

Capítulo

4

Estudo Longitudinal em Pesquisas em Informática na Educação

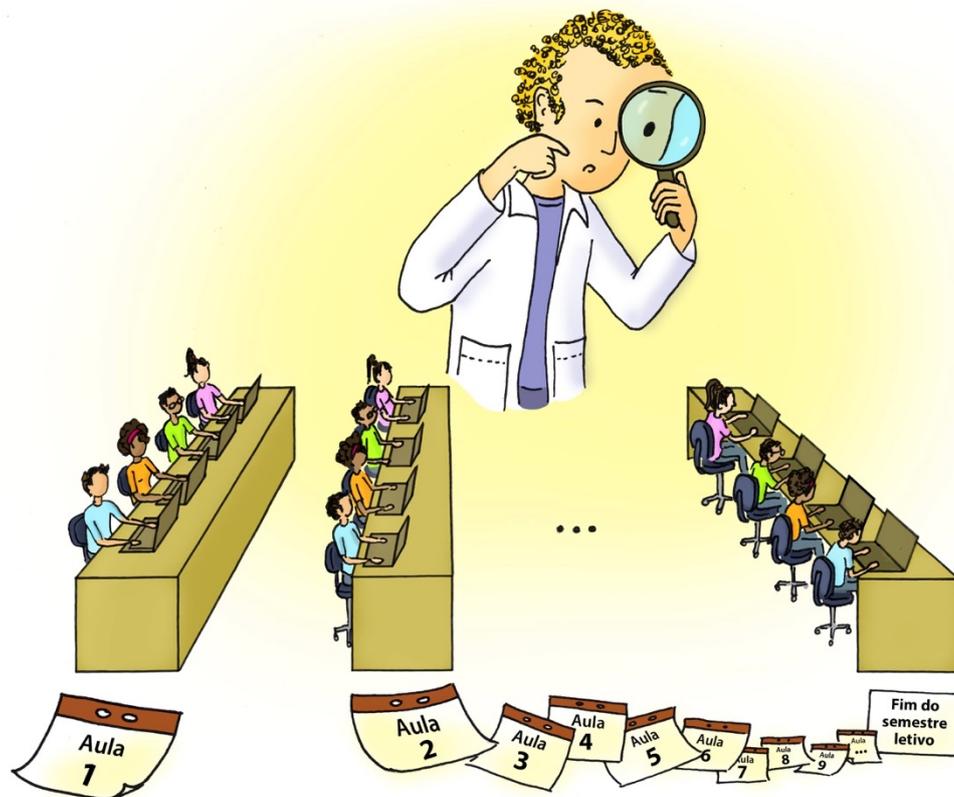
Aline C. R. de Souza (DES-UEM)

Willian L. de Oliveira (DES-UEM)

Objetivo do Capítulo

Este capítulo tem o objetivo de apresentar os estudos longitudinais como método de pesquisa, mostrando suas principais características e aplicações na área de informática na educação. Ao final da leitura deste capítulo, você deve ser capaz de:

- Identificar problemas envolvendo estudos longitudinais.
- Aplicar as técnicas apropriadas para resolução dos problemas.



Era uma vez... Pedro, um aluno da pós-graduação de uma renomada universidade que, com o objetivo de aumentar o rendimento e a participação de alunos dos cursos de graduação, desenvolveu um conjunto de metodologias pedagógicas destinadas ao ensino de disciplinas básicas que serão cursadas em sua maioria por alunos ingressantes. Em seu trabalho, Pedro propôs o uso da metodologia “sala de aula invertida”, que permite uma maior interação entre os alunos e os conteúdos a serem estudados. Entre as atividades sugeridas por Pedro, está a elaboração e a aplicação de testes disponibilizados em alguma plataforma digital, que devem ser realizados antes de cada aula. Esta estratégia possibilita que o aluno se familiarize com o conteúdo antes deste ser abordado pelo professor, o que pode facilitar o processo de aprendizagem. Desta forma, ao final do estudo, cada aluno terá um conjunto de notas que foram coletadas ao longo do tempo, caracterizando um conjunto de dados longitudinais. Para verificar a eficácia da metodologia proposta em disciplinas de base, e conseqüentemente a melhoria da aprendizagem em outras disciplinas do curso de graduação, Pedro deverá utilizar técnicas de estudos longitudinais, que serão abordadas neste capítulo.

1 Introdução

Em diversas áreas do conhecimento, o levantamento de dúvidas relacionadas a um determinado tema, bem como a busca por soluções ou respostas que possam colaborar em seu desenvolvimento, são elementos chaves que conduzem uma pesquisa. Juntamente a estes fatores, é necessário eleger um conjunto de métodos de investigação que mais se adequem ao objetivo final do estudo, e que auxiliem o pesquisador na compreensão do problema abordado. Uma importante etapa deste processo refere-se à observação de um experimento, ou fenômeno de interesse, relacionados ao objeto de estudo. Isto porque através desta prática, geram-se dados que, se obtidos de maneira estratégica, poderão fornecer informações de grande relevância ao pesquisador.

Consideremos, por exemplo, a pesquisa desenvolvida por Pedro. Para investigar a aprendizagem através da aplicação dos novos métodos propostos, dois grupos de alunos escolhidos por Pedro de forma aleatória, são selecionados para que seus desempenhos sejam observados. Em uma primeira sugestão de estudo deste caso, a aprendizagem do grupo de alunos submetidos à nova metodologia de ensino pode ser confrontada com relação a aprendizagem do outro grupo, que foi submetido às práticas de ensino convencionais. Desta forma, Pedro poderá verificar a eficácia dos novos métodos propostos. Para tanto, será necessário coletar, em um determinado instante de tempo, os resultados que reflitam os desempenhos dos alunos dos dois grupos analisados, como por exemplo, as notas de um teste aplicado, a média de um conjunto de testes realizados em algum período, entre outros.

Esta abordagem, em que a variável de interesse é observada uma vez para a coleta, é conhecida como **estudo transversal** (ver Capítulo 2 deste livro para mais detalhes sobre Estudos Experimentais). Neste tipo de estudo, as informações sobre a variável de interesse são medidas simultaneamente em um determinado instante de tempo T para todos os indivíduos envolvidos. Assim, uma característica intrínseca aos estudos transversais é que o processo de observação fornece uma “fotografia” de como as variáveis estão relacionados no momento da coleta.

Suponha que a variável observada por Pedro seja a pontuação média das avaliações bimestrais (obtida ao final do curso) de dois grupos de 10 alunos, que foram divididos aleatoriamente da seguinte maneira:

Grupo 1: Alunos que cursaram a disciplina aplicada com as novas metodologias propostas no trabalho de Pedro.

Grupo 2: Alunos que cursaram a disciplina aplicada com as metodologias tradicionais.

Considere ainda que as observações coletadas foram organizadas para estudo. Sob uma perspectiva mais simplória, podemos ter uma ideia se o desempenho dos alunos do grupo A foi maior do que o do grupo B, através de uma análise descritiva, calculando algumas medidas resumo, tais como 1º quartil, mediana, 3º quartil, média, mínimo e máximo, etc. Note, porém, que estas quantidades dizem respeito a um instante particular de tempo: o semestre no qual foi ministrado o curso. Qualquer informação fora deste período é desconsiderada. Na situação hipotética em que Pedro deseja verificar a relação entre a aprendizagem dos alunos em disciplinas de base, e o desempenho ao longo do curso, este tipo de dados não será de muita utilidade.

Imaginemos agora que em sua pesquisa Pedro queira justificar a importância da implementação de novas metodologias de ensino, investigando o impacto da aprendizagem de uma disciplina de base ao longo de todo curso. Ou seja, as seguintes perguntas devem ser respondidas: Uma falha na aprendizagem de disciplinas básicas afeta o desempenho de um aluno em outras disciplinas? Para que novas medidas possam ser tomadas, de que forma isto acontece ao longo do curso de graduação?

Observe que a estratégia abordada anteriormente para a coleta de dados não é apropriada para se conseguir responder a estas perguntas. O acompanhamento dos alunos por vários períodos pode ser fundamental para que se conheça gradativamente o seu grau de aprendizagem. Com este propósito, o processo de observação do desempenho deverá ser feito ao longo de diferentes instantes de tempo t_1, t_2, \dots, t_n , através de um tipo de estudo denominado **estudo longitudinal**. De maneira ilustrativa, é como se, ao longo do processo de observação, acompanhássemos a trajetória das variáveis de interesse obtendo um perfil de comportamento para cada indivíduo observado.

Estudo longitudinal é sempre recomendável?

Apesar de vantajoso para a análise da evolução temporal de variáveis de interesse, em muitos problemas o estudo longitudinal é bastante custoso e de difícil implementação, uma vez que o acompanhamento dos indivíduos pode ser demasiadamente longo, sendo necessário anos de observação.

2 Análise descritiva para dados longitudinais

As técnicas usadas na análise descritiva de dados obtidos através de um estudo transversal podem ser facilmente aplicadas aos dados longitudinais (ver Capítulo 9 deste livro para mais detalhes sobre Estatística Descritiva). Consideremos como motivação o estudo longitudinal aplicado ao problema de Pedro, apresentado na Seção Era uma vez... desse capítulo. Suponha que para atender às necessidades de sua pesquisa, 10 alunos escolhidos de maneira aleatória tenham sido acompanhados ao longo dos quatro semestres subsequentes ao semestre em que a disciplina básica analisada foi ofertada, através da observação das médias gerais de cada aluno. Tais pontuações, que foram obtidas calculando-se a média aritmética das notas finais das disciplinas que foram cursadas no semestre no qual realizou-se a coleta, encontram-se apresentadas na Tabela 1.

Grupo	Aluno	Semestre	Nota	Grupo	Aluno	Semestre	Nota
1	1	1	7,20	2	6	1	6,20
1	1	2	7,41	2	6	2	6,21
1	1	3	7,59	2	6	3	6,16
1	1	4	7,90	2	6	4	6,06
1	2	1	7,68	2	7	1	6,11
1	2	2	8,11	2	7	2	6,26
1	2	3	8,46	2	7	3	6,43
1	2	4	8,83	2	7	4	6,72
1	3	1	6,66	2	8	1	5,41
1	3	2	7,15	2	8	2	5,81
1	3	3	7,61	2	8	3	6,19
1	3	4	8,03	2	8	4	6,59
1	4	1	6,41	2	9	1	5,20
1	4	2	6,86	2	9	2	5,07
1	4	3	7,28	2	9	3	4,96
1	4	4	7,73	2	9	4	4,94
1	5	1	7,21	2	10	1	4,91
1	5	2	7,64	2	10	2	5,18
1	5	3	8,12	2	10	3	5,39
1	5	4	8,64	2	10	4	5,51

Tabela 1 - Dados longitudinais coletados por Pedro.

Para fazer a análise descritiva dos dados abordados, utilizaremos a linguagem R (R CORE TEAM, 2016). Os dados descritos podem ser organizados em um data frame no R. A função `head` permite observar as primeiras linhas deste conjunto e assim verificar a forma em que estão dispostos.

```
> head(dadosNotas)
grupo aluno semestre nota
1      1      1      1 7.20
2      1      1      2 7.41
3      1      1      3 7.59
4      1      1      4 7.90
5      1      2      1 7.68
6      1      2      2 8.11
```

De maneira geral, podemos obter a média e a variância de uma resposta longitudinal relacionada a um indivíduo. Para tanto, basta considerar as medidas longitudinais correspondentes ao mesmo. Suponha, por exemplo, que desejamos verificar tais medidas relacionadas ao aluno 5. Este procedimento é realizado através das seguintes linhas de comando:

```

> #Média Notas indivíduo 5 nos 4 semestres
> mean(dadosNotas$nota [dadosNotas$aluno==5])
[1] 7.9025

> #Variância Notas indivíduo 5 nos 4 semestres
> var(dadosNotas$nota[dadosNotas$aluno==5])
[1] 0.3798917

```

Considere agora que o interesse consiste em obter a média e a variância das notas de todos os alunos envolvidos no estudo no primeiro semestre de coleta. Isto pode ser realizado em R da seguinte forma:

```

> #Média Notas alunos no semestre 1
> mean(dadosNotas$nota[dadosNotas$semestre==1])
[1] 6.299
> #Variância Notas alunos no semestre 1
> var(dadosNotas$nota[dadosNotas$semestre==1])
[1] 0.85161

```

Na pesquisa de Pedro, os alunos foram divididos em dois grupos. O grupo A é composto por estudantes que cursaram a disciplina proposta e o grupo B é composto por alunos que não a cursaram. Podemos verificar o valor médio e comparar as notas dos indivíduos pertencentes a cada um dos grupos de interesse.

```

> #Média e variancia das Notas do grupo 1 nos 4 semestres
> mean(dadosNotas$nota[dadosNotas$grupo==1])
[1] 7.626
> var(dadosNotas$nota[dadosNotas$grupo==1])
[1] 0.4027095
> #Média e variancia das Notas do grupo 2 nos 4 semestres
> mean(dadosNotas$nota[dadosNotas$grupo==2])
[1] 5.7655
> var(dadosNotas$nota[dadosNotas$grupo==2])
[1] 0.3551418

```

Observe que no primeiro semestre após a oferta da disciplina proposta por Pedro, a média do grupo A foi de 7,63 ao passo que a do grupo B foi 5,77. Uma pergunta de interesse seria: Qual o comportamento da média de cada grupo nos quatro semestres do estudo? As tendências de comportamento destas quantidades ao longo do tempo podem ser detectadas através dos **gráficos de perfil**. Neles, marcam-se os valores da variável observada para cada indivíduo, em cada instante de tempo de acompanhamento. O gráfico de perfil para as notas dos alunos encontra-se exposto na Figura 1.

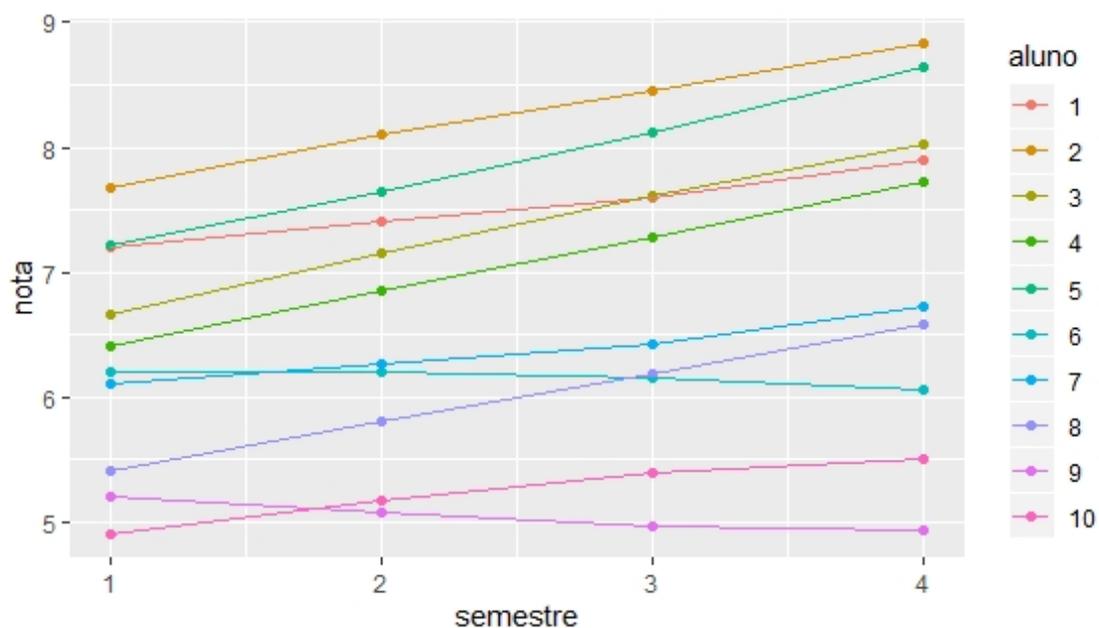


Figura 1: Trajetórias longitudinais das notas dos alunos.

O gráfico de perfil para as notas dos alunos foi gerado em R através do pacote gráfico *ggplot2*, utilizando as seguintes linhas de comando:

```
> require(ggplot2)
> p <- ggplot(data = dadosNotas, aes(x = semestre, y = nota, group =
aluno, color=aluno))
> p + geom_point(aes(colour = aluno)) + geom_line()
```

Nota-se que, em geral, as notas dos alunos tendem a aumentar ao longo dos semestres, seguindo um comportamento linear. Em alguns casos, no entanto, observa-se uma queda de rendimento nos semestres subsequentes ao início do acompanhamento dos alunos. Separando-se os gráficos de perfil pelos grupos 1 e 2, temos o seguinte resultado gráfico, apresentados na Figura 2.

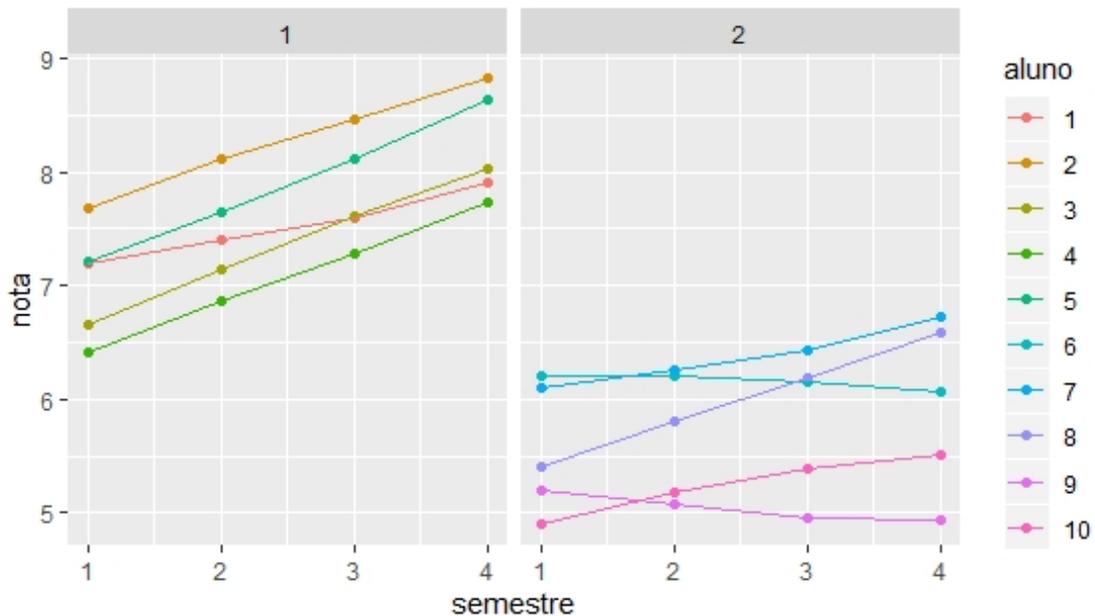


Figura 2: Trajetórias longitudinais das notas dos alunos divididas por grupos.

Para a geração desse gráfico em R foi utilizada a seguinte linha de comando:

```
> p + geom_point(aes(colour = aluno)) + geom_line() + facet_wrap(~grupo)
```

Observamos que os alunos do grupo 1 apresentaram rendimentos crescentes ao longo dos semestres, com notas maiores quando comparadas as notas dos alunos do grupo 2. Além disso, nota-se uma evolução semelhante nas notas dos alunos, com uma tendência linear. No grupo 2, no entanto, verificamos maiores diferenças entre os alunos, com indivíduos que apresentam queda de rendimento acadêmico no período de estudo.

3 Modelagem para dados longitudinais através dos modelos lineares de efeitos mistos

Para uma análise mais completa dos dados longitudinais, uma abordagem usando modelos estatísticos é aconselhável. Nesta, o comportamento médio da resposta analisada pode ser explicado através de um modelo de regressão linear, ou não linear, a depender da relação entre as variáveis. Tal componente descreve um efeito comum a todos os indivíduos, o qual é denominado de **efeito fixo**.

Como observamos no exemplo anterior, o comportamento individual de uma variável, como o rendimento acadêmico, pode ser diferente para cada aluno. Este aspecto do problema é modelado considerando componentes aleatórias que descrevem a variabilidade de cada indivíduo, que são denominadas de **efeitos aleatórios**.

Um modelo linear que considera estes dois tipos de efeitos é conhecido por

modelo linear de efeitos mistos. Esta classe de modelos é de grande utilidade na análise de dados longitudinais, tratando-se de uma extensão dos modelos de regressão linear abordados no Capítulo 11 deste livro.

3.1 Formulação do modelo

Na especificação de um modelo linear com efeitos mistos seja \mathbf{y}_i um vetor aleatório cujos elementos são as respostas $y_{i1}, y_{i2}, \dots, y_{in}$, correspondentes a um indivíduo i . Este vetor está associado às medidas de interesse observadas ao longo do tempo. Considere, por exemplo, que se deseja verificar a evolução temporal da altura de uma amostra aleatória de crianças, na faixa etária de 5 a 10 anos. Para tanto, as alturas dos indivíduos envolvidos no estudo são medidas nos seguintes instantes de acompanhamento, dados em meses após o início do estudo: $t_1 = 0$, $t_2 = 12$, $t_3 = 24$, $t_4 = 36$. Assim, um possível vetor de respostas correspondentes ao indivíduo i é

$$\mathbf{y}_i = (108, 113, 119, 125),$$

em que cada elemento j corresponde à altura, em centímetros, medida no instante t_j .

Ainda sobre a especificação do modelo em questão, seja \mathbf{X}_i uma matriz de variáveis $x_{i1}, x_{i2}, \dots, x_{ik}$ que se relacionam linearmente com \mathbf{y}_i . Imagine, por exemplo, que no estudo recém mencionado a altura se relacione de maneira linear com as variáveis *idade*, dada em anos, e *sexo*, que assume o valor 0 se o *sexo* for feminino e 1 se for masculino. Uma matriz de covariáveis associada a \mathbf{y}_i é dada por

$$\mathbf{X}_i = \begin{pmatrix} 5 & 0 \\ 6 & 0 \\ 7 & 0 \\ 8 & 0 \end{pmatrix}$$

em que cada linha j contém a *idade* e o *sexo* da criança observada, no instante de acompanhamento t_j .

Finalmente, seja $\boldsymbol{\epsilon}$ um vetor de erros aleatórios que mede a distância entre a resposta medida e o seu valor médio. Em geral, considera-se por suposição que estes erros são independentes entre indivíduos, e que estão igualmente distribuídos de acordo com uma distribuição normal, com média localizada no ponto zero, e variância desconhecida denotada pelo parâmetro σ^2 . Assim, o vetor $\boldsymbol{\epsilon}$ segue distribuição normal multivariada, com vetor de média zero e matriz de covariâncias $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$, em que \mathbf{I} é a matriz identidade de dimensão $n \times n$.

Para modelagem do comportamento médio da i -ésima variável resposta, no j -ésimo instante de tempo, podemos utilizar o seguinte modelo linear, considerando apenas os efeitos fixos:

$$y_{ij} = \beta_0 + x_{ij1}\beta_1 + \dots + x_{ijk}\beta_k + \epsilon_{ij},$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)$ é um vetor de coeficientes de regressão.

Voltando ao exemplo anterior, a modelagem da altura média das crianças pode ser feita considerando as covariáveis *idade* e *sexo* da seguinte maneira:

$$y_{ij} = \beta_0 + idade_{ij}\beta_1 + sexo_i\beta_2 + \epsilon_{ij},$$

Observe que *idade* é uma variável que depende do tempo. Desta forma, é indexada pelo índice *i*, que indica o indivíduo, e pelo índice *j*, que indica o instante de coleta *t_j*. Já a variável *sexo* se mantém fixa após a observação, necessitando apenas do índice relacionado ao indivíduo correspondente.

Quando especificamos o modelo desta forma não levamos em conta as características individuais de cada criança. Pode acontecer que algum elemento da amostra observada apresente um crescimento maior do que os demais. Em uma outra situação, poderíamos observar crianças que crescem de forma mais lenta ao longo do tempo. Este comportamento dependerá de características intrínsecas a cada indivíduo como, por exemplo, altura média de familiares, prática ou não de esportes, aspectos relacionados à saúde da criança, renda familiar, entre outros.

Para que o modelo de regressão consiga capturar a variabilidade relacionada aos elementos da amostra, podemos fazer a inclusão dos efeitos aleatórios. Um efeito individual que apenas desloque cada perfil de variável no eixo *y* resulta no seguinte modelo:

$$y_{ij} = \beta_0 + b_{i0} + x_{ij1}\beta_1 + \dots + x_{ijk}\beta_k + \epsilon_{ij},$$

Note que, agora, a característica particular de cada indivíduo será descrita através de uma componente aleatória b_{i0} , chamada de **intercepto aleatório**. Suponha, por exemplo, que a taxa de crescimento das crianças em estudo é a mesma, e o que muda são apenas os valores de suas alturas ao longo do tempo, isto é, observamos crianças mais altas, outras mais baixas, porém todas crescem na mesma velocidade ao longo dos anos. Um modelo adequado para esta situação é o modelo linear com intercepto aleatório:

$$y_{ij} = \beta_0 + b_{i0} + idade_{ij}\beta_1 + sexo_i\beta_2 + \epsilon_{ij}.$$

Em algumas situações é comum que a variável de interesse se comporte de maneira diferente para cada indivíduo ao longo do tempo. Nestes casos, outra componente deve ser adicionada ao modelo, capaz de modificar também a inclinação da trajetória de cada perfil. Tal modelo, chamado de **modelo linear com intercepto e coeficiente angular aleatório**, é especificado como:

$$y_{ij} = \beta_0 + b_{i0} + x_{ij1}(\beta_1 + b_{i1}) + \dots + x_{ijk}(\beta_k + b_{ik}) + \epsilon_{ij},$$

em que $x_{ij1}, x_{ij2}, \dots, x_{ijk}$ são os valores das *k* covariáveis (variáveis explicativas) associadas ao indivíduo *i* e observadas no instante *j* (instante de coleta *t_j*), $\mathbf{b}_i = (b_{i0}, b_{i1}, \dots, b_{ik})^T$ é um vetor de efeitos aleatórios adicionados ao intercepto e aos demais coeficientes β_1, \dots, β_k , respectivamente, para o indivíduo *i*, em que $i=1,2,\dots,n$, e *n* é o número total de indivíduos analisados no estudo.

Em geral, assume-se que o vetor de efeitos aleatórios \mathbf{b}_i segue distribuição normal multivariada, com vetor de médias iguais a zero e matriz de variâncias **D**. Diferentes escolhas para Σ e **D** induzem diferentes estruturas de dependência para o vetor de respostas.

Consideremos, no caso das crianças, que as taxas de crescimento das mesmas

não sejam iguais. As trajetórias das alturas ao longo do tempo serão retas com interceptos e coeficientes angular diferentes, para cada uma das crianças observadas. Um modelo adequado para esta situação é o modelo linear com intercepto e coeficientes angular aleatórios, especificado da seguinte forma:

$$y_{ij} = \beta_0 + b_{i0} + idade_{ij}(\beta_1 + b_{i1}) + sexo_i(\beta_2 + b_{i2}) + \epsilon_{ij}.$$

Na seção “Cenário Ilustrativo”, especificaremos um modelo linear de efeitos mistos, e obteremos estimativas para os valores de seus parâmetros através de pacotes do R.

O que representa cada parâmetro?

Nos modelos lineares de efeitos mistos, os parâmetros β 's são responsáveis em descrever os comportamentos **populacionais** enquanto que os termos $(\beta + b_i)$'s são responsáveis em descrever os comportamentos **individuais**.

3.2 Técnicas de Ajuste

Depois de selecionado um modelo estatístico adequado, o passo seguinte é descobrir (estimar) os valores de todos os parâmetros do modelo. Esta etapa é de extrema importância e, no caso dos modelos lineares com efeitos mistos, pode ser realizada utilizando algumas abordagens, denotadas por **formulação de dois estágios**, **abordagem marginal** e **abordagem condicional**.

Na formulação de dois estágios, considera-se no estágio 1 o ajuste de um modelo de regressão linear para cada indivíduo. Então, adotando as estimativas obtidas como sendo covariáveis, no estágio 2 ajusta-se um novo modelo de regressão multivariado. A abordagem condicional é utilizada principalmente quando o objetivo está em um subconjunto de coeficientes de regressão que não varia entre os indivíduos e, desta forma, o vetor b_i é tratado como efeitos perturbadores. Já a abordagem marginal é geralmente utilizada quando o grande objetivo é determinar os coeficientes específicos do indivíduo, sendo assim o vetor b_i é considerado como uma amostra aleatória independente de alguma distribuição. Para mais detalhes, ver Molenberghs and Verbeke (2000).

4 Cenário Ilustrativo

Nesta seção vamos mostrar como que a técnica introduzida nas seções anteriores pode auxiliar a resolução da situação problema abordada por Pedro.

Realizada a análise descritiva, é necessário especificar um modelo estatístico que melhor represente os dados observados. Os gráficos das Figuras 1 e 2 indicam uma **trajetória linear** das notas ao longo dos semestres. Observamos também que os perfis das notas se comportam de **maneira distinta para cada aluno**: algumas com trajetórias paralelas com relação ao eixo y, e outras com trajetórias de inclinações diferentes. Além

disso, é possível notar que o fator `grupo` exerce influência no comportamento longitudinal do rendimento acadêmico: alunos do grupo 1 em geral apresentam, em média, notas maiores comparadas a de alunos do grupo 2.

Considerando estas características dos dados começamos a nossa análise com um modelo linear misto simples, que pressupõe uma evolução média linear das notas no tempo, levando-se em conta o efeito do grupo na resposta média, bem como um intercepto aleatório para cada estudante envolvido no estudo. Este último termo permitirá explicar as diferenças de notas observadas entre os alunos. Assim, obtemos o seguinte modelo

$$y_{ij} = \beta_0 + b_{i0} + \beta_1 t_j + \beta_2 grupo_i + \epsilon_{ij} \quad (E.1)$$

em que a variável $grupo_i = 0$, se o aluno i pertence ao grupo 1 e $grupo_i = 1$, se o aluno i pertence ao grupo 2. A variável t_j é o j -ésimo instante de acompanhamento do aluno i , ou seja, o semestre de interesse (1, 2, 3 ou 4).

Utilizaremos nesta seção a função `lme` do pacote `nlme` do R, que ajusta o modelo linear de efeitos mistos. Para utilizá-la, começamos instalando e carregando o pacote na qual esta função se encontra:

```
> install.packages("nlme")
> require("nlme")
```

Para obter as estimativas dos parâmetros do modelo especificado na Equação (E.1), deve-se especificar no primeiro argumento da função utilizada a fórmula que relaciona a variável resposta com as covariáveis. Neste caso, assumimos que a nota se relaciona de maneira linear com as variáveis `semestre` e `grupo`, o que é representado pela sintaxe `nota ~ semestre + grupo`. Outro argumento importante a ser definido é o parâmetro `random`, utilizado para determinação dos efeitos aleatórios. Para a inclusão de um intercepto aleatório para cada aluno utilizamos a sintaxe `random = ~1|aluno`. Já a inclusão de um intercepto aleatório e coeficiente angular aleatório, por exemplo, na variável `semestre` é feita através da sintaxe `random = ~semestre|aluno`. Por fim, o argumento `data` é determinado com o nome do conjunto de dados analisado no R. No nosso exemplo, este conjunto é o `dadosNotas`. Desta forma, o modelo linear com intercepto aleatório é ajustado da seguinte maneira:

```
> fit.intercepto = lme(nota~semestre + grupo, random=~1|aluno,
data=dadosNotas)
```

```
> summary(fit.intercepto)
```

```
Linear mixed-effects model fit by REML
```

```
Data: dadosNotas
```

```
      AIC      BIC    logLik
46.73109 54.78568 -18.36555
```

```
Random effects:
```

```
Formula: ~1 | aluno
```

```
(Intercept) Residual
```

```
StdDev:    0.5155955 0.2555043
```

```

Fixed effects: nota ~ semestre + grupo
              Value Std.Error DF   t-value p-value
(Intercept)  6.96675  0.2541499 29  27.411971  0e+00
semestre     0.26370  0.0361338 29   7.297884  0e+00
grupo2      -1.86050  0.3359520  8  -5.537993  5e-04
Correlation:
      (Intr) semstr
semestre -0.355
grupo2   -0.661  0.000

Standardized Within-Group Residuals:
      Min      Q1      Med      Q3      Max
-2.11295737 -0.43598770  0.03934983  0.41335919  2.00086824

Number of Observations: 40
Number of Groups: 10

```

A função `summary()` retornará informações detalhadas a respeito do ajuste obtido. Uma das primeiras saídas apresentada pela função `summary()` são os valores do critério de informação de Akaike (AIC) e do critério de informação Bayesiano (BIC), que são medidas utilizadas para selecionar o “melhor” modelo, dentre um conjunto de modelos analisados. Sucintamente, tanto o AIC quanto o BIC comparam o ajuste do modelo com a quantidade de parâmetros e dados envolvidos. Considera-se como “melhor” modelo, dentre um conjunto de modelos analisados, aquele que tiver um menor valor de AIC ou BIC. Para mais detalhes, consultar o Capítulo “Modelos de Regressão aplicados em pesquisas em Informática na Educação”, Capítulo 11 desse livro. No exemplo, o valor do AIC é de 46.73109 enquanto que o valor do BIC é de 54.78568.

Os efeitos fixos são especificados em `Fixed effects`, em que são apresentados os valores das estimativas do intercepto, do coeficiente de regressão associado à variável `semestre` e do coeficiente de regressão associado à categoria `grupo2` do fator `grupo`, os valores dos erros padrão das respectivas estimativas, além dos valores da estatística `t` e os `p`-valores associados. A interpretação destas quantidades é análoga à apresentada no contexto dos modelos de regressão linear (ver Capítulo “Modelos de Regressão aplicados em pesquisas em Informática na Educação”, Capítulo 11 desse livro).

De fato, o intercepto descreve o comportamento populacional, sem levar em consideração características individuais dos alunos. Assim, o valor de 6.96675 para o intercepto indica a estimativa da nota média dos alunos do `grupo1` (grupo de referência, escolhida automaticamente pelo software estatística R) no início do estudo, sem levar em consideração a variável `semestre` e as características individuais de cada aluno. No caso de alunos do `grupo2` na mesma situação, isto é, desconsiderando o `semestre` e as características individuais, deve-se adicionar o valor -1.86050 (valor associado ao `grupo2` e que indica que há uma diminuição na estimativa da nota média dos alunos pertencentes ao grupo 2) à estimativa da nota média dada ao grupo 1. Desta forma, tem-se que $6.96675 - 1.86050 = 5.10625$ é a estimativa da nota média dos alunos do `grupo2` no início do estudo, desconsiderando o `semestre` e as características individuais.

O valor de 0.26370 para o coeficiente associado à variável *semestre* indica que ao “passar” de um semestre para o outro, a nota média estimada de cada aluno aumenta em 0.26370, desconsiderando as características individuais de cada aluno.

Os valores dos erros padrões apresentados são utilizados nos testes estatísticos para a verificação da significância das variáveis no estudo, isto é, se de fato a variável (fator) é importante (necessária) para estimar a resposta média. Os testes estatísticos são baseados em estatísticas de teste e os valores dessas estatísticas, considerando os dados observados, são apresentados (*t-value*). Em geral, valor grande dessa estatística, desconsiderando o sinal, indica que a variável (ou fator) é importante para o modelo. Já valor próximo de zero indica que a variável (ou fator) em questão não contribui para a resposta média e, portanto, pode ser desconsiderada do modelo. O valor ser considerado “grande” ou “pequeno” depende do número de observações, do número de parâmetros do modelo e do erro adotado no teste em questão. Em geral adota-se um erro de 5%, 10%.

Uma maneira mais direta de avaliar a significância das variáveis (fatores) é utilizando o valor *p* (*p-value*), que é calculado a partir do *t-value*. Um *p-value* menor do que o erro adotado indica que a variável (fator) é importante para o modelo (pois nesse caso, o *t-value* é considerado estatisticamente grande). Já *p-value* maior ou igual do que o erro adotado indica que a variável (fator) pouco contribui para a variável resposta (pois nesse caso, o *t-value* é considerado estatisticamente pequeno) e, portanto, pode ser desconsiderada do modelo. No exemplo, temos que o *p-value* para *Intercept*, *semestre* e *grupo2* são 0e+00, 0e+00 e 5e-04, respectivamente. Considerando um erro de 5%, dizemos que o intercepto, a variável *semestre* e o fator *grupo2* são significativos para o modelo e, portanto, influenciam na nota média estimada de cada aluno.

Na seção *Random effects* é possível obter o desvio padrão do intercepto aleatório (*Intercept*) 0.5155955, e dos erros aleatórios (*Residual*) 0.2555043. Para obter as previsões dos efeitos aleatórios (os valores dos interceptos aleatórios) para cada indivíduo basta utilizar a função *ranef()* :

```
> ranef(fit.intercepto)
  (Intercept)
1 -0.09515798
2  0.60674988
3 -0.24825868
4 -0.52383996
5  0.26050674
6  0.36932601
7  0.57895621
8  0.22093610
9 -0.68118037
10 -0.48803794
```

Conforme já descrito, o valor de 6.96675 para o intercepto indica a estimativa da nota média dos alunos no grupo1 no início do estudo, não levando em conta a variável *semestre* e as características individuais de cada aluno. Com a inserção dos

efeitos aleatórios, tais características são consideradas. Através da função `coef()` é possível extrair os valores dos coeficientes já somados as suas respectivas componentes aleatórias.

```
> coef(fit.intercepto)
      (Intercept) semestre grupo2
1      6.871592      0.2637 -1.8605
2      7.573500      0.2637 -1.8605
3      6.718491      0.2637 -1.8605
4      6.442910      0.2637 -1.8605
5      7.227257      0.2637 -1.8605
6      7.336076      0.2637 -1.8605
7      7.545706      0.2637 -1.8605
8      7.187686      0.2637 -1.8605
9      6.285570      0.2637 -1.8605
10     6.478712      0.2637 -1.8605
```

Observe que no grupo 1 o indivíduo 2, por exemplo, inicia o estudo (não considerando o semestre) com uma nota média de 7.573500 ao passo que o indivíduo 4, no mesmo grupo, possui uma média menor, com valor de 6.442910. Esta diferença é devida ao fato de que as suas características individuais estão sendo inseridas na modelagem. Ao considerar indivíduos no grupo 2 (alunos 6, 7, 8, 9 e 10), a análise deve ser repetida, porém, o valor -1.8605 deve ser adicionado à estimativa da nota média de cada indivíduo no grupo 1. Por exemplo, o indivíduo 8 inicia o estudo com uma nota média de $7.187686 - 1.8605 = 5.327186$. Já o indivíduo 10 no grupo 2 possui uma nota média inicial de $6.478712 - 1.8605 = 4.618212$.

Para estimar a nota média de um indivíduo no grupo 2 em um determinado semestre, basta adicionar $(0.2637 \times \text{Número do semestre correspondente})$ aos valores obtidos anteriormente. Por exemplo, a estimativa da nota média do indivíduo 9, pertencente ao grupo 2, no terceiro semestre, seria $6.285570 + (0.2637 \times 3) - 1.8605 = 5.21617$.

Adicionaremos agora à nossa análise um efeito aleatório no coeficiente angular para explicar as diferentes angulações de perfis detectadas na análise descritiva. Assim, o modelo linear com intercepto e coeficiente angular aleatório para o problema proposto é especificado da seguinte forma:

$$y_{ij} = (\beta_0 + b_{i0}) + (\beta_1 + b_{i1})t_j + \beta_2 \text{grupo}_i + \epsilon_{ij}, \quad (\text{E.2})$$

em que a variável $\text{grupo}_i = 0$, se o aluno i pertence ao grupo 1 e $\text{grupo}_i = 1$, se o aluno i pertence ao grupo 2, e seu ajuste é dado através da seguinte linha de comando:

```
> fit.int.coef <- lme(nota~semestre + grupo, random=~semestre|aluno,
data=dadosNotas)
> summary(fit.int.coef)
Linear mixed-effects model fit by REML
Data: dadosNotas
      AIC      BIC    logLik
```

-17.52806 -6.251639 15.76403

Random effects:

Formula: ~semestre | aluno

Structure: General positive-definite, Log-Cholesky parametrization

	StdDev	Corr
(Intercept)	0.62900947	(Intr)
semestre	0.20305389	-0.568
Residual	0.03610401	

Fixed effects: nota ~ semestre + grupo

	Value	Std.Error	DF	t-value	p-value
(Intercept)	6.869213	0.2581256	29	26.611898	0e+00
semestre	0.263700	0.0644140	29	4.093833	3e-04
grupo2	-1.665426	0.3278294	8	-5.080162	1e-03

Correlation:

	(Intr)	semstr
semestre	-0.440	
grupo2	-0.635	0.000

Standardized Within-Group Residuals:

	Min	Q1	Med	Q3	Max
	-1.377072780	-0.541337807	0.001028381	0.675780258	1.144364429

Number of Observations: 40

Number of Groups: 10

A interpretação da saída desta função é análoga àquela apresentada para o modelo (E.1). A diferença é que podemos extrair os valores referentes ao coeficiente angular aleatório:

```
> coef(fit.int.coef)
  (Intercept)  semestre  grupo2
1    6.954371  0.22824066 -1.665426
2    7.319859  0.37973838 -1.665426
3    6.223962  0.45556479 -1.665426
4    5.979500  0.43648496 -1.665426
5    6.712700  0.47581487 -1.665426
6    7.933975 -0.04467331 -1.665426
7    7.542733  0.20080364 -1.665426
8    6.687315  0.39122418 -1.665426
9    6.926568 -0.08718487 -1.665426
10   6.411150  0.20098670 -1.665426
```

Observe que, além de interceptos que captam a variabilidade entre os indivíduos, temos também diferentes coeficientes angulares relacionados à variável semestre, o que permite que o modelo comporte as diferentes tendências lineares (trajetórias com diferentes origens e inclinações) de cada aluno. Assim, a contribuição da variável semestre muda de indivíduo para indivíduo, para um mesmo instante de tempo analisado, independente do grupo ao qual pertence.

Note que o modelo (E.2) possui valores de AIC de -17.52806 e de BIC de -

6.251639. Como esses valores são menores do que os obtidos no modelo (E.1), concluimos que o modelo (E.2) é mais adequado a este tipo de dados.

Apesar do modelo linear com intercepto e coeficiente angular aleatório ter sido indicado pelos critérios AIC e BIC como o mais apropriado dentre os dois modelos analisados, é necessário verificar a qualidade do ajuste. Isto é possível através de uma análise de resíduos cujo objetivo é verificar se as suposições básicas a respeito das distribuições para os modelos são adequadas. Para verificar a normalidade dos resíduos utilizamos a seguinte função:

```
> qqnorm(fit.int.coef, ~ resid(.) | grupo, pch=20)
```

O resultado é apresentado na Figura 3. Quanto mais próximos os pontos estão de uma reta, mais próximos os resíduos estão da distribuição normal.

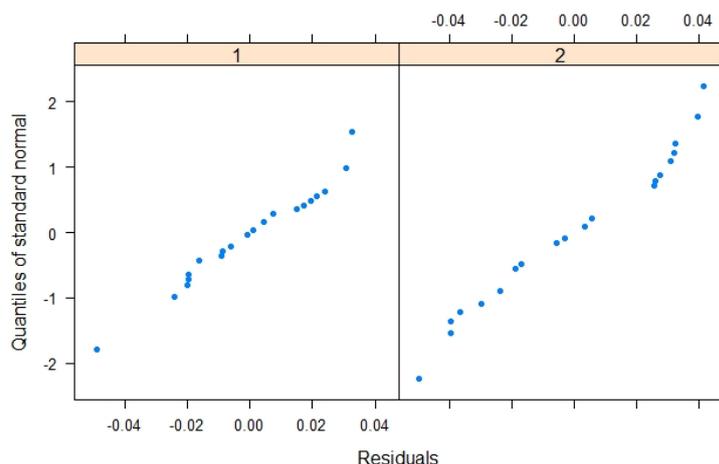


Figura 3: Gráfico Quantil-Quantil dos resíduos divididos por grupos.

Notamos pela Figura 3 que os pontos seguem uma tendência linear, dando indícios de normalidade para os resíduos.

Outro gráfico bastante utilizado para verificar o ajuste do modelo é o gráfico dos resíduos *versus* valores ajustados. Espera-se que os resíduos estejam distribuídos ao longo do zero e em uma faixa constante variando entre -2 e 2, ou seja, tenham variância constante. Para construirmos esse gráfico, utilizamos a seguinte função:

```
> plot(fit.int.coef, ylim = c(-2,2), pch=20)
```

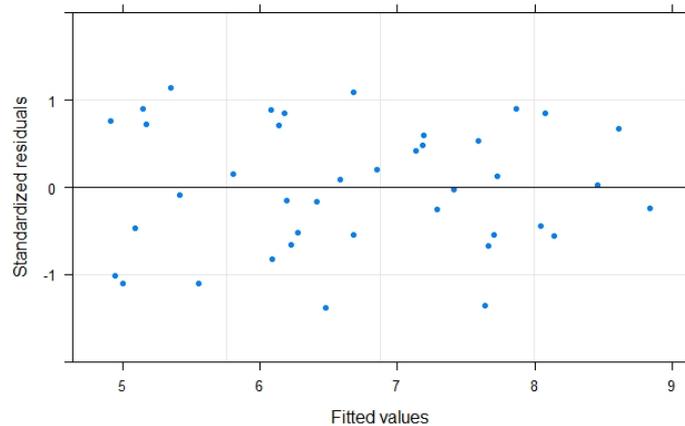


Figura 4: Gráfico dos resíduos *versus* valores ajustados, considerando o modelo (E.2).

Pelo gráfico da Figura 4 é possível notar que os valores dos resíduos se concentram em torno do zero e estão distribuídos em uma faixa entre -2 e 2, indicando que o modelo (E.2) se ajusta bem ao conjunto de dados proposto. Desta forma, é possível utilizar com confiança os resultados obtidos no ajuste (função `coef(fit.int.coef)`) para a análise final do problema.

De uma maneira geral, pode-se notar que alunos no grupo 2 tem notas médias estimadas em geral, menores do que alunos no grupo 1, já que a variável `grupo2` assume um valor negativo. Além disso, no grupo 2 o desempenho de dois alunos (indivíduos 6 e 9) diminuem com o passar dos semestres (valores negativos para a variável `semestre` desses indivíduos), o que não ocorre no grupo 1, indicando que o uso da metodologia “sala de aula invertida” além de produzir notas médias em geral maiores, faz com que o desempenho aumente com o passar do tempo, sendo uma ótima ferramenta para o aumento do aprendizado. Portanto, a metodologia proposta por Pedro se mostrou eficaz para o objetivo proposto.

5 Resumo

Nesse capítulo foram apresentadas as principais características dos estudos longitudinais assim como técnicas para resolução de problemas neste contexto. Inicialmente foram definidos os estudos transversais, que são adequados e aplicáveis a diversos tipos de situações e têm como atributo principal a sua menor duração de tempo quando comparado aos estudos longitudinais. Além disso, podem ser considerados de mais fácil execução e, conseqüentemente, de menor custo financeiro. Na sequência, foram definidos os estudos longitudinais que conseguem captar informações referentes à variabilidade de cada indivíduo, e apresentar uma tendência de comportamento da variável observada em um período de tempo. Foram apresentadas as principais técnicas descritivas utilizadas para análise de dados longitudinais e, posteriormente, a modelagem utilizada para tais problemas. Uma vez definido o modelo matemático (estatístico), técnicas de estimação foram sucintamente mencionadas. Pacotes do

software estatístico R, úteis para ajustar modelos desta natureza, foram apresentados destacando-se suas principais características. Toda a metodologia apresentada foi exemplificada em um exemplo real associado à Seção “Era uma vez”.

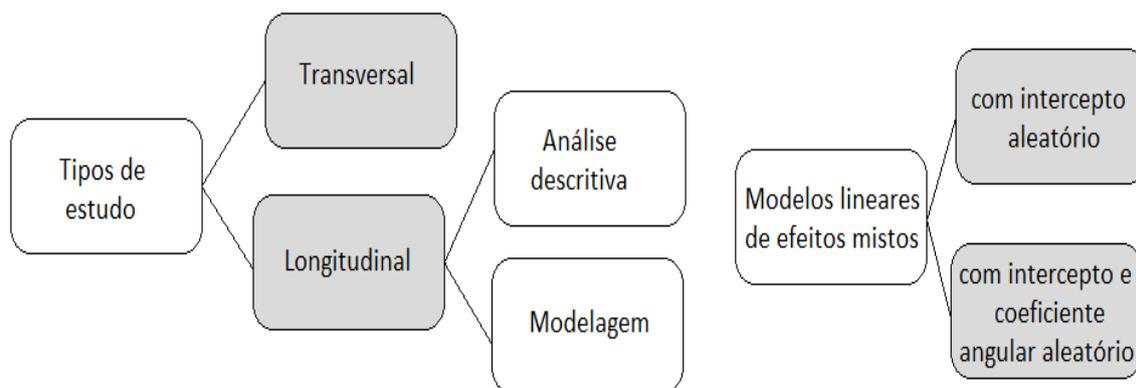


Figura: Mapa mental das teorias e diretrizes para estudo longitudinal.

6 Leituras Recomendadas

- **Análise de Dados Longitudinais** (versão parcial preliminar) (SINGER et al., 2018). Este material ilustra diversos estudos com estrutura longitudinal, apresentando os principais conceitos que caracterizam esse tipo de dados. São apresentados métodos básicos para a análise estatística, sendo assim uma boa referência para leitores que estão iniciando na área.
- **Mixed-effects Models in S and S-Plus** (PINHEIRO; BATES, 2000). Este livro aborda de maneira geral a teoria de modelos de efeitos mistos lineares e não-lineares na análise de dados agrupados, com destaque para dados longitudinais e medidas repetidas. A fim de simplificar a teoria, é apresentada uma estratégia de construção dos modelos além de algumas aplicações a dados reais. Códigos computacionais na linguagem R são fornecidos, sendo assim um bom guia para iniciantes na área.
- **Applied Longitudinal Analysis** (FITZMAURICE et al., 2011). Nesse livro, são apresentados os conceitos básicos e alguns problemas reais, das mais variadas áreas do conhecimento, envolvendo estudos longitudinais, possibilitando ao leitor um melhor entendimento das técnicas utilizadas em tais estudos.
- **Linear mixed models for longitudinal data** (MOLENBERGHS; VERBEKE, 2000). Assim como a referência anterior, neste livro são apresentados conceitos e exemplos de problemas em que se utilizam dados longitudinais. Recomenda-se a leitura para um aprofundamento do assunto e técnicas envolvidas.

7 Artigos exemplos

- **High school curriculum structure: Effects of course-taking and achievement in mathematics for high school graduates. An examination of data from the National Education Longitudinal Study of 1988** (LEE et al., 1998). Neste estudo, os autores investigam como a estrutura do currículo do ensino médio influencia o quão longe os graduados recebem o curso de matemática secundária e seu nível de realização nesse assunto até o final do ensino médio. Utilizam um estudo longitudinal de base ampla sobre as experiências dos adolescentes dos EUA de escolas secundárias públicas e privadas nas 30 maiores cidades do país. Assim, é uma importante aplicação da metodologia apresentada neste capítulo.
- **College preparatory curriculum for all: Academic consequences of requiring algebra and English I for ninth graders in Chicago.** (ALLENSWORTH et al., 2009). Este artigo avalia uma política implantada em Chicago, para alunos do ensino médio, que retirou as aulas de reforço, priorizando apenas o trabalho nos cursos preparatórios. Para a realização deste estudo, foi utilizado um arquivo de dados longitudinais contendo registros administrativos completos para cada aluno em cada semestre, transcrições de cursos semestrais e secundários e resultados de testes de conclusão do ensino fundamental e médio, entre outras informações.
- **Quality and Equality in Brazilian Secondary Schools: A Multilevel Cross-national School Effects Study.** (LEE et al., 2009). Este estudo identifica características das escolas secundárias brasileiras ligadas ao alto sucesso (qualidade) e a uma distribuição social equitativa de sucesso (igualdade). É destacada a importância de os projetos de coleta de dados serem longitudinais para estimar com precisão os efeitos da escola. Neste caso, o pesquisador precisaria de dados sobre os alunos (realização, em particular) quando eles entram e quando eles saem da escola.

8 Checklist

Resumidamente, para realizar uma pesquisa utilizando os conceitos e técnicas de estudos longitudinais, você terá que realizar as seguintes atividades:

- Observar as variáveis de interesse, nos indivíduos do estudo, ao longo do tempo;
- Realizar uma análise descritiva, apresentando as medidas resumo e gráficos;
- Extrair o máximo de informação dos dados tais como as médias dentro de cada grupo (tratamento, técnica), média geral dos indivíduos, variância entre os indivíduos do mesmo grupo (tratamento, técnica), variância de cada indivíduo, valores atípicos (outliers), entre outros;
- Eleger o modelo mais adequado ao conjunto de dados em estudo:

- Inserir os efeitos aleatórios apenas nos termos necessários, de acordo com os resultados da análise descritiva;
- Ajustar o modelo através de alguma técnica de estimação, utilizando um dos pacotes do software estatístico R;
- Responder às perguntas de interesse, de acordo com o modelo final ajustado.

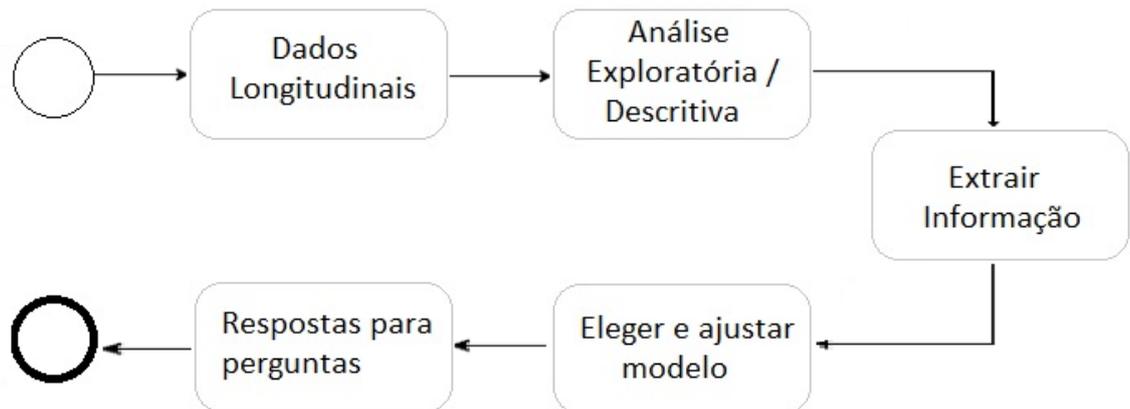


Figura: Fluxograma de atividades para estudo longitudinal.

9 Exercícios

- 1) Enumere as principais diferenças entre os estudos transversal e longitudinal.
- 2) Um professor do departamento de informática de uma universidade brasileira verificou que os alunos ingressantes no curso de informática de sua universidade, nos últimos anos, tinham dificuldades em aprender os conceitos básicos de programação, o que resultava em notas ruins nas avaliações. Pensando em melhorar o nível de aprendizado, o professor desenvolveu uma nova técnica de ensino com o auxílio de recursos multimídias. Para testar se a técnica de fato é eficaz, ou seja, se com a aplicação de um novo método de ensino o aluno consegue melhorar o desempenho acadêmico, foram selecionados de maneira aleatória oito alunos que foram divididos em dois grupos (A e B) contendo cada grupo quatro alunos. Os alunos do grupo A receberam o método de ensino novo, proposto pelo professor, enquanto os alunos do grupo B continuaram a aprender os conceitos de programação da forma tradicional. A cada 15 dias de aprendizado, um teste era aplicado para todos os alunos e as notas eram anotadas. Assim, a variável resposta de interesse é **Nota**. Note que os objetivos deste estudo são comparar os grupos e avaliar as mudanças ao longo dos dias. Os dados são apresentados a seguir:

Grupo A			Grupo B		
Aluno	Tempo (dias)	Nota	Aluno	Tempo (dias)	Nota
1	15	7,0	1	15	6,3
1	30	8,2	1	30	6,4
1	45	7,9	1	45	6,2
1	60	8,5	1	60	6,6
2	15	7,3	2	15	6,1
2	30	7,9	2	30	6,3
2	45	8,2	2	45	6,4
2	60	9,3	2	60	6,3
3	15	8,0	3	15	6,3
3	30	7,9	3	30	7,0
3	45	7,9	3	45	7,1
3	60	8,1	3	60	6,8
4	15	8,3	4	15	6,0
4	30	8,2	4	30	6,3
4	45	8,0	4	45	6,7
4	60	8,0	4	60	6,5

Faça uma análise descritiva dos dados referente à variável **Nota**.

- 3) Baseado nos resultados obtidos no exercício anterior, proponha um ou mais modelos para estes dados. Ajuste os modelos e interprete os resultados. Existem diferenças entre os grupos? O novo método de aprendizagem é eficaz?
- 4) A **média do tempo de estudo** diário em cada semestre, dos alunos de duas turmas do curso de Informática de uma universidade norte-americana, são analisados ao longo dos 3 primeiros anos da graduação. A turma 1 (T_1) tem aulas durante o período matutino, enquanto que a turma 2 (T_2) tem aulas no período da noite. Cada turma é composta por 6 alunos sendo que em cada semestre o **tempo médio de estudo**, em horas, de cada estudante em cada uma das turmas é registrado. A suspeita é de que alunos que cursam a graduação pela manhã em geral gastam mais tempo dedicado aos estudos. Além disso, espera-se que a medida que os semestres avancem, o tempo médio gasto com estudos também aumentem. Os dados são apresentados a seguir.

Turma 1					
Aluno	Semestre	Tempo Médio (Horas)	Aluno	Semestre	Tempo Médio (Horas)
1	1	1,26	4	1	3,53
1	2	0,73	4	2	2,38

1	3	2,45	4	3	5,74
1	4	1,71	4	4	6,72
1	5	1,59	4	5	6,82
1	6	3,58	4	6	9,02
2	1	4,17	5	1	2,67
2	2	3,35	5	2	2,84
2	3	4,27	5	3	4,35
2	4	6,01	5	4	5,94
2	5	7,53	5	5	5,02
2	6	7,49	5	6	7,96
3	1	1,21	6	1	3,69
3	2	3,77	6	2	3,12
3	3	3,51	6	3	4,82
3	4	5,12	6	4	5,43
3	5	5,48	6	5	8,03
3	6	6,63	6	6	7,81
Turma 2					
Aluno	Semestre	Tempo Médio (Horas)	Aluno	Semestre	Tempo Médio (Horas)
1	1	0,08	4	1	1,12
1	2	0,12	4	2	1,34
1	3	2,18	4	3	1,98
1	4	3,13	4	4	5,64
1	5	3,42	4	5	6,27
1	6	3,83	4	6	7,34
2	1	1,54	5	1	1,76
2	2	2,71	5	2	1,92
2	3	3,12	5	3	4,01
2	4	4,63	5	4	4,34
2	5	4,55	5	5	5,87
2	6	4,14	5	6	5,17
3	1	3,2	6	1	1,96
3	2	1,09	6	2	2,94
3	3	3,2	6	3	4,92
3	4	3,04	6	4	4,09
3	5	5,44	6	5	4,19
3	6	5,69	6	6	5,75

As suspeitas iniciais são verificadas nos dados? Para responder a essa pergunta,

faça uma análise exploratória dos dados, considerando a variável de interesse **tempo médio de estudo**, em horas.

- 5) Baseado nos resultados obtidos no exercício anterior, proponha algum modelo para representar os dados. Ajuste o modelo e interprete os resultados.

10 Referências

- ALLENSWORTH, E.; NOMI, T.; MONTGOMERY, N.; LEE, V.E. College preparatory curriculum for all: Academic consequences of requiring algebra and English I for ninth graders in Chicago. **Educational Evaluation and Policy Analysis**, 2009. V. 31, n. 4, p. 367-391.
- FITZMAURICE, G.M.; LAIRD, N.M.; WARE, J.H. **Applied Longitudinal Analysis**, 2nd ed. New York: Wiley, 2011.
- JIANG, J. **Linear and generalized linear mixed models and their applications**. Springer, New York, New York, USA, 2007.
- LEE, V.E.; BURKAM, D.T.; CHOW-HOY, T.; SMERDON, B.A.; GOVERDT, D. **High school curriculum structure: Effects of course-taking and achievement in mathematics for high school graduates. An examination of data from the National Education Longitudinal Study of 1988** (Working Paper No. 98-09). Washington, DC: U.S. Department of Education, Office of Educational Research and Development, National Center for Educational Statistics, 1998.
- LEE, V.E.; FRANCO, C.; ALBERNAZ, A. Quality and equality in Brazilian secondary schools: A multilevel cross-national school effects study. **International Review of Comparative Sociology**, 2009. V. 1, p. 25–61.
- MOLENBERGHS, G.; VERBEKE, G. **Linear mixed models for longitudinal data**, New York: Springer, 2000.
- PINHEIRO, J. C., BATES, D. M. **Mixed-effects Models in S and S-Plus**. Springer, 2000.
- R CORE TEAM. R: A language and environment for statistical computing. **R Foundation for Statistical Computing**, Vienna, Austria, 2016. Disponível em: <http://www.R-project.org>
- SINGER, J.M.; NOBRE, J.S.; ROCHA, F.M.M. **Análise de Dados Longitudinais** (versão parcial preliminar). São Paulo: Universidade de São Paulo, 2018. Disponível em: <http://www.ime.usp.br/~jmsinger/MAE0610/Singer&Nobre&Rocha2018jun.pdf>

Sobre os autores



Aline Campos Reis de Souza

<http://lattes.cnpq.br/1026084833654918>

Doutoranda em Estatística pelo Instituto de Matemática e Estatística da Universidade de São Paulo, mestre em Estatística pela Universidade de São Paulo e Universidade Federal de São Carlos (2016) e graduada em Matemática Aplicada e Computação Científica pela Universidade de São Paulo (2012). Atualmente é professora assistente na Universidade Estadual de Maringá (UEM).



Willian Luís de Oliveira

<http://lattes.cnpq.br/4423792653643417>

Doutor em Estatística pela Universidade Federal de São Carlos (2016), mestre em Ciências de Computação e Matemática Computacional na área de Estatística pela Universidade de São Paulo (2011) e graduado em Matemática Aplicada e Computação Científica pela Universidade de São Paulo (2008). Possui pós-doutorado em Estatística pela Universidade de São Paulo (2017). Atualmente é professor adjunto na Universidade Estadual de Maringá (UEM). Tem experiência na área de Matemática, Programação, Probabilidade e Estatística, atuando principalmente nos seguintes temas: Modelos de Regressão, Modelos Bivariados, Modelos Semiparamétricos e Estatística Computacional.