

Análise Quantitativa de Dados em Linguística

Modelos de Efeitos Mistos

Ronaldo Lima Jr.

`ronaldojr@letras.ufc.br`

`ronaldolimajr.github.io`

Universidade Federal do Ceará

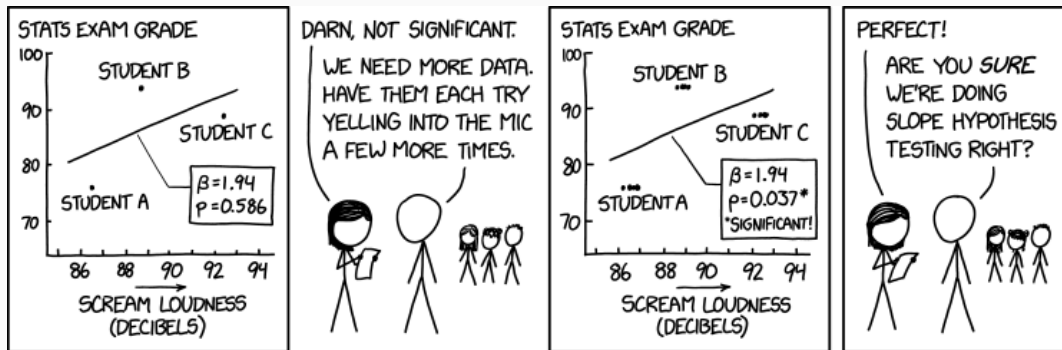
1. Conceito
2. Modelos
3. Recomendações

Conceito



- Modelo de Efeitos Mistos
- Modelo Misto
- Modelo de Efeitos Aleatórios
- Modelo Hierárquico
- Modelo Multinível

Motivação



<https://xkcd.com/2533/>

- Modelos sem efeitos aleatórios (mistos) assumem **independência dos dados** (como nas jogadas de um dado ou de uma moeda)
- Mas dados (principalmente em linguística) raramente são independentes:
ex.: cada participante normalmente contribui com mais de um dado (fazemos várias coletas de cada participante), gerando **dependência** entre dados
- Precisamos informar o modelo que há dados advindos de um mesmo participante (ou de um mesmo item lexical, mesmo texto, mesmo autor, etc.)
- Violar o pressuposto da independência dos dados aumenta as chances de **erro de tipo I** (como na tirinha anterior)

Tentativas de solução

- Uma tentativa de solucionar a dependência dos dados é de usar as médias dos participantes em testes de hipótese (com `aggregate()`, por exemplo)
 - O problema é que quando tiramos médias, perdemos informações importantes, principalmente informação sobre **variação** – se o modelo só “enxerga” médias, ele vai subestimar a quantidade de variação (além das outras desvantagens de testes de hipótese diante de modelos de regressão)
- Outra tentativa é de usar *repeated-measures ANOVAs*
 - Ao contrários das ANOVAs, os modelos de efeitos mistos permitem múltiplos agrupamentos e lidam bem com dados desbalanceados (além das demais vantagens de modelos sobre testes de hipótese)

Modelos de Efeitos Mistos

- Modelos de efeitos mistos combinam **efeitos fixos** e **efeitos aleatórios**
- De maneira geral, continuamos interessados nos efeitos fixos, mas precisamos dar conta da variância individual sobre os dados
- **Efeitos fixos**: *intercept* e *slopes* dos efeitos de interesse (como já vimos)
 - fixos porque podemos testá-los novamente em um novo estudo
- **Efeitos aleatórios**: quanto cada indivíduo/item/etc. varia em relação ao efeitos fixos
 - aleatórios porque são efeitos idiossincráticos que mudam com mais coletas ou um novo estudo

Exemplo 1

- Efeitos de uma prática pedagógica específica (tratamento) sobre o desempenho de alunos em um teste (variável resposta, medição)
- Uma turma com intervenção pedagógica (experimental) outra sem (controle)
- Ambas fazem pré-teste e pós-testes

Variável resposta: diferença das notas pós - pré

Variável preditora: Turma (intervenção vs controle)

- Queremos generalizar este efeito para novos dados
- Em novo estudo, continuamos interessado neste efeito
- **Efeito fixo**

Exemplo 1

- Efeitos de uma prática pedagógica específica (tratamento) sobre o desempenho de alunos em um teste (variável resposta, medição)
- Uma turma com intervenção pedagógica (experimental) outra sem (controle)
- Ambas fazem pré-teste e pós-testes

Alunos:

- **Dependência:** cada aluno contribui com mais de um dado (vários itens dos testes); alunos diferentes têm características intrínsecas que afetam suas notas
- **Repetição/replicação/generalização:** não estamos interessados no efeito de ser Marcos, Alice, Juca, Lúcia, etc. – com mais dados ou em outro estudo serão outros alunos (amostras aleatórias da população de alunos)
- **Efeito aleatório!**

Exemplo 1

- Efeitos de uma prática pedagógica específica (tratamento) sobre o desempenho de alunos em um teste (variável resposta, medição)
- Uma turma com intervenção pedagógica (experimental) outra sem (controle)
- Ambas fazem pré-teste e pós-testes

Itens:

- **Dependência:** cada item contribui com mais de um dado (vários alunos fazendo cada item); itens diferentes têm características intrínsecas que afetam os resultados
- **Repetição/replicação/generalização:** não estamos interessados no efeito do item 1, 2, 3, etc. – com mais dados ou em outro estudo podem ser outros itens (amostras aleatórias da população de possíveis itens)
- **Efeito aleatório!**

Exemplo 2

- Efeito de (i) tempo de residência no Brasil, (ii) idade de chegada ao Brasil, (iii) quantidade de exposição ao PB; e (iv) ter aulas de PB sobre a pronúncia das vogais médias de imigrantes hispanofalantes
- Imigrantes com diferentes características (ii), (iii) e (iv) gravados mensalmente durante 1 ano
- Participantes gravados lendo palavras com as vogais médias (e palavras distratoras)

Variável resposta: pista acústica da gravação (F1, F2, duração, distância euclidiana)

Variáveis preditoras: tempo, idade de chegada, exposição ao PB, aulas de PB

- Queremos generalizar esses efeitos para novos dados
- Em novo estudo, continuamos interessado nesses efeitos
- Efeitos fixos

Exemplo 2

- Efeito de (i) tempo de residência no Brasil, (ii) idade de chegada ao Brasil, (iii) quantidade de exposição ao PB; e (iv) ter aulas de PB sobre a pronúncia das vogais médias de imigrantes hispanofalantes

Participantes:

- **Dependência:** cada participante contribui com mais de um dado (várias palavras); participantes diferentes têm características intrínsecas que afetam os resultados
- **Repetição/replicação/generalização:** não estamos interessados no efeito de ser Alejandro, Gabriela, Juan, Ana, etc. – com mais dados ou em outro estudo serão outros participantes (amostras aleatórias da população de hispanofalantes)
- **Efeito aleatório!**

Exemplo 2

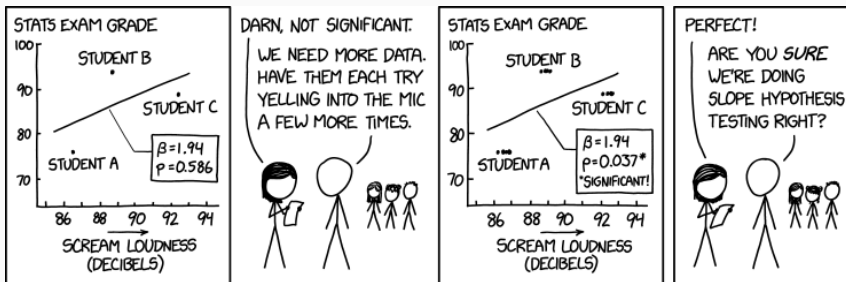
- Efeito de (i) tempo de residência no Brasil, (ii) idade de chegada ao Brasil, (iii) quantidade de exposição ao PB; e (iv) ter aulas de PB sobre a pronúncia das vogais médias de imigrantes hispanofalantes

Palavras:

- **Dependência:** cada palavra contribui com mais de um dado (vários participantes lendo a mesma palavra); palavras diferentes têm características intrínsecas que afetam os resultados
- **Repetição/replicação/generalização:** não estamos interessados no efeito da palavra avó, avô, rego, r(ê)go, etc. – com mais dados ou em outro estudo podem ser outras palavras (amostras aleatórias da população de palavras com vogais médias)
- **Efeito aleatório!**

Modelos (de efeitos) mistos

- Não queremos modelar os participantes ou itens específicos (*overfitting*)
- Não queremos ignorar o efeito da variação intrínseca de participantes ou itens (erro de tipo I)



Modelos (de efeitos) mistos

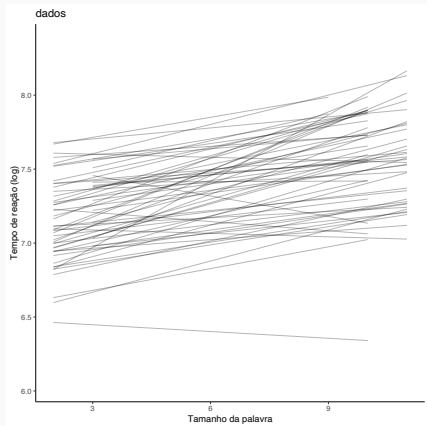
- Queremos um meio do caminho: modelar para além dos participantes (itens, etc.) específicos a fim de generalizar para novos dados, mas informando o modelo de que há dados dependentes/aleatórios

→ Efeitos fixos + Efeitos aleatórios = Efeitos mistos

Modelos (de efeitos) mistos

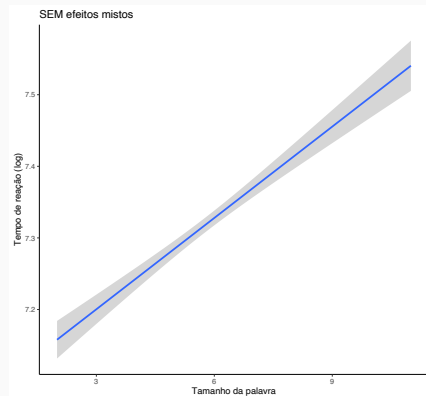
Efeito do tamanho da palavra (em número de letras) sobre tempo de reação

Dados



MODELO
MISTO

Regressão Linear SEM efeitos mistos



Modelos (de efeitos) mistos

- Lembrando:
 - **Efeitos fixos**: *intercept* e *slopes* (como temos feito)
 - **Efeitos aleatórios**: quanto cada participante (item, etc.) varia em relação aos efeitos fixos (em desvio-padrão)

Modelos

Dados beginningReaders do pacote languageR

- 7.923 dados de tempo de reação/resposta (*lexical decision latencies*) de 59 crianças holandesas de 8 anos de idade a 184 palavras
- Trata-se de um desenho experimental de medidas repetidas (*repeated measures design*)
- Cada palavra tem 59 medidas repetidas (1 de cada participante)
- Cada participante tem 184 medidas repetidas (1 para cada palavra)
- Consequentemente, **Subject** e **Word** serão variáveis aleatórias no modelo

Dados beginningReaders do pacote languageR

Possíveis variáveis preditoras:

- Trial: rank na lista do experimento (*integer*)
- OrthLength: tamanho da palavra em número de letras (*integer*)
- LogFrequency: log da frequência da palavra em materiais de leitura para crianças (*number*)
- LogFamilySize: log do tamanho da família morfológica da palavra (*number*)
- ReadingScore: nota em um teste de leitura (*integer*)
- ProportionOfErrors: proporção de erro por palavra (*number*)

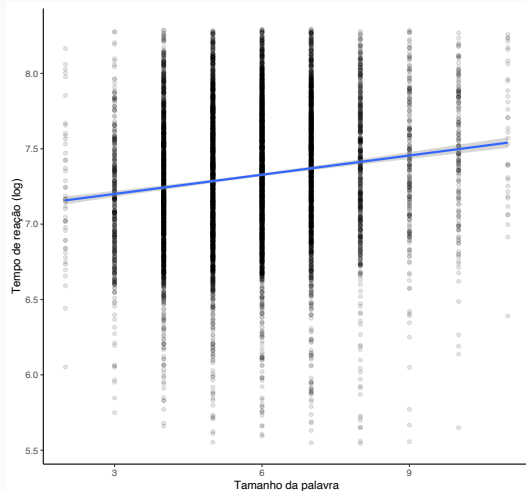
Modelo sem efeitos aleatórios

Efeito do tamanho da palavra (em número de letras) sobre tempo de reação:

```
1 mod1 = lm(LogRT ~ OrthLength, data = readers)
2 summary(mod1)

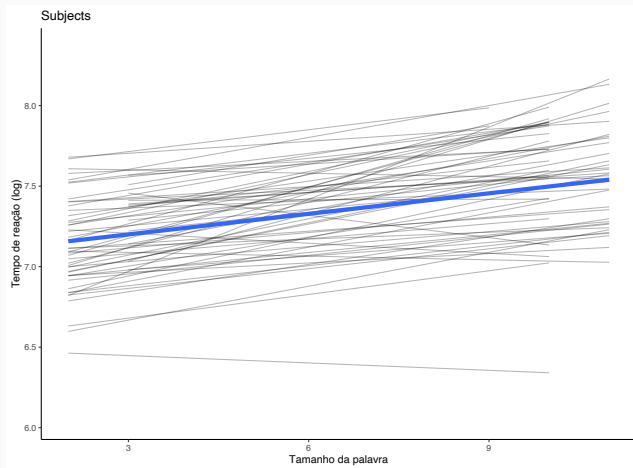
3 Coefficients:
4             Estimate Std. Error t value Pr(>|t|)
5 (Intercept)  7.072771   0.019593  360.99  <2e-16 ***
6 OrthLength   0.042520   0.003275  12.98  <2e-16 ***
```

→ O tempo de reação aumenta (significativamente) com o aumento do tamanho da palavra

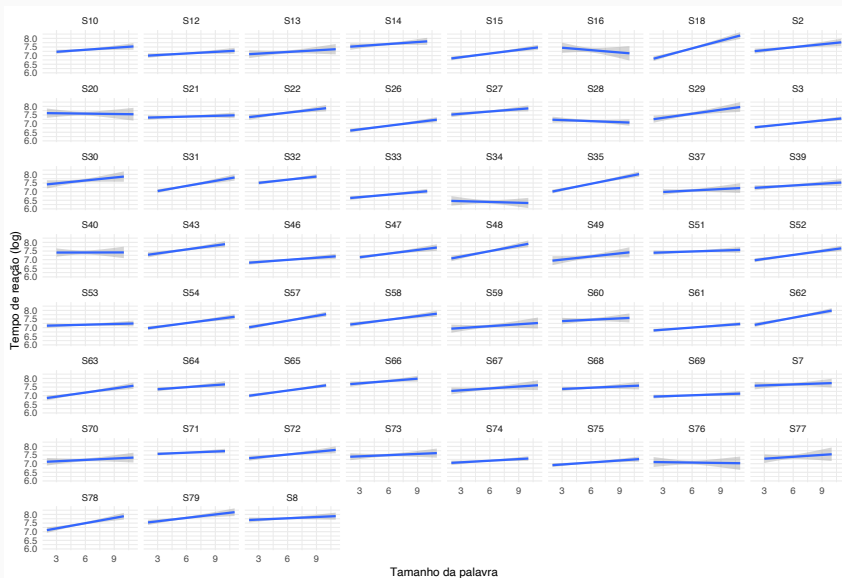


Modelo sem efeitos aleatórios

O modelo sem efeitos mistos “força” todos os participantes (e todas as palavras) a terem o mesmo *intercept* e o mesmo *slope* (linha azul)



Variação individual



Modelo com intercepts aleatórios

- Como primeiro passo, podemos informar ao modelo que participantes podem ter valores de partida variáveis (*random/varying intercepts*)
- Utilizamos a função `lmer` do pacote `lme4` (utilizaremos `glmer` quando for uma regressão logística de efeitos mistos)
- Utilizamos `(1 | Subject)` para isso
(o 1 indica “*intercept*”)

```
1 | mod2 = lmer(LogRT ~ OrthLength + (1 | Subject), data = readers)
2 | summary(mod2)
```

Modelo com intercepts aleatórios

```
1 Random effects:
2   Groups      Name                Variance Std.Dev.
3   Subject    (Intercept) 0.07106   0.2666
4   Residual                    0.15114   0.3888
5   Number of obs: 7923, groups: Subject, 59

6 Fixed effects:
7             Estimate Std. Error t value
8 (Intercept) 7.062632   0.038415 183.85
9 OrthLength  0.047717   0.002756 17.31
```

- **Efeitos Fixos:** *intercept* e *slope*

- **Efeitos aleatórios:**

- variância (e DP) no intercept por participante:
68% dos *intercepts* estão ± 1 DP
95% dos *intercepts* estão $\pm 1,96$ DP
- variação (e DP) dos resíduos (variância não explicada pelo modelo). DP dos resíduos no modelo sem efeitos mistos foi 0,46 – com efeitos mistos há menos variância dos resíduos

- **E os valores de p?**

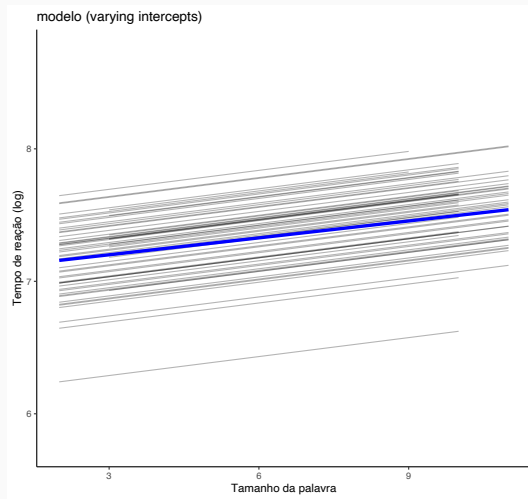
- Podemos 'forçar' um valor de p (com o pacote `lmerTest` ou com a função `tab_model` do pacote `sjPlot`)
- Podemos ignorar os valores de p e reportar intervalos de confiança (com `confint`)

Modelo com intercepts aleatórios

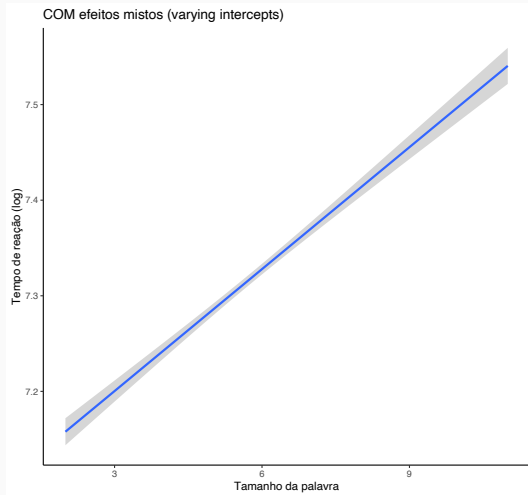
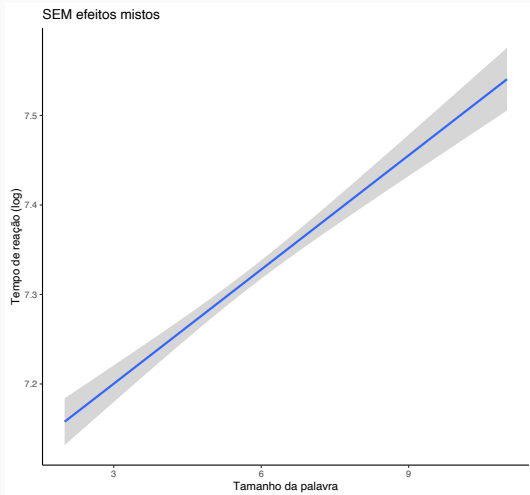
```
1 library(lmerTest)
2 mod2 = lmer(LogRT ~ OrthLength + (1 | Subject), data = readers)
3 summary(mod2)
4 Random effects:
5   Groups   Name                Variance Std.Dev.
6   Subject  (Intercept) 0.07106   0.2666
7   Residual                    0.15114   0.3888
8 Number of obs: 7923, groups:  Subject, 59
9 Fixed effects:
10              Estimate Std. Error      df t value Pr(>|t|)
11 (Intercept) 7.063e+00  3.841e-02 8.393e+01 183.85  <2e-16 ***
12 OrthLength  4.772e-02  2.756e-03 7.864e+03  17.31  <2e-16 ***
```

```
1 confint(mod2)
2              2.5 %      97.5 %
3 .sig01      0.22204436 0.32112631
4 .sigma      0.38274676 0.39489956
5 (Intercept) 6.98701241 7.13823642
6 OrthLength  0.04231222 0.05311755
```

Modelo com intercepts aleatórios



Modelo com intercepts aleatórios



Modelo com intercepts e slopes aleatórios

- Como próximo passo, podemos informar ao modelo que participantes podem ter tanto valores de partida variáveis (*random/varying intercepts*) como também tendências de aumento/diminuição variáveis (*random/varying slopes*)
- Queremos permitir que o aumento/diminuição no tempo de reação em função do tamanho da palavra possa variar por participante
- Utilizamos para isso $(1 + \text{OrthLength} \mid \text{Subject})$, ou simplesmente $(\text{OrthLength} \mid \text{Subject})$

```
1 | mod3 = lmer(LogRT ~ OrthLength + (OrthLength | Subject) , data = readers)
2 | summary(mod3)
```

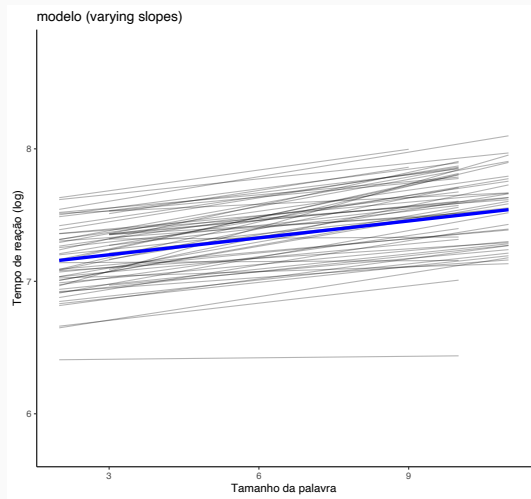
Modelo com intercepts e slopes aleatórios

```
1 Random effects:
2   Groups   Name      Variance Std.Dev. Corr
3   Subject  (Intercept) 0.0766184 0.27680
4             OrthLength 0.0007581 0.02753 -0.35
5   Residual                0.1493183 0.38642
6 Number of obs: 7923, groups: Subject, 59

7 Fixed effects:
8             Estimate Std. Error t value
9 (Intercept) 7.064822   0.039669   178.10
10 OrthLength  0.047468   0.004539    10.46
```

- **Efeitos aleatórios:** Além da variância (e DP) no *intercept* e nos resíduos, agora temos:
 - variância (e DP) no *slope* por participante
 - correlação prevista entre *intercepts* e *slopes*: o número negativo indica que *intercepts* mais altos tendem a ter *slopes* mais baixos (possível *ceiling effect*)

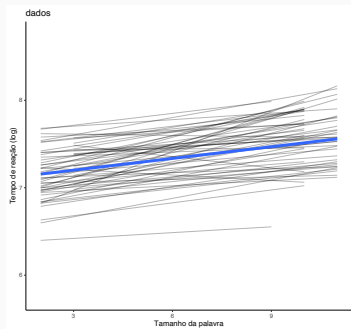
Modelo com intercepts e slopes aleatórios



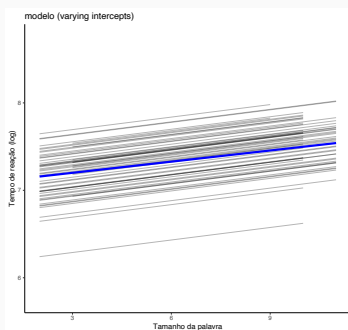
Compare

- Linha azul = tendência do grupo
- Linhas cinzas = dados/tendência de cada participante

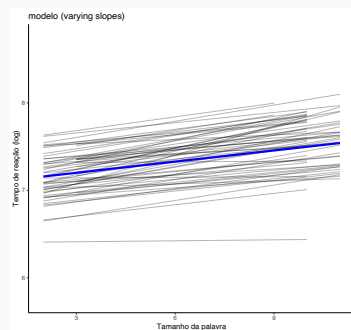
Dados



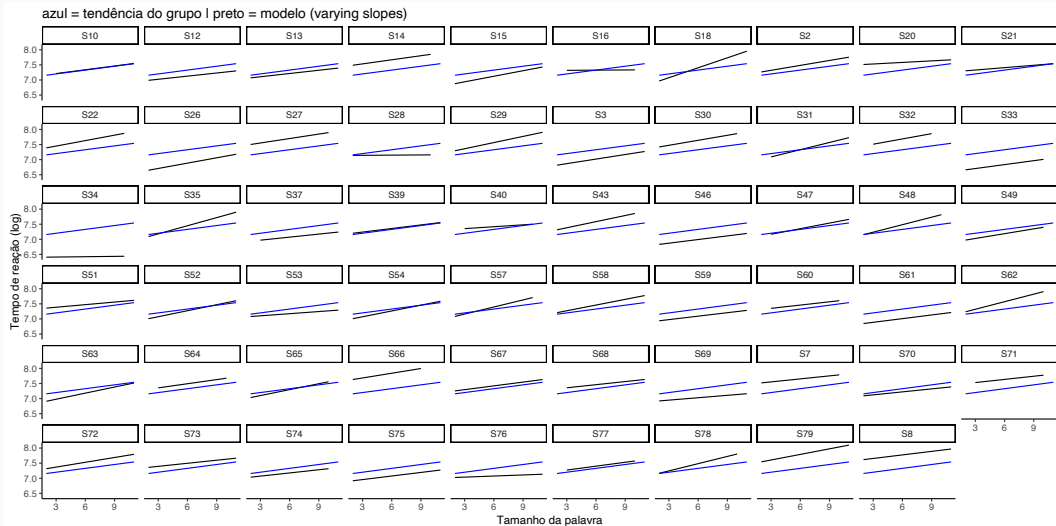
Intercepts aleatórios



Slopes aleatórios

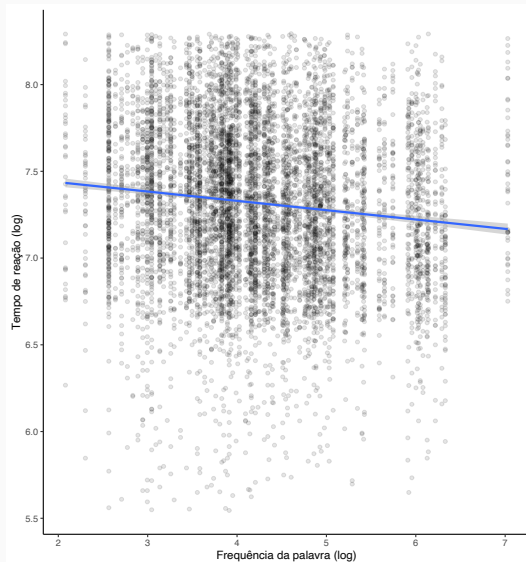
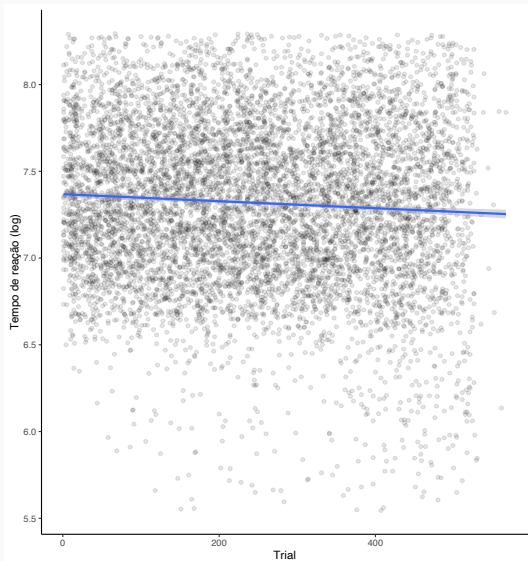


Intercepts e slopes variáveis

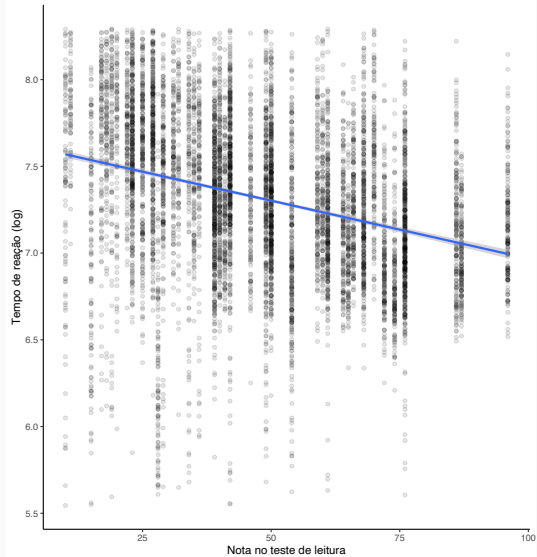
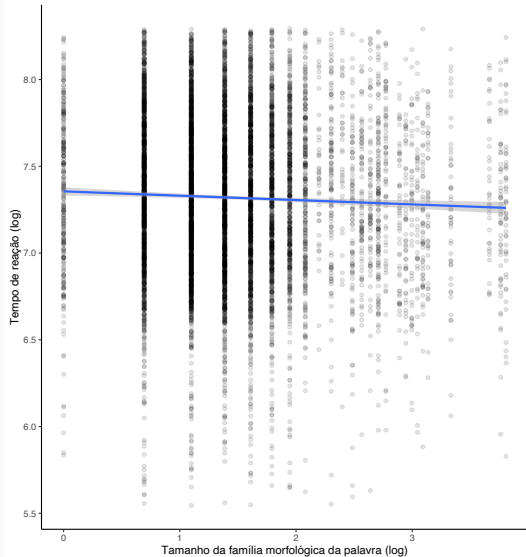


- Ainda temos as palavras como variável aleatória
- Há mais variáveis preditoras
- Precisamos verificar possíveis interações entre variáveis preditoras
- Começamos com uma análise exploratória das demais variáveis

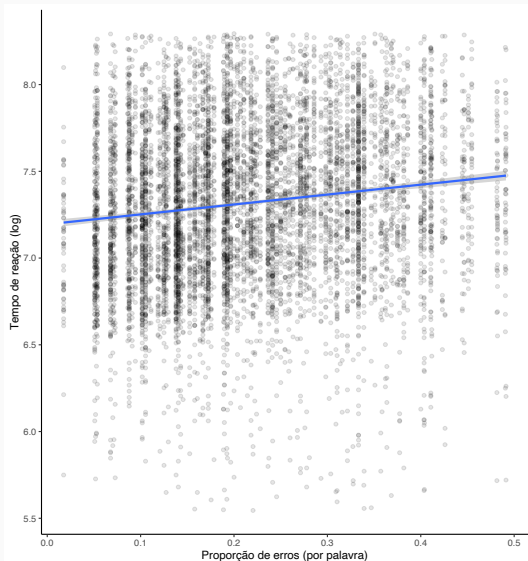
Modelo completo



Modelo completo



Modelo completo



→ Sem efeitos mistos

```
1 m0 = lm(LogRT ~ OrthLength + Trial +  
2       LogFrequency + LogFamilySize +  
3       ReadingScore + ProportionOfErrors,  
4       data = readers)
```

→ Com efeitos mistos

```
1 m1 = lmer(LogRT ~ OrthLength + Trial +  
2       LogFrequency + LogFamilySize +  
3       ReadingScore + ProportionOfErrors +  
4       (1 | Word) + (OrthLength | Subject),  
5       data = readers)
```

Modelo completo

→ LogFamilySize!

```
1 summary(m0)
2
3 Coefficients:
4             Estimate Pr(>|t|)
5 (Intercept)  7.580e+00 < 2e-16
6 OrthLength   3.545e-02 < 2e-16
7 Trial         -2.795e-04 < 2e-16
8 LogFrequency -2.691e-02 5.99e-06
9 LogFamilySize -1.566e-02 0.0204
10 ReadingScore -7.086e-03 < 2e-16
11 ProportionOfErrors 3.576e-01 5.92e-11
```

```
1 summary(m1)
2
3 Fixed effects:
4             Estimate Pr(>|t|)
5 (Intercept)  7.578e+00 < 2e-16
6 OrthLength   3.561e-02 2.99e-09
7 Trial         -2.381e-04 1.03e-15
8 LogFrequency -2.525e-02 0.00104
9 LogFamilySize -1.514e-02 0.08108
10 ReadingScore -7.618e-03 4.52e-07
11 ProportionOfErrors 3.958e-01 2.34e-08
```

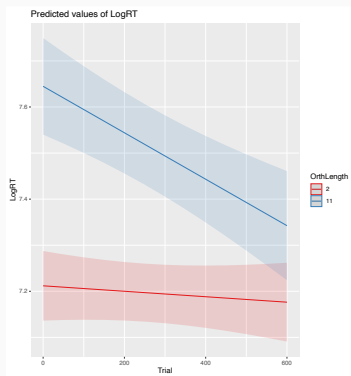
→ Verificando possíveis interações

```
1 m2 = lmer(LogRT ~ (Trial + OrthLength + LogFrequency + LogFamilySize +  
2           ReadingScore + ProportionOfErrors) ^ 2 +  
3           (1|Word) + (OrthLength|Subject),  
4           data = readers)
```

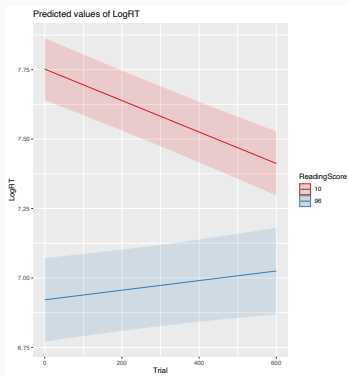
→ Mantendo interações significativas

```
1 m3 = lmer(LogRT ~ Trial + OrthLength + LogFrequency + LogFamilySize +  
2           ReadingScore + ProportionOfErrors +  
3           Trial:OrthLength + Trial:ReadingScore + Trial:ProportionOfErrors +  
4           (1|Word) + (OrthLength|Subject),  
5           data = readers)
```

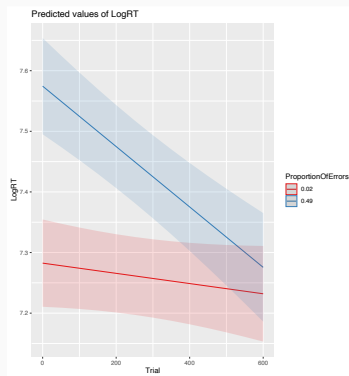

Trial:OrthLength



Trial:ReadingScore



Trial:ProportionOfErrors



Modelo completo

→ Trial!

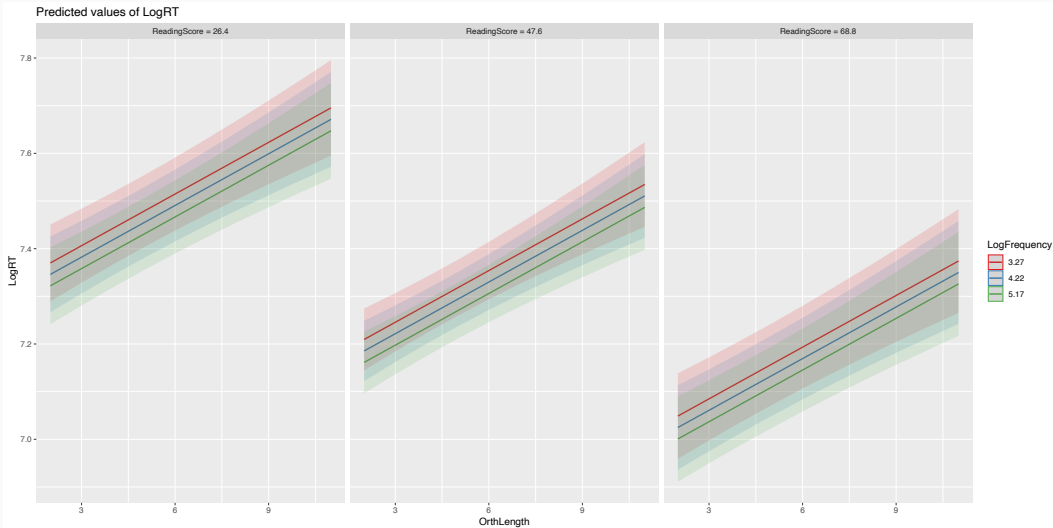
→ Sem interações

```
1 summary(m1)
2
3 Fixed effects:
4
5      Estimate  Pr(>|t|)
6 (Intercept)  7.578e+00 < 2e-16
7 OrthLength   3.561e-02 2.99e-09
8 Trial         -2.381e-04 1.03e-15
9 LogFrequency -2.525e-02 0.00104
10 LogFamilySize -1.514e-02 0.08108
11 ReadingScore  -7.618e-03 4.52e-07
12 ProportionOfErrors 3.958e-01 2.34e-08
```

→ Com interações

```
1 summary(m3)
2
3 Fixed effects:
4
5      Estimate  Pr(>|t|)
6 (Intercept)  7.555e+00 < 2e-16
7 Trial         -1.656e-04 0.21062
8 OrthLength   4.811e-02 6.99e-10
9 LogFrequency -2.529e-02 0.00100
10 LogFamilySize -1.483e-02 0.08685
11 ReadingScore  -9.659e-03 1.79e-09
12 ProportionOfErrors 6.212e-01 7.19e-10
```

Modelo completo



Recomendações

Recomendações finais

- Modelos mistos devem ser o padrão em estudos com medidas repetidas
- Comece com *intercepts* variáveis e depois adicione *slopes* variáveis também (de acordo / coerente com o desenho experimental)
- Nem sempre o modelo vai convergir (principalmente se mais complexo, com *slopes* aleatórios)
 - Reescalonar (*scale*) e centralizar as variáveis ajuda (tópico futuro)
 - Winter (2021) oferece muitas sugestões nesses casos
- É importante saber o que são os efeitos aleatórios e o que querem dizer seus coeficientes; contudo, normalmente continuamos tendo maior interesse de pesquisa pelos efeitos fixos

Perguntas?