valor do coeficiente for  $\leq 0,05$  a preditora será mantida, caso contrário, será removida.

```
modelo4 <- lmer(id.peso ~ time + as.factor(id.sexo) + as.factor(id.renda) + (time || id), data = dados)
```

```
> summary(modelo4)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: id.peso ~ time + as.factor(id.sexo) + as.factor(id.renda) + (time || id)
Data: dados

REML criterion at convergence: 3429.2

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-2.13201 -0.38114 -0.04621  0.24232  2.17056 

Random effects:
 Groups   Name        Variance Std.Dev. 
id       (Intercept) 10502   102.48  
id.1     time        18682   136.68  
Residual            5175    71.94  
Number of obs: 280, groups: id, 40

Fixed effects:
            Estimate Std. Error    df t value Pr(>|t|)    
(Intercept) 2985.79    47.95  35.01 62.267 <2e-16 ***
time         484.29    21.72  39.00 22.299 <2e-16 ***
as.factor(id.sexo)1 -64.12    37.44  35.00 -1.713  0.0956 .  
as.factor(id.renda)2 14.13    53.63  35.00  0.263  0.7938  
as.factor(id.renda)3 -25.26    57.22  35.00 -0.441  0.6616  
as.factor(id.renda)4 -48.01    56.31  35.00 -0.853  0.3997  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
              (Intr) time  a.(.)1 a.(.)2 a.(.)3 
time          -0.013 
as.fctr(.)1 -0.446  0.000 
as.fctr(.)2 -0.805  0.000  0.199 
as.fctr(.)3 -0.676  0.000  0.010  0.602 
as.fctr(.)4 -0.733  0.000  0.114  0.632  0.573 
> 1
```

A preditora **id.renda** tem o maior p-valor, portanto, será removida do modelo.

Executamos novamente o modelo 4.

```
modelo4 <- lmer(id.peso ~ time + as.factor(id.sexo) + (time || id), data = dados)
```

```
> summary(modelo4)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: id.peso ~ time + as.factor(id.sexo) + (time || id)
Data: dados

REML criterion at convergence: 3459.7

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-2.11761 -0.38286 -0.03505  0.22479  2.25705 

Random effects:
 Groups   Name        Variance Std.Dev. 
id       (Intercept) 10136   100.68  
id.1     time        18570   136.27  
Residual            5181    71.98  
Number of obs: 280, groups: id, 40

Fixed effects:
            Estimate Std. Error    df t value Pr(>|t|)    
(Intercept) 2975.00    23.34  38.06 127.438 <2e-16 ***
time         484.29    21.65  39.02 22.365 <2e-16 ***
as.factor(id.sexo)1 -68.69    35.80  38.00 -1.919  0.0625 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
              (Intr) time 
time          -0.027 
as.fctr(.)1 -0.652  0.000 
1
```

No output acima podemos observar que o coeficiente da preditora de efeito fixo **id.sexo** não é estatisticamente diferente de zero. Portanto, será removida e executamos novamente o modelo 4.

```
modelo4 <- lmer(id.peso ~ time + (time || id), data = dados)
```

```

> summary(modelo4)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['tmerModLmerTest']
Formula: id.peso ~ time + (time || id)
Data: dados

REML criterion at convergence: 3472.3

Scaled residuals:
    Min     1Q   Median     3Q    Max 
-2.14593 -0.37397 -0.03556  0.24693  2.21420 

Random effects:
 Groups   Name        Variance Std.Dev. 
id       (Intercept) 10981   104.79  
id.time      time      18223   134.99  
Residual           5200    72.11  
Number of obs: 280, groups: id, 40

Fixed effects:
            Estimate Std. Error   df t value Pr(>|t|)    
(Intercept) 2945.80    18.30 39.07 160.97 <2e-16 ***
time         484.29    21.45 39.07 22.57 <2e-16 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
  (Intr) 
time -0.035 
>

```

O modelo final é:

$$\hat{y} = 2945,8 + 484,29 \cdot time$$

Portanto, este modelo nos indica que o ganho médio mensal de peso observado no grupo das 40 crianças foi de 484,29g.

Os intervalos de confiança de 95% dos coeficientes de regressão podem ser obtidos com o comando **confint(modelo4)**.

```
confint(modelo4)
```

```

> confint(modelo4)
Computing profile confidence intervals
          2.5 %    97.5 %
.sig01    78.89013 136.1419
.sig02    108.36019 169.0654
.sigma    65.56060 79.8559
(Intercept) 2909.53175 2982.0737
time       441.77203 526.8110
>

```

Vamos avaliar estatisticamente os efeitos fixos e aleatórios do modelo4 pelos comandos **anova()** e **ranova()**.

```
anova(modelo4)
```

```
ranova(modelo4)
```

```

> anova(modelo4)
Type III Analysis of Variance Table with Satterthwaite's method
  Sum Sq Mean Sq NumDF DenDF F value    Pr(>F)    
time 2650011 2650011     1 39.07  509.62 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

```

> ranova(modelo4)
ANOVA-like table for random-effects: Single term deletions

Model:
id.peso ~ time + (1 | id) + (0 + time | id)
                    npar logLik   AIC    LRT Df Pr(>Chisq)
<none>                  5 -1736.1 3482.3
(1 | id)                 4 -1768.5 3545.0 64.73  1  8.568e-16 ***
time in (0 + time | id)  4 -2034.6 4077.2 596.97  1 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
>

```

Podemos comprovar nos outputs acima que os efeitos fixos e os efeitos aleatórios foram adequados para analisar os dados.

Até este ponto tudo parece muito bom, entretanto, temos que analisar o comportamento dos resíduos, etapa fundamental, visto que estes obrigatoriamente devem ter distribuição normal com média igual a zero. Este é um dos mais importantes pressupostos dos modelos lineares. Para tal, calculamos os resíduos usando o comando `residuals()` e testamos o seu ajuste à distribuição normal com o teste de Shapiro-Wilk.

```
summary(residuals(modelo4))
shapiro.test(residuals(modelo4))
```

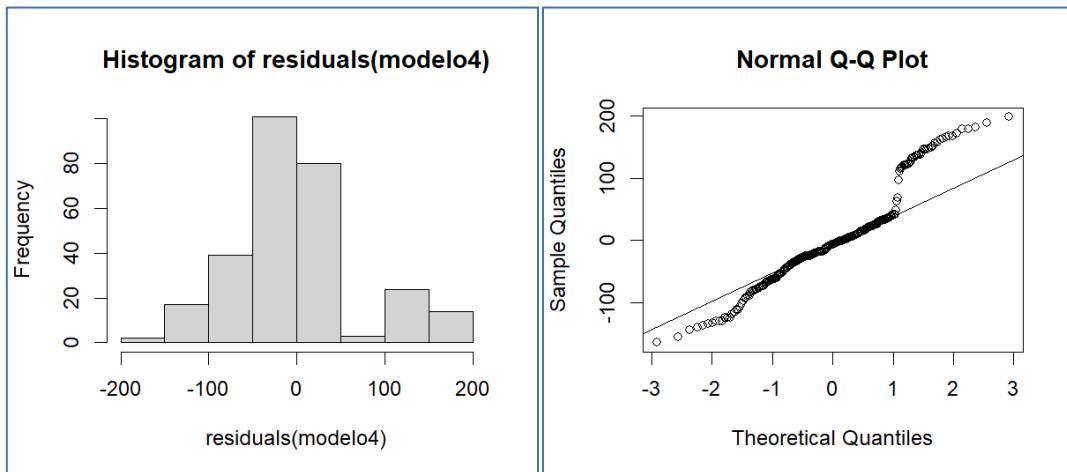
```
R 4.3.3 · ~/🔗
> summary(residuals(modelo4))
   Min. 1st Qu. Median Mean 3rd Qu. Max.
-154.744 -26.967 -2.564 0.000 17.806 159.667
> shapiro.test(residuals(modelo4))

Shapiro-Wilk normality test

data: residuals(modelo4)
W = 0.9186, p-value = 3.241e-11
```

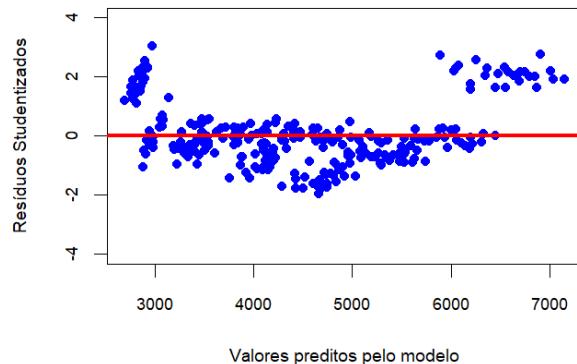
Veja que os resíduos têm média zero, porém. não se ajustam à distribuição normal, pois o p-valor do teste de Shapiro-Wilk é menor que 0,05. Isto também pode ser observado nos gráficos abaixo (histograma e qqnormal).

```
hist(residuals(modelo4))
qqnorm(residuals(modelo4))
qqline(residuals(modelo4))
```



Vamos avaliar a distribuição dos resíduos studentizados em relação aos valores preditos pelo modelo. Espera-se que os resíduos tenham uma distribuição aleatória e homogênea em relação à linha de referência no ponto zero. Observe no gráfico que a distribuição dos resíduos não é aleatória.

```
res1 <- rstudent(modelo4)
pred <- predict(modelo4)
plot(pred, res1, ylim = c(-4,4), ylab = "Resíduos Studentizados", xlab = "Valores preditos pelo modelo",
     pch=21, col="blue", bg="blue", cex=1, cex.lab= 0.8, cex.axis = 0.75)
abline(h = 0, v = 0, col = "red", lwd=3)
```



Considerando que os resíduos **não** se ajustam à distribuição normal, o modelo 4 não deve ser considerado adequado, visto que não temos confiança na precisão dos coeficientes de regressão calculados e seus intervalos de confiança.

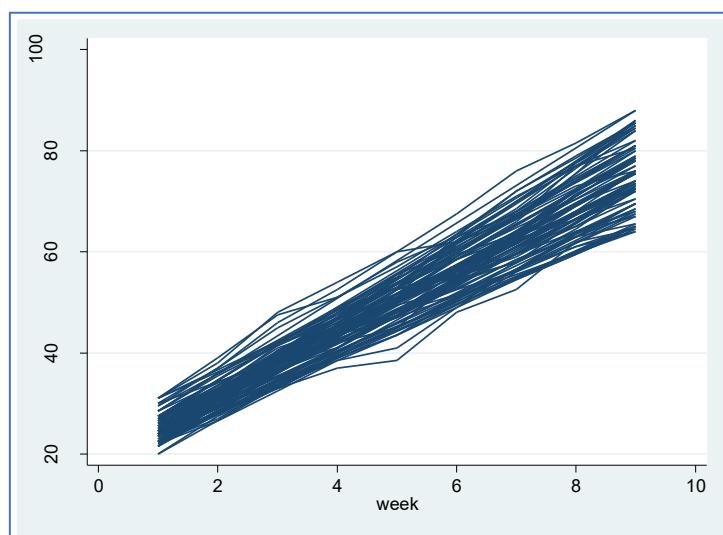
O que podemos fazer? Inicialmente reavaliar a variável desfecho (**id.peso**), procurando identificar outliers. Se eles não podem ser eliminados, podemos procurar alguma transformação (log ou  $\sqrt{}$ ) que resolva o problema dos outliers. Caso todas as tentativas sejam insatisfatórias, não devemos forçar o uso dos Modelos Lineares de Efeitos Mistos. Devemos procurar outra técnica de análise. Uma boa alternativa, que supera o problema dos outliers e consequentemente dos resíduos, é a Regressão Quantílica com Efeitos Mistos, que não tem os pressupostos dos modelos lineares e é muito eficiente para analisar dados longitudinais balanceados ou não.

Neste próximo exemplo, será analisado, um banco de dados obtido a partir do manual do Stata, disponível em <http://www.stata-press.com/data/r15/pig.dta>. Não se trata de uma pesquisa na área da Pediatria, porém, por sua simplicidade, torna-se um excelente exemplo.

Este banco de dados ([pig](#)) contém informações de peso de 48 porcos que foram acompanhados ao longo de 9 semanas, de tal forma que cada animal tinha 9 avaliações do peso. Como todos os animais têm o mesmo número de avaliações, trata-se de um banco de dados balanceado, com  $n = 48$  animais,  $t = 9$  avaliações semanais e  $N = 432$  medidas de peso. O objetivo da pesquisa que gerou este banco de dados foi modelar o ganho de peso semanal de porcos. Este é um banco de dados do tipo longo.

id	week	weight
1	1	24.0
2	1	32.0
3	1	39.0
4	1	42.5
5	1	48.0
6	1	54.5
7	1	61.0
8	1	65.0
9	1	72.0
10	2	22.5
11	2	30.5
12	2	40.5
13	2	45.0
14	2	51.0

A curva de ganho de peso de cada animal é apresentada abaixo.



Nesta pesquisa, **weight** é a variável desfecho (quantitativa) e **week** a preditora que expressa o momento em que o peso do animal foi aferido. Por último, **id** é a variável que identifica os animais.

Desenvolveremos quatro modelos considerando o efeito aleatório somente no intercepto, somente no slope, no intercepto e slope aleatório correlacionados e no intercepto e slope aleatório não correlacionados.

```
m1 <- lmer(weight ~ week + (1 | id), data = pig)
m2 <- lmer(weight ~ week + (0 + week | id), data = pig)
m3 <- lmer(weight ~ week + (week | id), data = pig)
m4 <- lmer(weight ~ week + (week || id), data = pig)
```

Para avaliar estes modelos empregamos o critério **AIC**.

```
AIC(m1, m2, m3, m4)
```

```
> AIC(m1, m2, m3, m4)
   df      AIC
m1  4 2041.797
m2  4 1925.919
m3  6 1752.871
m4  5 1751.029
> |
```

O modelo **m4**, com efeito aleatório no intercepto e no slope não correlacionados é o mais indicado para analisar estes dados, visto que tem o menor valor de AIC. Isto também é demonstrado pelo comando **anova()**.

```
anova(m1, m2, m3, m4)
```

```
> anova(m1, m2, m3, m4)
refitting model(s) with ML (instead of REML)
Data: pig
Models:
m1: weight ~ week + (1 | id)
m2: weight ~ week + (0 + week | id)
m4: weight ~ week + (week || id)
m3: weight ~ week + (week | id)
  npar    AIC      BIC  logLik deviance   Chisq Df Pr(>Chisq)
m1:  4 2037.8 2054.1 -1014.93    2029.8
m2:  4 1921.7 1938.0  -956.87    1913.7 116.1228  0      <2e-16 ***
m4:  5 1748.1 1768.4  -869.04    1738.1 175.6543  1      <2e-16 ***
m3:  6 1749.9 1774.3  -868.96    1737.9  0.1528  1      0.6959
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |
```

Desta forma, usaremos o modelo 4 para avaliar o ganho médio mensal de peso.

```
m4 <- lmer(weight ~ week + (week || id), data = pig)
```

```

> summary(m4)
Linear mixed model fit by REML. t-tests use Satterthwaite's method ['lmerModLmerTest']
Formula: weight ~ week + (week || id)
Data: pig

REML criterion at convergence: 1741

Scaled residuals:
    Min      1Q  Median      3Q     Max 
-3.6135 -0.5408  0.0195  0.5464  3.0117 

Random effects:
Groups   Name        Variance Std.Dev. 
id       (Intercept) 6.9176   2.6301  
id.week  week        0.3764   0.6135  
Residual            1.5988   1.2644  
Number of obs: 432, groups: id, 48

Fixed effects:
            Estimate Std. Error    df t value Pr(>|t|)    
(Intercept) 19.35561  0.40211 47.53764 48.13 <2e-16 ***
week         6.20990  0.09164 47.53736 67.77 <2e-16 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
          (Intr) 
week     -0.075

```

Tanto o intercepto quanto a preditora **week** tem p-valor menor que 0,05.

O modelo final é:

$$y = 19,35561 + 6,2099 \cdot \text{Week}$$

Isto significa que o ganho médio semanal de peso no período avaliado é de 6,2099 kg. Neste ponto, podemos calcular os intervalos de confiança de 95% dos coeficientes.

`confint(m4)`

```

> confint(m4)
Computing profile confidence intervals ...
              2.5 %    97.5 %
.sig01     2.0950626 3.286507
.sig02     0.4952488 0.760316
.sigma     1.1745051 1.366445
(Intercept) 18.5600258 20.151201
week       6.0285878 6.391204

```

## Análise dos resíduos

```

res <- residuals(m4)
summary(res)
shapiro.test(res)

```

```

> res <- residuals(m4)
> summary(res)
      Min. 1st Qu. Median Mean 3rd Qu. Max. 
-4.56907 -0.68380 0.02465 0.00000 0.69086 3.80803 
> shapiro.test(res)

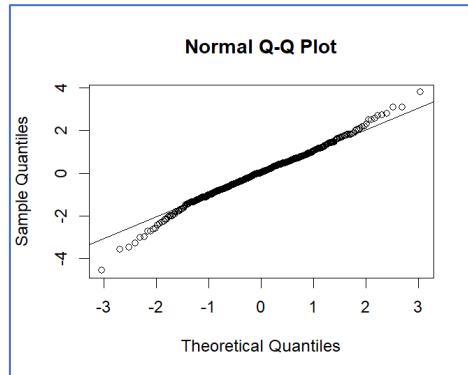
Shapiro-Wilk normality test

data: res
W = 0.9905, p-value = 0.006955

```

No output acima observamos que os resíduos têm média zero, porém não se ajustam à distribuição normal. Isso também pode ser observado no gráfico **qqnorm()**.

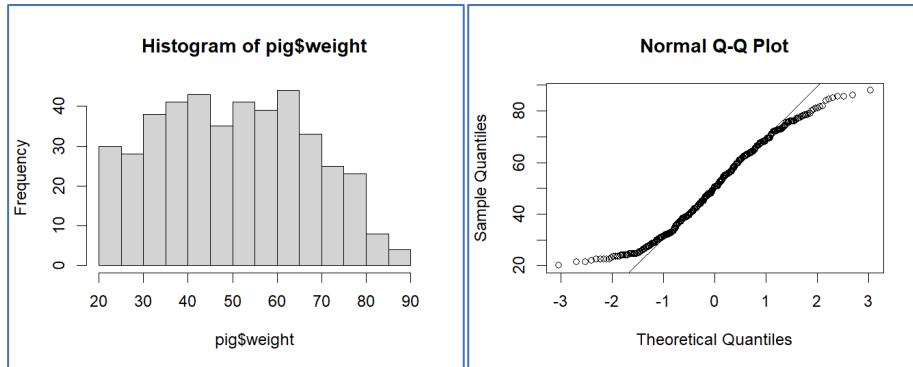
```
qqnorm(res) ; qqline(res)
```



Esta importante etapa da análise mostrou que os resíduos do modelo não têm distribuição normal. Portanto, este modelo não deve ser utilizado, visto que a estimativas dos coeficientes de regressão e seus intervalos de confiança não são confiáveis.

A análise da variável desfecho **weight** nos mostra que esta não se ajusta à distribuição normal, o que é evidente pelo teste de Shapiro-Wilk que tem  $p < 0,001$  e pelos gráficos histograma e **qqnorm()**. Neste caso a Regressão Quantílica com Efeitos Misto seria uma opção de análise muito melhor.

```
> shapiro.test(pig$weight)
Shapiro-Wilk normality test
data: pig$weight
W = 0.97236, p-value = 2.714e-07
```



## OBSERVAÇÃO

A concepção, elaboração do texto e execução dos exemplos é responsabilidade de AMM.

O segundo exemplo, baseado em uma simulação foi criado a partir de um script em R desenvolvido por AMM.

As IAs (Claude, ChatGPT e Manus) foram utilizadas somente na revisão gramatical e de estilo.

## BIBLIOGRAFIA

1. Amaral SSWG. Modelos lineares mistos para análise de dados longitudinais bivariados provenientes de ensaios agropecuários [tese]. Piracicaba: Escola Superior de Agricultura “Luiz de Queiroz”; 2013. 78 p.
2. Bates D, Mächler M, Bolker B, Walker S. “Fitting Linear Mixed-Effects Models Using lme4.” *Journal of Statistical Software*, 2015; 67(1), 1–48. doi:10.18637/jss.v067.i01.
3. Bates D, Maechler M, Bolker B, Walker S, Christensen RHB, Singmann H, Dai B, Scheipl F, Grothendieck G, Green P. Package ‘lme4’. Repository CRAN, 2017. Disponível em: <http://cran.r-project.org/web/packages/lme4/lme4.pdf>.
4. Carniel EF, Morcillo AM, Blotta HM, Silva MTN, Mazzola TN, Antonio MAGM, Zanolli ML, Netto AA, Higashi HG, Raw I, Vilela MMS. Immunogenicity and safety of combined intradermal recombinant Hepatitis B with BCG vaccines at birth. *Vaccine* (Guildford), 2008; 26:647-652.
5. Fausto MA, Carneiro M, Antunes CMF, Pinto JA, Colosimo EA. O modelo de regressão linear misto para dados longitudinais: uma aplicação na análise de dados antropométricos desbalanceados. *Cad. Saúde Pública*, 2008; 24(3):513-524.
6. Garcia TP, Marder K. Statistical Approaches to Longitudinal Data Analysis in Neurodegenerative Diseases: Huntington’s Disease as a Model. *Curr Neurol Neurosci Rep*. 2017; 17(2):14. doi: 10.1007/s11910-017-0723-4
7. Hair Jr JF, Fávero LP. Multilevel modeling for longitudinal data: concepts and applications. *RAUSP Management Journal*, 2018; 54(4):459-489.
8. Hickey GL, Mokhlesb MM, Chambersc DJ, Kolamunnage-Donaa R. Statistical primer: performing repeated-measures analysis. *Interactive CardioVascular and Thoracic Surgery*, 2018; 26(4):539–544.

9. Lundqvist S, Börjesson M, Cider A, Hagberg L, Ottehall CB, Sjöström J, Larsson MEH. Long-term physical activity on prescription intervention for patients with insufficient physical activity level—a randomized controlled trial. *Trials*. 2020; 21:793. doi: 10.1186/s13063-020-04727-y
10. Ma Y, Mazumdar M, Memtsoudis SG. Beyond Repeated measures ANOVA: advanced statistical methods for the analysis of longitudinal data in anesthesia research. *Reg Anesth Pain Med*. 2012; 37(1):99–105. doi: 10.1097/AAP.0b013e31823ebc74
11. Muhammad LN. Guidelines for repeated measures statistical analysis approaches with basic science research considerations. *J Clin Invest*. 2023; 133(11):e171058. doi: 10.1172/JCI171058
12. Oliveira LP, Lima CG, Castro VLSS, Siqueira MC, Maia AHN. Modelo linear misto para avaliar toxicidade por doses repetidas. *Revista da Estatística UFOP*, 2014; III(3):609 – 613.
13. Paulino MFVM, Lemos-Marini SHV, Guerra-Junior G, Morcillo AM. Crescimento e composição corporal de uma coorte de crianças e adolescentes com Diabetes tipo I. *Arquivos Brasileiros de Endocrinologia e Metabologia*, 2013; 57:623-631.
14. Pinheiro J, Bates D, DebRoy S, Sarkar D, Heisterkamp S, Van Willigen B, Ranke J. Package ‘nlme’. Repository CRAN, 2025. Disponível em: <https://cran.r-project.org/web/packages/nlme/nlme.pdf>.
15. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, 2024. Disponível em: <<https://www.R-project.org/>>.
16. Ramasamy N, Raj R, Patel N, C Palanivel. Longitudinal data analysis methods – a primer. *EFI Bulletin*. 2023; 4(1):12-17. DOI: 10.56450/EFIB.2023.v3i01.003
17. Repeated Measures Analysis with Stata. UCLA: Statistical Consulting Group. Disponível em: <https://stats.oarc.ucla.edu/stata/seminars/repeated-measures-analysis-with-stata/>
18. Santos CIS, Ribeiro MÂGO, Ribeiro JD, Morcillo AM. Respiratory physiotherapy in children with community-acquired pneumonia.. *Canadian Journal of Respiratory Therapy*, 2009; 45:23-28.
19. Schober P, Vetter YR. Repeated Measures Designs and Analysis of Longitudinal Data: If at First You Do Not Succeed—Try, Try Again. *Anesth Analg*. 2018 Jun 12;127(2):569–575. doi: 10.1213/ANE.0000000000003511
20. Spyrides MHC, Struchiner CJ, Barbora MTS, Kac G. Amamentação e crescimento infantil: um estudo longitudinal em crianças do Rio de Janeiro, Brasil, 1999/2001. *Cad. Saúde Pública*. 2005; 21(3):756-766.
21. StataCorp. Stata Longitudinal Data/Panel-Data Reference Manual. Release 12. College Station, TX: Stata Press, 2011.
22. Wang X, Andrinopoulou E-R, Veen KM, Bogers AJJC, Takkenberg JJM. Statistical primer: an introduction to the application of linear mixed-effects models in cardiothoracic surgery

outcomes research—a case study using homograft pulmonary valve replacement data. Eur J Cardiothorac Surg 2022; doi:10.1093/ejcts/ezac429.

23. Winter B. Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499, 2013. Disponível em: <https://arxiv.org/pdf/1308.5499>.