Modelos multinível: *turtles all the way down*

Luiz Max Carvalho

Motivação

O modelo multinível é uma extensão do modelo linear generalizado (GLM) que permite a análise de dados agrupados ou hierárquicos. Ele é especialmente útil quando os dados são coletados em diferentes níveis, como alunos dentro de escolas ou pacientes dentro de hospitais.

Por exemplo, considere um estudo sobre o desempenho acadêmico de alunos em diferentes escolas, onde o objetivo é prever as notas dos alunos, y, com base em um pré-teste, x, e outra informação. Um modelo de regressão pode ser ajustado considerando as características das escolas, como o tamanho da escola, a proporção de alunos por professor, etc. Para isso, podemos considerar que as notas dos alunos são influenciadas por características individuais e por características da escola:

- Modelo de intercepto aleatório:

$$\begin{aligned} y_{ij} &= \beta_0 + \beta_1 x_{ij} + u_{0j} + \epsilon_{ij}, \\ u_{0j} &\sim \mathrm{N}(0, \sigma_u^2), \\ \epsilon_{ij} &\sim \mathrm{N}(0, \sigma^2), \end{aligned}$$

  onde y_{ij} é a nota do aluno i na escola j, x_{ij} é o pré-teste do aluno i na escola j, β_0 e β_1 são os coeficientes de regressão, u_{0j} é o efeito aleatório da escola j e ϵ_{ij} é o erro aleatório do aluno i na escola j.

- Modelo de intercepto **e** inclinação aleatórios:

$$\begin{aligned} y_{ij} &= \beta_0 + u_{0j} + (\beta_1 + u_{1j}) x_{ij} + \epsilon_{ij}, \\ \begin{pmatrix} u_{0j} \\ u_{1j} \end{pmatrix} &\sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_u^2 & \sigma_{u0u1} \\ \sigma_{u0u1} & \sigma_{u1}^2 \end{pmatrix} \right), \\ \epsilon_{ij} &\sim \mathrm{N}(0, \sigma^2), \end{aligned}$$

  onde u_{1j} é o efeito aleatório da inclinação da escola j e σ_{u0u1} é a covariância entre os efeitos aleatórios de intercepto e inclinação.

Aqui, o modelo para o indivíduo i é o primeiro nível, enquanto o modelo para a escola j é o segundo nível.

"Multinível" ou "Hierárquico"

Modelos multiníveis são frequentemente chamados de modelos hierárquicos, pois eles lidam com dados que têm uma estrutura hierárquica. Por exemplo, em um estudo educacional, os alunos estão aninhados dentro de escolas, e as escolas podem estar aninhadas dentro de distritos escolares. Essa estrutura hierárquica é o que distingue os modelos multiníveis dos modelos de regressão tradicionais. Para mais, ver a discussão na página 245 de Gelman & Hill (2007).

Modelo linear generalizado misto

De modo geral, a modelagem multinível pode ser vista como uma extensão dos modelos lineares generalizados (GLMs) que incorpora efeitos aleatórios. Esses modelos são chamados de modelos lineares generalizados mistos (GLMMs). Eles permitem que os pesquisadores considerem tanto os efeitos fixos (como os coeficientes de regressão) quanto os efeitos aleatórios (como os interceptos e inclinações aleatórias) em um único modelo.

Considere observações $\boldsymbol{y}$ de uma variável resposta Y que segue uma distribuição da família exponencial, com uma função de ligação $g(\cdot)$; uma matriz de design $\boldsymbol{X}$ com n linhas e p colunas, onde cada linha corresponde a uma observação e cada coluna a uma variável preditora; uma matriz de design $\boldsymbol{Z}$ com n linhas e q colunas, onde cada linha corresponde a uma observação e cada coluna a um efeito aleatório; e um vetor de efeitos aleatórios $\boldsymbol{u}$ com q componentes. O modelo pode ser escrito como:

$$
\begin{aligned}
g(\mathbb{E}[\boldsymbol{Y} \mid \boldsymbol{U} = \boldsymbol{u}]) &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u}, \\
\boldsymbol{U} &\sim \mathrm{N}(0, \boldsymbol{\Sigma}_{\boldsymbol{U}}).
\end{aligned}
$$

Ajuste do modelo

Para ajustar um modelo multinível, podemos estimar os parâmetros usando máxima verossimilhança (ML) ou máxima verossimilhança restrita (REML). Em R, o pacote `lme4` é amplamente utilizado para ajustar modelos lineares generalizados mistos. A função `lmer()` é usada para ajustar modelos lineares mistos, enquanto a função `glmer()` é usada para ajustar modelos lineares generalizados mistos. Este link é um bom tutorial sobre o uso do `lme4`.

Exemplo

Vamos considerar um exemplo de dados de um estudo de privação de sono, onde o objetivo é modelar o tempo de reação dos sujeitos com base no número de dias de sono. Os dados estão disponíveis no pacote `lme4` e podem ser carregados com o seguinte código:

```
library(lme4)
```

```
Loading required package: Matrix
```

```
library(lattice) # Para visualização dos efeitos aleatórios
data("sleepstudy", package = "lme4")
head(sleepstudy)
```

```
  Reaction Days Subject
1 249.5600    0     308
2 258.7047    1     308
3 250.8006    2     308
4 321.4398    3     308
5 356.8519    4     308
6 414.6901    5     308
```

Neste exemplo, `Reaction` é a variável resposta (tempo de reação), `Days` é o pré-teste (número de dias de sono) e `Subject` é o identificador do sujeito. Podemos ajustar um modelo linear misto usando a função `lmer()`:

```
model <- lmer(Reaction ~ Days + (1 | Subject), data = sleepstudy)
```

Neste modelo, estamos modelando o tempo de reação dos sujeitos com base no número de dias de sono, permitindo que haja um intercepto aleatório para cada sujeito. O resultado do `summary(model)` nos dará os coeficientes estimados, incluindo o intercepto fixo e o efeito aleatório.

```
summary(model)
```

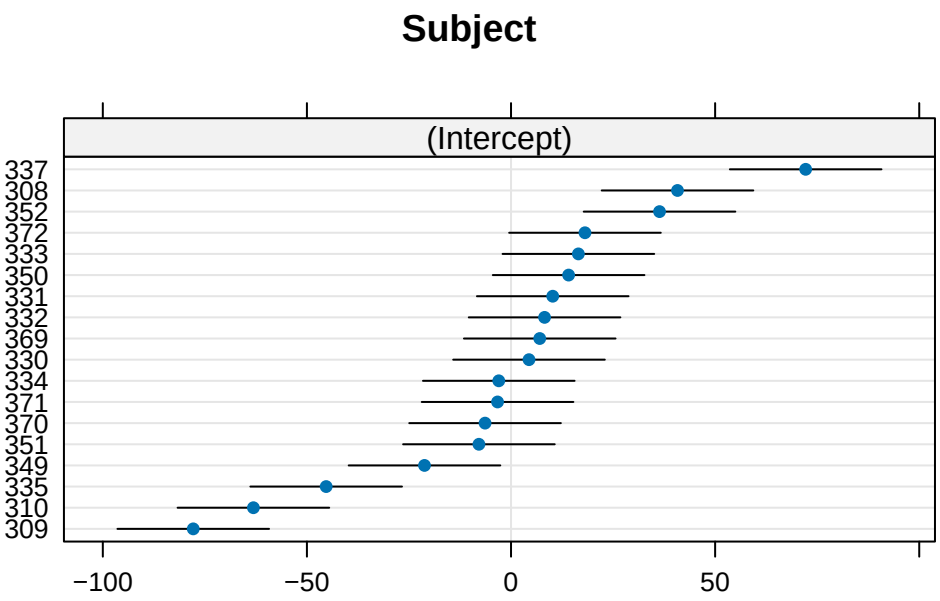
```
Linear mixed model fit by REML ['lmerMod']
Formula: Reaction ~ Days + (1 | Subject)
   Data: sleepstudy

REML criterion at convergence: 1786.5
```

```
Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.2257 -0.5529  0.0109  0.5188  4.2506

Random effects:
 Groups   Name        Variance Std.Dev.
 Subject  (Intercept) 1378.2   37.12
 Residual              960.5   30.99
Number of obs: 180, groups:  Subject, 18

Fixed effects:
            Estimate Std. Error t value
(Intercept) 251.4051     9.7467   25.79
Days         10.4673     0.8042   13.02

Correlation of Fixed Effects:
     (Intr)
Days -0.371
```

Podemos também visualizar os efeitos aleatórios usando a função `ranef()`:

```
dotplot(ranef(model))
```

```
$Subject
```



Exercícios de fixação

1. Suponha que $U_i \sim N\left(0, \sigma_u^2\right)$ e que dado $U_i = u_i$, as variáveis aleatórias binárias Y_{ij} são independentes com

$$P\left(Y_{ij} = 1 \mid U_i = u_i\right) = 1 - P\left(Y_{ij} = 0 \mid U_i = u_i\right) = \Phi\left(\beta_0 + \beta_1 x_{ij} + u_i\right).$$

   Aqui Φ é a distribuição normal acumulada padrão, e os x_{ij} são valores de alguma covariável conhecida. Este é um modelo multinível probit. Mostre que

$$P\left(Y_{ij} = 1\right) = 1 - P\left(Y_{ij} = 0\right) = \Phi\left(\gamma_0 + \gamma_1 x_{ij}\right),$$

   exibindo γ_0 e γ_1 em termos de β_0, β_1 e σ_u^2.

2. Agora, suponha que $U_i \sim N\left(0, \sigma_u^2\right)$ e que condicionalmente a $U_i = u_i$, as variáveis aleatórias Y_{ij} são independentes e seguem uma distribuição gama com $\mathrm{E}\left[Y_{ij} \mid U_i = u_i\right] = \exp\left(\beta_0 + \beta_1 x_{ij} + u_i\right)$. Encontre uma expressão para a média marginal $\mathrm{E}\left[Y_{ij}\right]$ e para a variância marginal $\mathrm{Var}\left[Y_{ij}\right]$.

Referências

- Gelman, A., & Hill, J. (2007). Data analysis using regression and multilevel/hierarchical models. Cambridge university press.