REVISTA DA **ABRALIN**

ENSAIO TEÓRICO

Uma comparação entre ANOVA e modelos lineares mistos para análise de dados de tempo de resposta

Mahayana Cristina GODOY
Universidade Federal do Rio Grande do Norte (UFRN)

Marcus Alexandre NUNES
Universidade Federal do Rio Grande do Norte (UFRN)

RESUMO

Neste artigo, argumentamos que modelos lineares mistos (MLMs) são mais adequados que Análise de Variância (ANOVA) para lidar com dados de tempo de reação. Com a análise de um conjunto de dados simulados, mostramos que MLMs têm menor chance de incorrer em erros do Tipo I por permitir a inclusão de mais de um efeito aleatório (geralmente participantes e itens) em um único modelo. Também apresentamos uma introdução à implementação e análise de dados por meio de MLMs usando R e sugerimos materiais adicionais para os pesquisadores que desejarem fazer esse tipo de análise. Nosso principal objetivo é fomentar o uso de MLMs na comunidade psicolinguística brasileira.

ABSTRACT

In this paper, we argue that linear mixed models (LMMs) are more appropriate than Analysis of Variance (ANOVA) for the treatment of reaction time data. We analyze simulated data to show that LMMs decrease the chance of Type I errors by allowing the inclusion of more than one random effect (usually participants and items) within a single model. We also provide an introduction to the implementation and the data analysis of LMMs in R and suggest additional materials for researchers who want to start using these models. Our main goal is to encourage the use of LMMs amongst Brazilian psycholinguists.

OPEN ACCESS

EDITADO POR
- Raquel Freitag (UFS)

AVALIADO POR
- Livia Oushiro (UNICAMP)
- Guilherme Duarte Garcia (BSU)

SOBRE OS AUTORES
- Mahayana Cristina Godoy
  Conceptualização, Curadoria de Dados, Análise Formal, Metodologia, Administração do Projeto, Recursos, Supervisão, Validação, Visualização, Escrita – rascunho original, Escrita – análise e edição.
- Marcus Alexandre Nunes
  Conceptualização, Curadoria de Dados, Análise Formal, Investigação, Metodologia, Recursos, Software, Validação, Visualização, Escrita – análise e edição.

DATAS
– Recebido: 18/02/2020
– Aceito: 02/04/2020
– Publicado: 18/07/2020

COMO CITAR
Godoy, M. C.; Nunes, M. A. (2020). Uma comparação entre ANOVA e modelos lineares mistos para análise de dados de tempo de resposta. *Revista da Abralin*, v. 19, n. 1, p. 1-23, 2020.

PALAVRAS-CHAVE
Modelos lineares mistos. ANOVA. Tempo de resposta. Psicolinguística

KEYWORDS
Linear mixed models. ANOVA. Reaction time. Psycholinguistics

Introdução

A pesquisa experimental em psicolinguística depende da análise de dados de experimentos que, frequentemente, têm algumas características comuns. Em primeiro lugar, o estudo recruta um conjunto de participantes, geralmente alunos universitários. Além disso, coleta-se o julgamento ou o tempo de resposta a itens linguísticos criados pelos experimentadores de acordo com as variáveis a serem testadas. No caso de dados resultantes de tempos de resposta - caso de que trataremos - é comum que o pesquisador conduza sua análise por meio de Análise de Variância.

Neste artigo, seguimos o exposto em Baayen, Davidson e Bates (2008) e argumentamos que uma análise por meio de modelos lineares mistos é mais adequada para lidar com os dados de medidas repetidas que são obtidos em experimentos psicolinguísticos. Nosso objetivo principal é apresentar as vantagens desse tipo de análise frente a uma Análise de Variância para que a prática se torne mais comum no país. Além disso, apresentamos um passo-a-passo básico de como se implementa um modelo linear misto por meio da linguagem R e como se interpretam seus resultados. A escolha pela linguagem R, aqui, não se deve a mera casualidade, mas segue a necessidade crescente de disponibilizar códigos para construir um ambiente acadêmico transparente quanto à coleta e análise de dados. Por fim, recomendamos outros tutoriais e recursos disponíveis em português que podem auxiliar o pesquisador brasileiro que queira começar a analisar seus dados por meio de modelos lineares mistos.

1. Psicolinguística: medidas repetidas e generalização

Para ilustrar as possibilidades de análise estatística de dados de tempo de resposta, consideremos um experimento simples de *priming* semântico. Imaginemos que um linguista selecionou 60 palavras alvo, que foram apresentadas a 20 participantes[1]. Esses itens foram apresentados após uma palavra semanticamente relacionada (e.g. leite > CAFÉ) ou não relacionada (pente > CAFÉ). O pesquisador queria testar se a apresentação de uma palavra semanticamente relacionada antes da palavra alvo

1 O número de participantes em um experimento de *priming* costuma ser maior, mas simplificamos aqui para fins de ilustração.

facilitaria seu reconhecimento, e, por isso, registrou quanto tempo os participantes levavam para decidir se a palavra apresentada era ou não uma palavra do português brasileiro. Ao final, o pesquisador registrou que a diferença entre as médias de tempo de reação para palavras em cada uma das duas condições foi de 68,90ms (690,58ms para a condição relacionada, 759,48ms para a condição não-relacionada). Nesse cenário, o linguista deve ponderar o que ele espera que os resultados desse experimento informem dadas as condições de obtenção dos dados.

É nesse sentido que Raaijmakers (2003) destaca que a diferença entre médias de dois grupos em um contexto experimental não deve ser tomada pelo seu valor absoluto, mas precisa levar em conta se essa distinção permanecerá caso o experimento seja repetido. Portanto, um modelo estatístico deve considerar as variáveis que podem influenciar o resultado em uma possível replicação. Um dos fatores que pode influenciar os tempos de resposta é justamente a variável manipulada. A variável experimental preditora (ou independente) é um fator cujos níveis de variação são fixos (e.g., há dois níveis: relacionados e não relacionados), determinados pelo pesquisador e repetidos (esses níveis podem ser repetidos para cada novo item). Essa variável representa aquilo que chamamos de efeito fixo. No entanto, há ainda uma série de fatores - chamados efeitos aleatórios - que inserem variabilidade no conjunto de dados sem que sejam controlados pelo pesquisador. No experimento em tela, dois desses efeitos aleatórios são os itens experimentais e os participantes.

É razoável assumir que características intrínsecas a cada participante influenciam suas respostas de maneiras distintas. Como cada participante é diferente, alguns serão mais rápidos, outros mais lentos, alguns estarão desconfortáveis com a tarefa, outros já terão participado de outros experimentos e estarão mais à vontade, e todos terão características próprias no modo como construíram seu vocabulário e como usam a língua no dia a dia. Além disso, os itens do experimento - no caso, as palavras selecionadas - também são diferentes entre si. Por mais que os estímulos experimentais sejam controlados, cada um tem características próprias que interferem nos tempos de resposta. Isso significa que uma replicação do experimento com outros itens e outros participantes pode levar a resultados diferentes. Por fim, há ainda variabilidade decorrente das interações participante x condição e item x condição. É possível que o efeito da variável testada seja maior para um ou outro indivíduo ou item específico, o que contribui para a distribuição de dados.

Tanto no caso dos participantes quanto dos itens experimentais temos apenas um pequeno conjunto do fenômeno que queremos testar: os participantes são uma amostra dos falantes de português, e as palavras selecionadas são apenas uma amostra da língua portuguesa. Ao conduzir um experimento e aplicar modelos estatísticos para analisar seus dados, o pesquisador certamente imagina que seus resultados possam ser interpretados como representativos do processamento da linguagem na população, embora o experimento tenha sido aplicado em apenas duas dezenas de participantes. Além disso, o pesquisador também espera que qualquer efeito identificado através de análise quantitativa diga respeito ao fenômeno linguístico estudado, ou seja, ao fenômeno do *priming* semântico. Isso significa que o pesquisador quer ter certeza de que o efeito encontrado possa ser generalizado para a língua, não sendo mero reflexo do pequeno conjunto de 60 itens selecionados para a tarefa experimental.

A reflexão de que participantes e itens introduzem uma variabilidade aleatória no conjunto de dados coletados deve levar o experimentador a adotar modelos estatísticos que prevejam esse efeito. Por esse motivo, a análise de dados dessa natureza, na psicolinguística, geralmente se dá por meio de um de dois tipos de modelos lineares: uma Análise de Variância (ou ANOVA) ou um Modelo Linear Misto. Para argumentarmos por que acreditamos que a segunda opção é melhor para analisar dados de tempos de resposta com medidas repetidas, consideremos inicialmente como se faz esse tipo de análise a partir de uma ANOVA.

2. Modelo de Análise de Variância

Na psicolinguística, uma ANOVA é geralmente empregada para averiguar se, em dados coletados experimentalmente, há diferença em ao menos um par de médias entre as condições comparadas. Esse teste calcula uma estatística F com fins de verificar se a variância associada à manipulação de uma variável (no nosso caso, tipo de *prime*) é maior que a variância causada por fatores aleatórios não controlados pelo experimentador (para maiores detalhes sobre o cálculo das variâncias e da estatística F, cf. KUTNER *et al.*, 2004).

A fim de que a análise quantitativa dê conta das variações causadas por participantes e itens específicos, é preciso que esses dois elementos sejam tratados como variáveis que introduzem uma variação aleatória em um modelo estatístico. Como lembram Baayen, Davidson e Bates (2008), a psicolinguística tem adotado a proposta de Clark (1973), que consiste no cálculo de uma estatística quasi-F a partir do agrupamento de dados por participante (F_1) e por item (F_2). Para o cálculo de F_1, as observações de cada participante são agregadas para produzir uma média para cada condição. Para o cálculo de F_2, as observações de cada item são agregadas também para produzir uma média para cada condição. Devido à dificuldade do cálculo de quasi-F, Clark (1973) propõe que um valor mínimo de quasi-F, chamado minF', seja obtido a partir de F_1 e F_2 através da equação $\frac{F_1 F_2}{F_1 + F_2}$.

A crítica de Clark (1973) foi incorporada aos modos de análise da psicolinguística, uma vez que o cálculo de F_1 e F_2 é prática corrente até os dias atuais e refletem uma preocupação em lidar com itens e participantes como efeitos aleatórios em um modelo estatístico. No entanto, Raaijmakers, Schrijnemakers e Gremmen (1999) reportam uma mudança no modo como a análise é feita para o cálculo da estatística do teste: embora a maioria dos artigos publicados no *Journal of Memory and Language* reportasse também o valor de minF' em conjunto com F_1 e F_2 na década de 70, vinte anos depois quase nenhum artigo o fazia. Segundo os autores, ao longo do tempo, F_1 e F_2 deixaram de ser vistos como um passo intermediário para o cálculo de minF', e foram reinterpretados como testes realizados para generalizar os resultados, respectivamente, por participantes e por itens. Raaijmakers (2003) sugere que um dos motivos para essa mudança foi o fato de que minF' pode não ser significativo mesmo quando F_1 e F_2 o são. Isso, aliado à crítica de alguns autores de que o uso de

minF' seria um teste muito conservador (SMITH, 1976; WIKE; CHURCH, 1976; *apud* RAAIJMAKERS, 2003) fez com que a prática padrão da área se limitasse ao cálculo de F_1 e F_2.

Há uma discussão sobre quão conservador de fato é o uso de minF', com alguns autores defendendo que o teste fornece uma boa aproximação da estatística F (DAVENPORT e WEBSTER, 1973; *apud* RAAIJMAKERS, 2003). Porém, aqui chamamos atenção para outras consequências do uso da ANOVA para análise de dados de experimentos psicolinguísticos. Ao lançar mão de um valor de F_1 e outro de F_2, esses testes se baseiam em duas análises distintas para derivar um único resultado, e nenhuma delas engloba, em um único modelo, todos os efeitos aleatórios que influenciam o experimento. A própria prática de calcular dois valores de F advém do fato de que uma ANOVA pode considerar apenas um efeito aleatório por vez.

Há pelo menos dois problemas que derivam dessa prática. Em primeiro lugar, a realização de dois testes aumenta as taxas de falsos positivos, também conhecidos como erros do tipo I, principalmente quando os p-valores não são corrigidos (KUTNER *et al.*, 2004). Além disso, há situações em que apenas uma das análises se mostra estatisticamente significativa. Embora haja entendimento de que um resultado só é considerado estatisticamente significativo quando os dois valores de F estão abaixo do nível de α assumido, isso não impede que alguns trabalhos reportem os resultados como significativos mesmo quando apenas um dos valores cumpre esse critério.

Se ainda não houvesse possibilidade de inclusão de itens e participantes como efeitos aleatórios em um mesmo modelo, uma ANOVA de fato se mostraria como melhor alternativa para análise dos dados. Porém, novas ferramentas para análise estatística foram desenvolvidas desde a publicação de Clark (1973), em grande parte graças ao maior poder computacional disponível atualmente. Hoje, a inclusão de itens e participantes como efeitos aleatórios cruzados em um mesmo modelo é possível em uma classe de modelos lineares chamados modelos lineares mistos (MLMs). Esse tipo de análise tem se tornado popular para a análise de dados em pesquisa linguística desde a publicação de Baayen, Davidson e Bates (2008), mas ainda é subutilizada na comunidade linguística brasileira (algumas exceções são COSTA, 2013, e GODOY *et al.*, 2017). Na próxima seção, detalharemos as vantagens do um modelo misto e suas semelhanças e diferenças em relação à ANOVA. Para tanto, utilizaremos um conjunto de dados simulados para o experimento de *priming* descrito na primeira seção[2].

3. O conjunto de dados analisado

Os dados analisados neste trabalho foram gerados de maneira aleatória, utilizando o método proposto por Debruine e Barr (2019). Muito comuns da literatura estatística, as simulações de Monte

2 O dados, bem como os códigos para análise do experimento por meio de uma ANOVA e um modelo linear misto, estão disponíveis em https://osf.io/efxt4.

Carlo são utilizadas para avaliar o desempenho de modelos estatísticos em situações computacionais controladas. Estas situações replicam o comportamento real dos dados, permitindo que estudos em larga escala sejam realizados sem a necessidade da coleta de dados reais. Tais métodos são utilizados desde casos mais simples, como a estimativa das probabilidades de uma roleta (MORETTIN; BUSSAB, 2004), até aplicações complexas de estatística, como a utilização de séries temporais para modelagem de sequências de DNA (LOPES e NUNES, 2006).

O conjunto de dados utilizado neste trabalho foi gerado a partir de uma simulação de Monte Carlo. Em primeiro lugar, definimos o número de sujeitos (20) e itens (60). Como nosso objetivo neste trabalho é mostrar que os resultados de uma ANOVA, tal qual realizada rotineiramente, podem levar a aumento da detecção de falsos positivos, foram definidas médias iguais para as condições não-relacionada e relacionada do *prime*. Note que, por ser um processo de Monte Carlo, as estimações pontuais dessas médias podem divergir, mas elas não são estatisticamente diferentes. Desta forma, estamos condicionando os dados a não possuírem diferença entre os níveis da variável *prime*.

Há dois tipos principais de eventos na natureza: eventos determinísticos e eventos estocásticos. Eventos determinístico são aqueles em que, mantidas as condições iniciais, o resultado final será sempre o mesmo, não importando quantas vezes forem repetidos. Por exemplo, o tempo que uma esfera de 1kg leva para cair de uma altura de 20m será sempre o mesmo, desde que este evento seja repetido sob exatamente as mesmas condições, como presença ou ausência de vácuo, temperatura do local do experimento, pressão atmosférica, valor da aceleração gravitacional e tudo o mais que pode ser controlado pelo experimentador.

Eventos estocásticos, por outro lado, variam seu resultado mesmo quando as condições iniciais são mantidas. O evento estocástico mais simples possível é o lançamento de uma moeda. Embora saibamos que os únicos resultados possíveis são cara ou coroa, o resultado dos lançamentos individuais de uma moeda são imprevisíveis. Entretanto, é possível modelar com alguma incerteza o comportamento do lançamento de um número grande de moedas.

Graças à participação de seres humanos, experimentos de psicolinguística são estocásticos. Desta forma, inserimos uma variabilidade geral no modelo, pois estamos tratando com dados estocásticos. Além da variabilidade geral do modelo, inserimos variabilidades extras por item e por participante para simular a variabilidade introduzida por esses termos na distribuição dos dados. Feitos esses ajustes, obtivemos uma distribuição normal dos dados de tempo de resposta. Reconhecemos que, em geral, dados de tempo de reação tem uma distribuição com cauda longa à direita e, com frequência, passam por uma transformação logarítmica para que possam ser analisados (GODOY, 2019). Aqui optamos por criar uma distribuição normal para que o exemplo fosse mais simples e não tirasse o foco do objetivo do artigo, que é comparar o desempenho de MLMs e ANOVAs na análise de dados de experimentação linguística.

Antes de seguirmos para a análise dos dados por meio de um MLM, analisamos o mesmo conjunto com uma ANOVA que toma o tipo de *prime* como variável preditora. Não nos aprofundaremos em explicar os passos dessa ANOVA por acreditarmos que esse tipo de análise é conhecido pela comunidade psicolinguística do Brasil a despeito do *software* utilizado. A Tabela 1 resume os

resultados de F_1 e F_2, que sugerem um efeito significativo[3] de *prime* por participante e por item (F1(1, 38) = 0,0301; F2(1, 58) = 0,0449). Dito de outro modo, uma ANOVA encontrou efeito de *prime* no conjunto de dados analisado ao fazer uma análise considerando a variabilidade de itens e de participantes separadamente. Contudo, como veremos a seguir, ao inserirmos esses dois termos em um único modelo linear, o resultado é diferente.

Teste	DF	SS	MS	F	p-valor
F1	1	39109	39109	5,0789	0,0301
	38	292609	7700	-	-
F2	1	58664	58664	4,2011	0,0449
	58	809907	13964	-	-

TABELA 1 – Resultados de F_1 e F_2 para os dados analisados.
Fonte: Elaborada pelos autores.

4. Modelos Lineares Mistos

Modelos Lineares Mistos (MLMs) são uma classe de modelos estatísticos que recebem esse nome por especificarem, em sua equação, dois tipos de efeitos: efeitos fixos e aleatórios. Para entender melhor a construção de um MLM, consideremos o exemplo de experimento apresentado. Segundo a hipótese levantada, há uma variável - *prime* - que pode influenciar o tempo de resposta dos participantes. Representamos isso formalmente com a equação abaixo, que pode ser lida como "tempo varia em função de *prime*".

$$\text{tempo} \sim \textit{prime}$$

Na Figura 1, vemos a distribuição dos valores de tempo de resposta para cada uma das condições de *prime* de um dos participantes do experimento. A reta ajustada parte do valor médio da condição relacionada (690,58ms) para o valor médio da condição não-relacionada (759,48ms).

3 Ao longo deste artigo, assumimos a posição da estatística frequentista em adotar um nível de significância (aqui determinado em 0,05) como ponto de referência para rejeitarmos uma hipótese nula. Essa visão é a mais comum em pesquisas experimentais, mas não é a única possível para realização de análises estatísticas e é criticada inclusive por pesquisadores que atuam em pesquisa linguística (GARCIA, 2019). Foge ao escopo deste trabalho discutirmos os problemas envolvidos no uso dicotômico de p-valores. Para essa discussão, veja Kruschke e Liddell (2018). Para uma discussão mais detalhada sobre o que significa exatamente um p-valor em análises frequentistas, cf. Winter (2019, Capítulos 9 e 10).

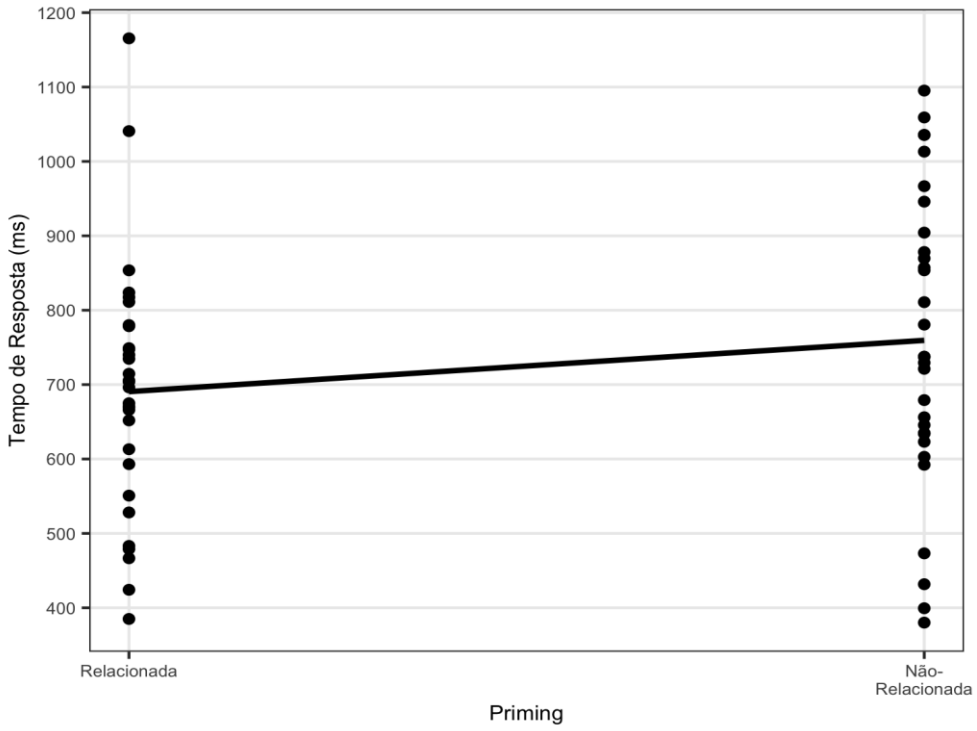

FIGURA 1 - Distribuição do tempo de resposta (ms) para as condições relacionada (`rel`) e não-relacionada (`nrel`) do prime.
Fonte: Elaborada pelos autores.

```
##
## Call:
## lm(formula = tempo ~ prime, data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -379.40 -124.78   -4.98  112.35  474.87
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   690.58      33.75  20.460   <2e-16 ***
## primenrel      68.90      47.73   1.443    0.154
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 184.9 on 58 degrees of freedom
## Multiple R-squared:  0.03468,    Adjusted R-squared:  0.01803
## F-statistic: 2.083 on 1 and 58 DF,  p-value: 0.1543
```

Os valores estimados para o intercepto e o *slope* – ou α e β – estão no *output* acima, na tabela de coeficientes (*Coefficients*, em inglês). O valor de α é 690,58 e de β é 68,90. Desta forma, a equação do modelo linear deste exemplo pode ser resumida em (2).

$$\text{tempo} = 690{,}58 + 68{,}90 \times \textit{prime} \qquad (2)$$

Para entender como obtivemos os valores de α e β reportadas na tabela de coeficientes, vamos assumir que o *prime* relacionado é nosso nível de referência[4]. Assim, a variável *prime* assume dois níveis: 0 para *prime* relacionado e 1 para *prime* não-relacionado. Desta forma, podemos resolver a equação (2) substituindo os valores de *prime* por seus equivalentes numéricos como em (3):

$$\begin{array}{rl} \textit{prime} \text{ relacionado} & : 690{,}58 + 68{,}90 \times 0 = 690{,}58 \\ \textit{prime} \text{ não-relacionado} & : 690{,}58 + 68{,}90 \times 1 = 759{,}48 \end{array} \qquad (3)$$

A média de tempo de resposta para a condição de *prime* relacionado intercepta o eixo y no valor 690,58. Dizemos que esse é o valor do nosso intercepto para um modelo linear, e a partir desse valor de referência, podemos ver a variação do outro nível da condição *prime*. Para a condição não-relacionada, a média foi de 759,48; portanto, dizemos que essa condição apresenta um coeficiente β de 68,90 em comparação ao valor do intercepto, o que resulta na média calculada (*i.e.*, 690,58 + 68,90).

No entanto, conforme já indicamos, os tempos de resposta também estão sujeitos à variabilidade dos participantes. Dizemos que temos medidas repetidas por participante porque cada um deles contribuiu com 60 medidas de tempos de resposta, e, independentemente do tipo de condição, suas médias são bem diferentes e refletem idiossincrasias de cada um. Na Figura 1 vemos o comportamento das respostas de um dos participantes, enquanto na Figura 2 podemos comparar o desempenho de todos eles simultaneamente. Enquanto alguns participantes tiveram tempos de respostas mais rápidos, outros foram mais lentos. Enquanto alguns participantes foram mais rápidos nas respostas para a condição relacionada, outros participantes foram mais rápidos na execução da tarefa não-relacionada. Além disso, a inclinação da reta - ou seja, o *slope* - é diferente para cada um deles.

4 Em geral, o R utiliza ordem alfabética para determinar o nível de referência. Isso faz com que, por default, o nível de referência seja não-relacionado (nrel). Para fins didáticos, acreditamos que seria mais interessante ter como nível de referência o nível que nos interessa mais: o de prime relacionado. A mudança foi feita por meio do código priming$prime <- factor(priming$prime, levels = c("rel", "nrel")), que pode ser consultado em nossos materiais suplementares.

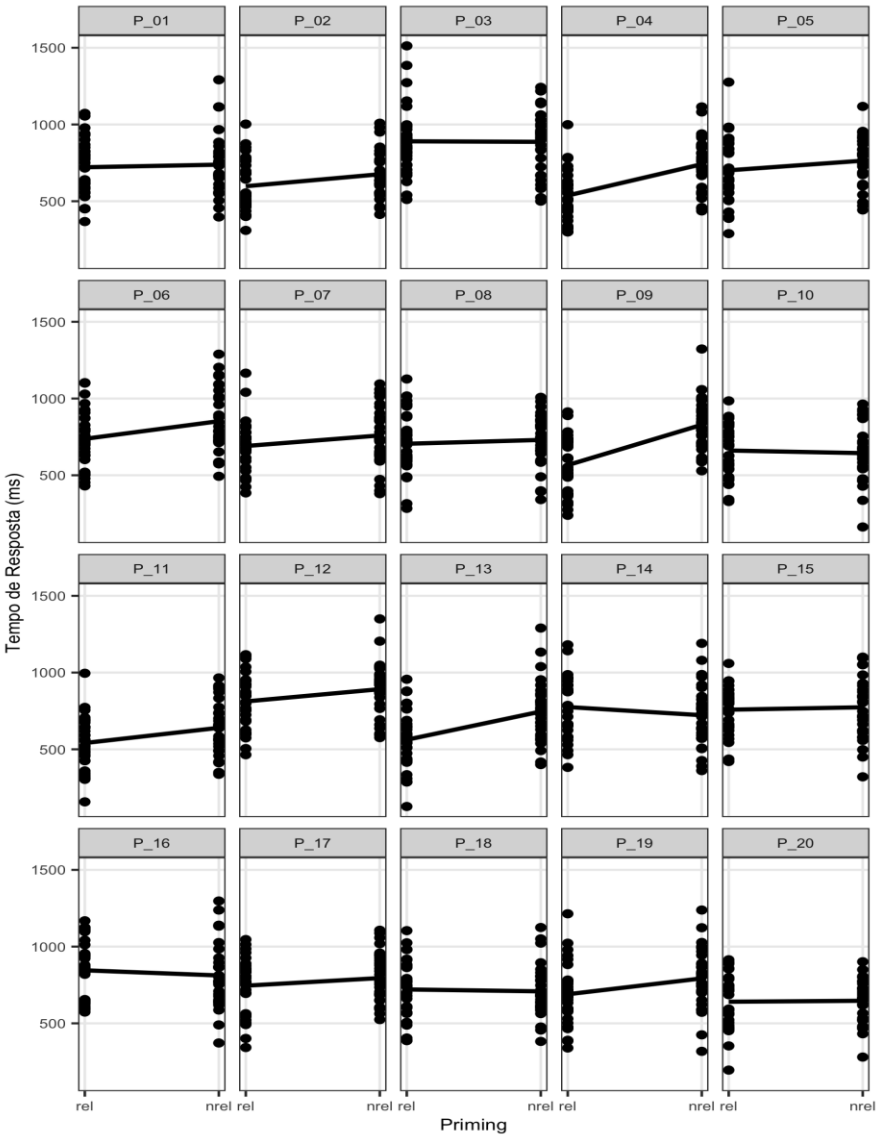

FIGURA 2: Comparação do tempo de resposta (ms) para as condições relacionada e não-relacionada do *prime* para todos os participantes do experimento.
Fonte: Elaborada pelos autores.

Ao construirmos um modelo estatístico que tente explicar como a condição *prime* afeta o tempo de resposta, seria adequado que esse modelo levasse em conta a variabilidade de cada participante. Ao utilizar o modelo ANOVA proposto por Clark (1973) e calcular as médias por item e participante, o pesquisador não considera estas variabilidades e acaba enfraquecendo a sua modelagem, descartando informações importantes sobre a variabilidade dos dados coletados.

Usando a sintaxe do R, o efeito aleatório do participante pode ser inserido no modelo da seguinte maneira:

```
tempo ~ prime + (1|participante)
```

O que a notação acima indica é que o modelo deve considerar interceptos diferentes para cada participante (o número 1, aqui, representa o intercepto). Como modela um tipo de variabilidade aleatória, esse intercepto é chamado de intercepto aleatório. Considerando ainda que nossos dados têm outra fonte de variável aleatória – os itens – acrescentamos mais um termo a nosso modelo para que ele leve em conta esse efeito: interceptos aleatórios por item.

```
tempo ~ prime + (1|participante) + (1|item)
```

Por fim, a Figura 2 nos mostrou também que o declive da reta entre as duas condições difere para cada um dos participantes, e podemos assumir que isso também seja verdade para todos os 60 itens experimentais. Para considerar as diferenças dos declives – *slopes* – associados às condições experimentais, inserimos um novo termo no modelo chamado de *slope* aleatório, pode ser representado em nosso modelo da seguinte maneira:

```
tempo ~ prime + (1+prime|participante) + (1+prime|item)
```

A sintaxe `1 + prime` pode ser lida como "intercepto aleatório + *slope* aleatório por *prime*". De modo geral, essa equação se lê como "um modelo linear misto com tempo como variável resposta, *prime* como efeito fixo, interceptos aleatórios para participantes e itens e *slopes* aleatórios por *prime* para participantes e itens". A partir dessa fórmula, criamos um MLM que tem como efeito fixo a condição *prime*, mas que considera, ao mesmo tempo, variabilidade por item e por participante. Ainda que uma ANOVA se proponha a fazer o mesmo, a análise por MLM permite considerar, em um único modelo, mais de uma fonte de variabilidade aleatória.

Uma vez introduzidos os parâmetros de um modelo misto, vejamos como se dá sua implementação e análise de resultados no R.

5. Modelos Lineares Mistos com um Efeito Fixo

Faremos um modelo misto considerando o conjunto de dados que vemos abaixo e que contém 4 colunas: tempo, com o tempo de resposta dos participantes em milissegundos; e *prime*, com identificação das condições de *prime* relacionado (`rel`) e não-relacionado (`nrel`); participante, com identificação dos 20 participantes da pesquisa; e item, com identificação dos 60 itens experimentais.

```
##        tempo   prime  participante    item
## 1   397.3359    nrel          P_01 item_01
## 2   789.1405    nrel          P_01 item_02
## 3   651.0004    nrel          P_01 item_03
## 4   455.5234    nrel          P_01 item_04
## 5   586.7212    nrel          P_01 item_05
## 6  1113.3861    nrel          P_01 item_06
```

```
## 7   966.7943    nrel          P_01 item_07
## 8   734.0415    nrel          P_01 item_08
## 9   681.3087    nrel          P_01 item_09
## 10  859.2883    nrel          P_01 item_10
```

Com a função `lmer` do pacote `lme4`[5], construímos um modelo linear chamado `modelo.prime`. Os argumentos de `lmer` são, nessa ordem, a fórmula do MLM a ser ajustada – e que construímos acima – e o nome do conjunto de dados a partir do qual se deve construir o modelo. Como resultado, temos o seguinte código:

```
modelo.prime <- lmer(tempo ~ prime + (1+prime|participante) + (1+prime|item), data =
priming)

## Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, :
## unable to evaluate scaled gradient
```

Note que recebemos um aviso após rodar este ajuste, identificado por "*Warning in checkConv(attr(opt, "derivs"), opt$par, ctrl = control$checkConv, : unable to evaluate scaled gradient*". Este erro se refere à função gradiente. O gradiente é uma técnica utilizada para encontrar o valor máximo da função de verossimilhança do modelo. A função de verossimilhança é importante para a teoria Estatística porque o seu máximo determina as estimativas dos parâmetros do modelo ajustado aos dados. Portanto, é fundamental que a procura pelo máximo desta função convirja numericamente. Assim, temos a garantia de termos chegado em um modelo com estimativas corretas para os seus parâmetros.

Recebemos apenas um aviso no ajuste de nosso modelo. Outro aviso muito comum que pode surgir neste tipo de análise é identificado por "*Model failed to converge: degenerate Hessian with 1 negative eigenvalues*". Este aviso diz respeito à matriz Hessiana dos dados. Esta matriz determina a curvatura local da função que está sendo maximizada. Entretanto, ela necessita ser positiva semi-definida, ou seja, seus autovalores devem ser todos não-negativos. Somente assim é possível garantir que a estimativa do gradiente encontrou, de fato, um ponto de máximo global. De acordo com os resultados que obtivemos, com um autovalor negativo, não há garantia de que a estimativa encontrada para o modelo seja, de fato, o máximo procurado.

De maneira simplificada, isso indica que o modelo não foi ajustado a contento: as estimações numéricas realizadas não convergiram a um valor final referente a uma máxima global. Ou seja, os resultados obtidos pelo modelo não são confiáveis. Como a função `lmer` é baseada em um algoritmo iterativo, não há a garantia de que uma resposta sempre será encontrada, devido a problemas de convergência numérica. Infelizmente, esta característica surge devido a limitações computacionais que, até o momento da publicação deste trabalho, ainda não haviam sido sanadas (BOLKER *et al.*, 2013; *apud* BATES *et al.*, 2015).

Para evitar problemas de convergência do algoritmo numérico iterativo, Barr *et al.* (2013) sugerem que comecemos a análise da maneira que iniciamos, com a estrutura mais complexa possível

5 Os códigos usados a partir desta seção estão disponíveis com os materiais suplementares do artigo.

para os efeitos aleatórios. Caso o modelo tenha problemas de convergência no método iterativo, devemos simplificar esta estrutura até atingir uma convergência sem problemas. Portanto, simplificaremos a estrutura dos efeitos aleatórios dos participantes, retirando o *slope* aleatório dos itens, mantendo apenas o intercepto aleatório para este efeito. Manteremos a estrutura dos participantes da maneira original.

```
modelo.prime <- lmer(tempo ~ prime + (1+prime|participante) + (1|item), data = priming)
```

Dessa vez, com as estruturas dos efeitos aleatórios para os participantes mantida como no modelo mais complexo, mas simplificando a estrutura aleatória definida para os itens, conseguimos que o modelo ajustado convergisse sem problemas. Vemos os resultados desse modelo através do comando `summary`.

```
summary(modelo.prime)

## Linear mixed model fit by REML ['lmerMod']
## Formula: tempo ~ prime + (1 + prime | participante) + (1 | item)
##   Data: priming
##
## REML criterion at convergence: 15575.6
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.7733 -0.6842 -0.0081  0.6540  3.2723
##
## Random effects:
##  Groups       Name        Variance Std.Dev. Corr
##  item         (Intercept) 12906    113.60
##  participante (Intercept)  9145     95.63
##               primenrel    5367     73.26   -0.69
##  Residual                  21161    145.47
## Number of obs: 1200, groups:  item, 60; participante, 20
##
## Fixed effects:
##             Estimate Std. Error t value
## (Intercept)   695.01      30.38  22.880
## primenrel      62.54      34.63   1.806
##
## Correlation of Fixed Effects:
##           (Intr)
## primenrel -0.672
```

Há, em primeiro lugar, informações sobre os efeitos aleatórios (*random effects*). Nele, vemos a variabilidade dos dados que pode ser explicada pelos participantes e itens. Há, ainda, informação de variabilidade residual (*residuals*), que indica a variação que não pode ser explicada pelos parâmetros

explícitos no modelo. A tabela de efeitos fixos nos informa os efeitos da nossa variável independente. Vemos que há duas linhas: `(Intercept)` e `primenrel`. Isso significa que o modelo assumiu a condição relacionada como intercepto, e está comparando a condição relacionada com ela. Por esse motivo, o coeficiente estimado para *intercept* corresponde à média de tempo de resposta para as palavras de *prime* relacionado (695,01ms). O p-valor do *intercept* é pouco informativo nesse caso: ele assume a hipótese nula de que o valor do intercepto é igual a zero, o que não diz nada sob o efeito investigado, pois nunca se esperaria que um tempo de resposta foi de zero milissegundos.

A média para as palavras de *prime* relacionado foi 757,50ms. Isso é 62,54ms a mais que a média do intercepto, como evidencia o valor positivo do coeficiente para a linha `primenrel`. Essa linha, portanto, nos dá o valor do coeficiente estimado para a condição relacionada em comparação aos valores do intercepto, bem como as estatísticas associadas a esse coeficiente.

6. Comparação de modelos aninhados

Até o momento, vimos que um MLM consegue lidar com a variabilidade de itens e participantes em um único modelo, tornando desnecessária uma análise que o faça a partir da agregação de médias para cada um desses termos, como uma ANOVA. Após a construção do nosso modelo na Seção 5, analisamos seus coeficientes para explicarmos como devemos lê-los. No entanto, esse artifício foi apenas um recurso didático, pois em geral se parte para a análise de coeficientes de MLMs apenas quando se encontra o melhor modelo ajustado. Esse modelo é aquele que consegue explicar a distribuição dos dados de modo mais parcimonioso, i.e., com o menor número de efeitos fixos possíveis. Pra encontrar o melhor modelo ajustado aos dados, fazemos o que se chama de comparação de modelos aninhados para realizar um teste de razão de verossimilhança. Nas palavras de Winter (2013, p. 12), "verossimilhança é a probabilidade de observar seu conjunto de dados dado o seu modelo. A lógica do teste de razão de verossimilhança é comparar a verossimilhança de dois modelos entre si. Primeiro, o modelo sem o fator de interesse (...), e depois o modelo com o fator em que se está interessado". No caso do experimento que estamos analisando, perguntamo-nos se os dados têm maior probabilidade de serem observados dado um modelo que tenha *prime* como efeito fixo em comparação a um modelo que não o tenha. Fazemos isso ao compararmos `modelo.prime` (que repetimos abaixo para incluir o argumento REML[6]) e `modelo.nulo`. A função anova é então usada para realizar o teste de razão de verossimilhança e averiguar se os dois modelos explicam os dados de maneiras significativamente distintas. Como `modelo.nulo` está aninhado (i.e., contido) em

6 REML é a sigla, em inglês, para REstricted Maximum Likelihood. É uma técnica utilizada no ajuste de modelos lineares mistos para a estimação de parâmetros perturbadores. Os parâmetros perturbadores não são o objetivo final da análise, mas devem ser levados em conta na estimação dos parâmetros de interesse, como os coeficientes dos modelos lineares mistos. Exemplos de parâmetros perturbadores no contexto deste trabalho são as variabilidades por participante e por item dos experimentos de prime considerados.

`modelo.prime` exceto pela variável que estamos investigando, dizemos que esta é uma comparação de modelos aninhados.

```
modelo.nulo  <- lmer(tempo ~ 1 + (1+prime|participante) + (1|item),
                     data = priming,
                     REML = FALSE)
modelo.prime <- lmer(tempo ~ prime + (1+prime|participante) + (1|item),
                     data = priming,
                     REML = FALSE)

anova(modelo.nulo, modelo.prime)

## Data: priming
## Models:
## modelo.nulo: tempo ~ 1 + (1 + prime | participante) + (1 | item)
## modelo.prime: tempo ~ prime + (1 + prime | participante) + (1 | item)
##              Df   AIC   BIC  logLik deviance  Chisq Chi Df Pr(>Chisq)
## modelo.nulo   6 15608 15638 -7797.9    15596
## modelo.prime  7 15606 15642 -7796.3    15592 3.2545      1    0.07123 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Note que o próprio *output* explica quais foram os modelos considerados nesta análise. Para que os modelos sejam comparáveis, eles devem estar aninhados: `modelo.nulo` indica que este MLM não possui efeito de *prime*, enquanto `modelo.prime` possui este efeito.

Além disso, cada coluna da tabela resposta neste *output* possui seu próprio significado. `Df` indica o número de graus de liberdade de cada modelo. As colunas AIC e BIC são estatísticas chamadas *Akaike Information Criterion* e *Bayaesian Information Criterion*, respectivamente. Elas servem como auxiliares na escolha do melhor modelo para os dados analisados. Não existe um valor de referência para estas estatísticas. Vamos escolher o modelo que possua o menor valor de AIC ou BIC. Entretanto, esta decisão precisa ser tomada a partir de um Teste de Razão de Verossimilhanças, sobre o qual comentamos no próximo parágrafo. A coluna `logLik` reporta o logaritmo da máxima verossimilhança dos modelos ajustados. Este valor é utilizado no cálculo de AIC e BIC.

Para determinar se os valores de AIC e BIC reportados são estatisticamente diferentes, o indicado, estamos interessados em testar as hipóteses

$$\begin{aligned} H_0 &: \theta = 0 \\ H_1 &: \theta \neq 0, \end{aligned}$$

em que θ é o efeito procurado na análise. Em nosso caso, θ é o parâmetro que representa o efeito de *prime*. Sob a hipótese nula (H_0), assumimos que este efeito é nulo, enquanto sob a hipótese alternativa (H_1) assumimos o complementar disso. De acordo com Kutner *et al.* (2004), o Teste de Razão de Verossimilhanças Λ é definido por

$$\Lambda = -2\ln\frac{\sup_{\theta\in\Theta_0}\mathcal{L}(\theta)}{\sup_{\theta\in\Theta_1}\mathcal{L}(\theta)}, (4)$$

em que $\sup_{\theta\in\Theta_0}\mathcal{L}(\theta)$ indica o máximo do logaritmo da função de verossimilhança para os valores de θ sob a hipótese H_0 e $\sup_{\theta\in\Theta_1}\mathcal{L}(\theta)$ indica o máximo do logaritmo da função de verossimilhança para os valores de θ sob a hipótese H_1. Sob a H_0, a estatística Λ possui distribuição χ^2 com $p_1 - p_0$ graus de liberdade, em que p_1 e p_0 são o número de parâmetros a serem estimados pelos modelos *prime* e nulo, respectivamente.

As últimas três colunas deste *output* indicam o resultado do Teste de Razão de Verossimilhanças aplicado nos dois modelos. A coluna `Chisq` informa o valor da estatística calculada pela equação (4), enquanto `Chi Df` indica quantos graus de liberdade esta estatística possui. A coluna `Pr(>Chisq)` reporta o p-valor do teste aplicado.

Como o p-valor reportado (0,0712) está acima da taxa de erros do Tipo I considerada (em nosso caso, $\alpha = 0{,}05$), a comparação indica que não houve diferença na verossimilhança entre os modelos. Nesse caso, nosso pesquisador fictício deveria reportar os resultados com um texto como

"Ajustamos um Modelo Linear Misto com *prime* como variável preditora, interceptos aleatórios para participantes e itens e *slope* aleatório por *prime* para item. Uma comparação com modelos aninhados indicou que *prime* não contribui significativamente para o modelo ($\chi^2 = 3{,}2545$, p-valor = 0,0712)".

Supondo que o p-valor da comparação entre modelos fosse significativo, o indicado seria dizer que o melhor modelo ajustado foi o modelo com *prime* como efeito fixo (além de todos os efeitos aleatórios já citados). Além de reportar a estatística da comparação que atestaria essa significância, o pesquisador poderia ainda apresentar a tabela com os coeficientes do melhor modelo ajustado, indicando o β, o erro padrão (*Std. Error*) e o valor-t associado ao *prime* não-relacionado[7].

O objetivo deste artigo é apresentar uma alternativa ao uso da ANOVA nos casos em que o pesquisador encontra medidas repetidas provenientes de experimentos de tempo de resposta. Por isso, o exemplo de experimento apresentado contém apenas uma variável preditora e é bastante simples. Foge de nosso escopo detalhar os procedimentos para interpretação de resultados de modelos mais complexos, que contenham duas ou mais variáveis preditoras e suas interações. No entanto, destacamos que, nesse caso, também se mantém o método de comparação de modelos aninhados. Em vez de reportar os coeficientes do modelo mais complexo, o pesquisador primeiro testaria a significância dos efeitos fixos por meio de comparação de modelos, e então reportaria os coeficientes do modelo mais simples que melhor se ajusta aos dados. Para um passo-a-passo mais detalhado sobre como conduzir a análise de MLMs, recomendamos os tutoriais de Winter (2013, em inglês) e Godoy (2019, em português).

---

7 No caso de precisar relatar os p-valores dos coeficientes de um modelo com efeitos fixos, considere o uso do pacote lmertest conforme indicado em Godoy (2019).

7. Discussão

Conforme visto nas análises feitas neste artigo, o modelo ANOVA apresentou resultados conflitantes em relação ao MLM. Enquanto a análise de dados tradicional, baseada nas estatísticas F_1 e F_2, apresentou efeito significativo de *prime* por participante e por item, o modelo proposto neste trabalho não encontrou esses efeitos ao tomar conjuntamente a variabilidade dos efeitos aleatórios. Dado este quadro, nos posicionamos junto a Baayen, Davidson e Bates (2008), Barr *et al.* (2013), Bates *et al.* (2015), Boker *et al.* (2013), Winter (2013) e Godoy (2019) ao assumirmos que os resultados oriundos de MLMs são preferíveis.

O primeiro argumento que podemos usar a favor do MLM é a sua idade. Assim como todas as ciências, a Estatística também evoluiu. Quando Clark (1973) propôs seu método de análise, a teoria de Modelos Lineares ainda estava incipiente. As últimas quatro décadas trouxeram muitos avanços para a análise de dados de experimentos, e é natural que ciências experimentais - como a psicolinguística - acompanhem essas mudanças.

Além disso, no começo da década de 1970, a computação científica estava em sua infância. Naquela época era desejável utilizar métodos estatísticos mais simples, pois o acesso a computadores não era largamente democratizado como nos dias atuais. Isto é particularmente importante em relação aos MLMs, pois estes são métodos que dependem da convergência de algoritmos iterativos. Estes algoritmos possuem instruções que são executadas centenas ou até milhares de vezes, o que tornava impraticável a aplicação destes métodos algumas décadas atrás. O pesquisador nos dias atuais não precisa limitar os métodos de análise que utiliza de acordo com as restrições de *hardware* existentes no passado.

Experimentos complexos exigem métodos de análise mais sofisticados. Embora à primeira vista não pareça que estamos trabalhando com um experimento deste tipo, lembremos que cada sujeito teve seu tempo de resposta medido mais de uma vez. Ou seja, a hipótese de independência das observações exigida pelo modelo ANOVA é violada, e isso aumenta a complexidade do experimento realizado. Se lembrarmos do tipo de cálculo necessário para agregar os tempos de resposta por participante e por item, percebemos que o pesquisador está descartando informações importantes coletadas durante o experimento. Utilizar apenas a média por participante ou por item despreza o comportamento que é mostrado na Figura 2, na qual vemos que cada participante reage do seu jeito ao experimento que foi aplicado.

Por fim, acreditamos que a resposta da análise de dados deve ser assertiva. O pesquisador deve ser capaz de concluir se o efeito de *prime* é significativo ou não. Ao reportar as estatísticas F_1 e F_2 e seus respectivos p-valores, é possível que apenas uma delas seja significativa. Assim, não é possível afirmar de maneira definitiva se o efeito de *prime* está presente ou não na análise realizada. Mais uma vez, uma análise por MLMs mostra vantagem por permitir que um único modelo combine dois ou mais efeitos aleatórios.

8. O uso de MLM em estudos linguísticos no Brasil

Há ainda poucos trabalhos no Brasil que fazem uso de MLM. Essa escassez se explica, em parte, pelo fato de a técnica ser relativamente nova nos estudos da linguagem. Podemos traçar sua popularização com a publicação de Baayen, Davidson e Bates (2008) no periódico *Journal of Memory and Language*, provavelmente responsável pelo status privilegiado de que esse tipo de análise goza atualmente na comunidade internacional. Contudo, a falta de material em português pode dificultar o acesso de novos pesquisadores a essa classe de modelos.

Em termos de materiais disponíveis para consulta e para introdução a modelos lineares, há duas opções gratuitas e com foco em pesquisa linguística no país. Há o material do curso de estatística de Oushiro (2017), voltado especialmente para dados oriundos de pesquisa sociolinguística e com duas sessões sobre regressão linear. O tutorial de Godoy (2019) é uma opção para pesquisadores da área de psicolinguística que queiram começar a fazer análises por meio de modelos lineares. Ambos os materiais fazem uso da linguagem R para análise de dados. Com o auxílio desses recursos e a argumentação construída ao longo deste artigo, esperamos fomentar o uso de análises mais atuais que, no geral, contornam as limitações presentes em uma ANOVA tradicional.

REFERÊNCIAS

BAAYEN, R. H.; DAVIDSON, D. J.; BATES, D. M. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, v. 59, n. 4, p. 390–412, 2008.

BARR, D. J. *et al.* Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, v. 68, n. 3, p. 255–278, abr. 2013.

BATES, D. *et al.* Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, v. 67, n. 1, p. 1–48, 2015.

BOLKER, B. M. *et al.* Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS. *Methods in Ecology and Evolution*, v. 4, n. 6, p. 501–512, abr. 2013.

CLARK, H. H. The Language-as-Fixed-Effect Fallacy: A Critique of Language Statistics in Psychological Research. *Journal of Verbal Learning and Verbal Behavior*, v. 12, p. 335–359, 1973.

COSTA, I. DE O. *Verbos meteorológicos no plural em orações relativas do português brasileiro: sintaxe e processamento*. 2013. Dissertação (Mestrado em Estudos da Linguagem). Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2013.

DAVENPORT, J. M.; WEBSTER, J. T. A comparison of some approximate *F*-tests. *Technometrics*, v. 15, p. 779–789, 1973.

DEBRUINE, L. M.; BARR, D. J. Understanding mixed effects models through data simulation *PsyArXiv*, jun. 2019. Disponível em: <psyarxiv.com/xp5cy>. Acesso em 9 jun. 2020.

GARCIA, G. D. When lexical statistics and the grammar conflict: Learning and repairing weight effects on stress. *Language* 95(4), p. 612-641, 2019.

GODOY, M. C. Introdução aos modelos lineares mistos para os estudos da linguagem. *PsyArXiv*, 2019. Disponível em < https://doi.org/10.17605/OSF.IO/9T8UR>. Acesso em 9 jun. 2020.

GODOY, M. C. *et al*. O papel do conhecimento de eventos no processamento de sentenças isoladas. *Letrônica*, v. 10, n. 2, p. 538–554, 2017.

KRUSCHKE, J. K., LIDDEL, T. M. Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 25(1), p. 155–177, 2018.

KUTNER, M. *et al. Applied Linear Statistical Models*. 5. ed. New York: McGraw-Hill/Irwin, 2004.

LOPES, S. R. C.; NUNES, M. A. Long memory analysis in DNA sequences. *Physica A*: Statistical Mechanics and its Applications, v. 361, n. 2, p. 569–588, mar. 2006.

MORETTIN, P. A.; BUSSAB, W. DE O. *Estatística básica*. São Paulo: Saraiva, 2004.

OUSHIRO, L. *Introdução à Estatística para Linguistas*. Zenodo, 2017. Disponível em https://zenodo.org/record/822070#.Xo9qHdNKjOQ. Acesso em 9 jun. 2020.

RAAIJMAKERS, J. G. W. A Further Look at the "Language-as-Fixed-Effect Fallacy". *Canadian Journal of Experimental Psychology*, v. 57, n. 3, p. 141–151, 2003.

RAAIJMAKERS, J. G. W.; SCHRIJNEMAKERS, J. M. C.; GREMMEN, F. How to Deal with "The Language-as-Fixed-Effect Fallacy": Common Misconceptions and Alternative Solutions. *Journal of Memory and Language*, v. 41, p. 416–426, 1999.

SMITH, J. E. K. The assuming-will-make-it-so fallacy. *Journal of Verbal Learning and Verbal Behavior*, v. 3, p. 262–263, 1976.

WIKE, E. L.; CHURCH, J. D. Comments on Clark's "The language-as-fixed-effect fallacy". *Journal of Verbal Learning and Verbal Behavior*, v. 15, n. 3, p. 249–255, 1976.

WINTER, B. Linear models and linear mixed effects models in R with linguistic applications. *CoRR*, v. abs/1308.5499, 2013.

WINTER, B. *Statistics for Linguists*: An Introduction Using R. New York: Routledge 2019.

PARECER DE GUILHERME DUARTE GARCIA NO ARTIGO "UMA COMPARAÇÃO ENTRE ANOVA E MODELOS LINEARES MISTOS PARA ANÁLISE DE DADOS DE TEMPO DE RESPOSTA"
DOI 10.25189/rabralin.v19i1.13881

O artigo *Uma comparação entre* ANOVA *e modelos lineares mistos para análise de dados de tempo de resposta* traz uma excelente contribuição metodológica não apenas à psicolinguística brasileira, mas também à linguística experimental no País. Os autores apresentam uma comparação didática e facilmente reproduzível que demonstra a superioridade de modelos mistos com relação a ANOVAs. Neste comentário, focarei nas principais qualidades do artigo, e adicionarei algumas observações sobre os dados e métodos empregados no estudo. Espera-se que este modelo de publicação comentada fomente discussões produtivas a respeito do tema abordado no artigo.

Comentários gerais: de ANOVAS a modelos mistos

A utilização de ANOVAs ainda é frequente em diferentes subáreas da linguística (e.g., ver Plonsky (2015) para uma revisão sobre métodos quantitativos em aquisição de segunda língua). Um exemplo clássico da má utilização de ANOVAs em nossa área envolve a transformação de respostas binárias em percentuais. Ou seja, parte-se de uma variável binária ou categórica (e.g., 0/1), cria-se uma variável pseudocontínua derivada (e.g., percentuais), e analisa-se a nova variável a partir de uma ANOVA tradicional. Embora saibamos há muito tempo que esse tipo de abordagem é problemática, o costume parece persistir em diversas subáreas da linguística.

Em um influente artigo publicado no *Journal of Memory and Language*, Jaeger (2008) demonstra que ANOVAs devem ser evitadas na análise de dados categóricos (mesmo que esses dados sejam transformados em percentuais). O autor demonstra que modelos mistos (neste caso, regressões logísticas) são sempre preferíveis e mais confiáveis com relação a ANOVAs tradicionais.

Com base em Jaeger (2008), um defensor de ANOVAs poderia, então, argumentar que o método ainda é preferível quando variáveis dependentes são de fato contínuas, como no caso de tempos de resposta. Contudo, também sabemos que modelos mistos serão, por definição, superiores a ANOVAs, simplesmente por comportarem uma estrutura mais complexa (i.e., hierárquica), que oferecem uma abordagem mais realista com relação à variabilidade dos dados com que lidamos.

O presente artigo (*Uma comparação entre* ANOVA *e modelos lineares mistos para análise de dados de tempo de resposta*) exemplifica a superioridade de modelos mistos a partir de dados simulados de tempos de resposta. Os autores demonstram que um modelo misto é superior mesmo quando implementamos uma abordagem mais conservadora em que *duas* ANOVAs levem em conta a variabilidade de itens e de participantes: embora obtenhamos um efeito significativo de *priming* em ambas as ANOVAs, o que nos leva a rejeitar a hipótese nula, um modelo misto que inclui variabilidade de itens e participantes em *uma única regressão* nos mostra que não há um efeito significativo de

priming ($p > 0.05$). Ou seja, uma abordagem com ANOVAs, neste caso, nos leva a cometer um erro de Tipo I. Um modelo misto, por outro lado, nos oferece uma análise mais conservadora, uma vez que considera diferentes graus de variabilidade nos dados *simultaneamente*.

Uma das grandes vantagens do artigo é a replicabilidade da análise implementada em R (R Core Team 2019), já que os dados simulados e o *script* em R são disponibilizados pelos autores. Dessa forma, mesmo leitores que não estejam acostumados à linguagem R poderão reproduzir a análise discutida no artigo.

A mensagem final deste artigo é a mesma encontrada em diversos outros artigos em linguística nos últimos vinte anos (e.g., Baayen et al (2008), Matuschek et al (2017)): abandone ANOVAs e utilize modelos mais robustos—preferencialmente mistos. Mesmo em situações em que uma estrutura hierárquica (i.e., mista) seja menos necessária (e.g., análise de corpus ou léxico), modelos mais completos ("*full-fledged statistical models*") proporcionarão ao pesquisador uma imagem mais abrangente dos dados analisados. Além disso, o *output* de um modelo estatístico é consideravelmente mais informativo do que o típico *output* de uma ANOVA (e.g., coeficientes e magnitudes de efeitos)—muito embora uma ANOVA seja essencialmente um modelo linear simples.

Comentários específicos

Os dados simulados no artigo representam tempos de resposta. Se criarmos um histograma (`hist(priming$tempo)`), veremos que os tempos de reação simulados seguem uma distribuição normal (algo comum em simulações de dados que utilizam amostras randômicas a partir de distribuições gaussianas). Tipicamente, contudo, tempos de resposta terão uma distribuição assimétrica: alguns participantes terão um tempo de resposta alto, mas nenhum terá um tempo de resposta negativo. Por essa razão, dados de tempo de resposta são comumente transformados (e.g., log) para que aproximem uma distribuição normal. Embora isso não afete o artigo nem a análise dos autores, uma aplicação realista da análise em questão envolverá distribuições não normais, que exigirão algum tipo de transformação.

Onde estão os valores *p*?

Os leitores do presente artigo devem ter percebido que o output de um modelo misto (lmer()) não contém valores p. Repito abaixo parte do output de modelo.prime. Nele, temos o coeficiente, o erro padrão, e o valor t para cada efeito fixo—não há uma coluna para valor p. O distanciamento de valores p em modelos mistos, ainda que superficial, é mais uma vantagem com relação a ANOVAs tradicionais, em que o foco costuma estar na rejeição de uma hipótese nula. Embora aqui também estejamos

efetuando testes de hipóteses, a magnitude do efeito de cada variável (Estimate, ou $\widehat{\boldsymbol{\beta}}$) é "promovida" no output gerado pelo modelo. Ou seja, intuitivamente, estamos mais interessados em estimar o tamanho de um efeito do que em rejeitar ou não a hipótese nula.

```
Fixed effects:
            Estimate Std. Error t value
(Intercept)   695.01      30.38  22.880
primenrel      62.54      34.63   1.806
```

Há diferentes aproximações de valores p para modelos mistos, e cada aproximação pode gerar valores p distintos. Não seria, portanto, prudente presumir que um desses valores é o valor "correto" com relação a outros valores gerados. Por essa razão, a sugestão é não aproximar o valor p utilizando pacotes disponíveis (e.g., lmerTest). Tecnicamente, o pesquisador poderia apenas reportar o valor t: quanto maior esse valor, maior será a evidência contra a hipótese nula. O valor t equivale ao coeficiente dividido pelo erro padrão: $t = \frac{Estimate}{Std.\ Error}$. Mais especificamente, se usarmos um valor α = 0.05, um valor |t| > 1.96 será equivalente a um valor p < 0.05. Ou seja, se desejamos de fato dicotomizar o resultado a partir de valores p, é suficiente reportar o valor t.

Como o valor |t| de modelo.prime está abaixo de 1.96, prevemos que uma aproximação de valor p gerará um valor acima de 0.05 neste caso. Portanto, não é surpreendente que a variável prime não resulte em um modelo significativamente melhor com relação a um modelo nulo na comparação de modelos aninhados reportada pelos autores.

Por fim, é importante ter em mente que valores p são problemáticos independentemente do tipo de modelo utilizado: imaginar que um resultado é definitivamente real apenas porque temos um valor p abaixo de 0.05, ou imaginar que esse efeito é definitivamente falso apenas porque temos um valor p acima de 0.05 é uma simplificação que deveríamos evitar, mas que está quase sempre presente na estatística frequentista. A literatura recente está repleta de discussões a respeito das desvantagens de focarmos nossas análises estatísticas em uma noção binária que contrasta algo "significativo" com algo "não significativo" com base em um valor α arbitrário (e.g., Kruschke (2015), Kruschke and Liddell (2018); para um exemplo recente de estatística bayesiana aplicada a estudos linguísticos, ver Garcia (2019)).

Espera-se que o artigo em questão contribua para o avanço dos métodos de análise utilizados em psicolinguística no Brasil. Mesmo em teoria linguística, a análise de dados tem um papel fundamental—afinal, teorias precisam ser testadas e examinadas cuidadosamente. Não é possível avaliar conclusivamente um modelo teórico sem determinar suas predições empíricas, e modelos mistos certamente contribuem com essa avaliação por parte do pesquisador, uma vez que levam em conta diferentes níveis de variabilidade, algo essencial se pretendemos entender a gramática para além de diferenças individuais.

REFERÊNCIAS

BAAYEN, R. H.; DAVIDSON, D. J.; BATES, D. M. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390–412, 2008.

GARCIA, G.D. When lexical statistics and the grammar conflict: Learning and repairing weight effects on stress. *Language* 95(4), 612-641, 2019.

KRUSCHKE, J. K. Doing Bayesian data analysis: a tutorial with R, JAGS, and Stan. 3ª ed. Academic Press, London, 2015.

KRUSCHKE, J. K.; LIDDELL, T. M. Bayesian data analysis for newcomers. *Psychonomic Bulletin & Review*, 25(1):155–177, 2018.

JAEGER, T. F. Categorical Data Analysis: Away from ANOVAs (transformation or not) and towards Logit Mixed Models. *Journal of Memory and Language*, 59(4), 434–446, 2008.

MATUSCHEK, H.; KLIEGL, R.; VASISHTH, S.; BAAYEN, R. H.; BATES, D. M. Balancing Type I Error and Power in Linear Mixed Models. *Journal of Memory and Language*, 94:305-315, 2017.

PLONSKY, L. *Advancing quantitative methods in second language research*. New York, Routledge, 2015.

R Core Team (2019). *R*: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2019.