Modelos Mistos: Conceitos iniciais e representação matemática

1 Regressão Linear de efeitos mistos

Ao contrário dos modelos de regressão de efeitos fixos, como os tradicionais modelos ANOVA, os modelos de efeitos mistos assumem uma estrutura hierárquica de dados, onde os pontos de dados são agrupados ou aninhados em categorias de níveis superiores. Isto pode incluir medidas repetidas dentro de cultivares, cultivares dentro de regiões geográficas ou qualquer outro arranjo aninhado ou agrupado. Os modelos de efeitos mistos são caracterizados como uma extensão dos modelos de regressão tradicionais, destacando-se pela habilidade de **modelar simultaneamente efeitos fixos e aleatórios**. Tal modelo utiliza componentes de variância para explicar tanto as partes aleatórias explicadas quanto as inexplicadas.

Ao se limitar apenas aos efeitos fixos, todos os pontos de dados são tratados como independentes, situando-se no mesmo nível hierárquico. No entanto, é comum que os dados estejam estruturados de forma aninhada, o que implica que os pontos de dados não são independentes entre si, podendo ser produzidos pela mesma cultivar ou agrupados por outras características semelhantes, por exemplo. Em tais casos, os **dados são considerados hierárquicos**, e os modelos estatísticos precisam levar em conta essas estruturas para uma análise mais precisa. Felizmente, o pacote *lme4* facilita a modelagem de estruturas de dados hierárquicos ou aninhados.

A regressão linear de efeitos mistos oferece diversas vantagens em comparação com modelos estatísticos mais simples:

- São multivariados, o que significa que avaliam o efeito de múltiplos preditores simultaneamente, enquanto controlam os efeitos de todos os outros preditores.
- Permitem incorporar estatisticamente a variabilidade dentro dos genótipos, portanto, são adequados para modelar estruturas de dados hierárquicos ou aninhados. Isso se aplica se várias observações forem produzidas por um mesmo genótipo, por exemplo.
- Permitem a inclusão estatística da variabilidade dentro das cultivares, tornando-os adequados para modelar estruturas de dados hierárquicos ou aninhados. Isso é particularmente útil quando várias observações são produzidas pela mesma cultivar.
- Oferecem uma variedade de estatísticas diagnósticas que permitem controlar problemas como multicolinearidade (correlações entre preditores) e testar a violação de condições ou requisitos, como homogeneidade de variância.

1.1 Representação matemática do modelo misto

Nos modelos lineares, existem dois principais objetivos. O primeiro é estimar os valores dos parâmetros do modelo, como α e β, que descrevem a relação entre as variáveis independentes e dependentes. O segundo é estimar as variâncias apropriadas, como a variância do erro e. Por exemplo, considerando um modelo linear simples:

$$y = \alpha + \beta X + e \tag{1}$$

Não estima-se apenas os valores de α e β, que representam o intercepto e a inclinação da reta de regressão, respectivamente, mas também a variância do erro e. É importante ressaltar que os parâmetros α e β em um modelo simples, são considerados constantes fixas, enquanto o erro e é considerado uma variável aleatória que é amostrada de uma distribuição de probabilidade. Geralmente, assume-se que o erro tem média zero e variância σ_e^2.

Ao utilizar um modelo linear misto, torna-se viável a predição dos efeitos aleatórios mesmo na presença de efeitos fixos. A equação geral do modelo misto desenvolvida por Henderson é representada da seguinte forma:

$$y = \underbrace{X\beta}_{\text{Fixo}} + \underbrace{Zu + e}_{\text{Aleatório}} \tag{2}$$

$$u \sim N(0, G), \qquad e \sim N(0, R) \tag{3}$$

Em que:
y = o vetor para uma variável de interesse (n x 1);
X = a matriz de incidência dos efeitos fixos conhecida (n x p);
β = o vetor dos efeitos desconhecidos (p x 1);
Z = a matriz de incidência dos efeitos aleatórios conhecida (n x q);
u = o vetor de efeitos aleatórios desconhecido (q x 1);
e = o vetor de erros aleatórios (n x 1).

A matriz de incidência X associa cada observação aos efeitos fixos, enquanto a matriz de incidência Z associa cada observação aos efeitos aleatórios, ambos são uma matriz de incidência. Os vetores de parâmetros β e u representam os coeficientes associados aos efeitos fixos e aleatórios, respectivamente. O vetor de erros e captura a variação não explicada no modelo. Reescrevevendo a equação (2) para um contexto multivariado, temos a seguinte formulação:

$$\begin{aligned} y_1 &= X_1\beta_1 + Z_1u_1 + e_1 \\ y_2 &= X_2\beta_2 + Z_2u_2 + e_2 \\ &\cdots \\ y_i &= X_i\beta_i + Z_iu_i + e_i \end{aligned} \tag{4}$$

Agora, reformulando as equações (4) para representar de forma matricial para os dados observados, vamos considerar que temos n observações. Nesse caso, a matriz X terá dimensão (n x p), a matriz Z terá dimensão (n $\times$ q), e os vetores y, β e e terão dimensão (n $\times$ 1), temos:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{12} & \cdots & X_{1p} \\ 1 & X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} Z_{11} & Z_{12} & \cdots & Z_{1q} \\ Z_{21} & Z_{22} & \cdots & Z_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{n1} & Z_{n2} & \cdots & Z_{nq} \end{bmatrix} \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_q \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} \tag{5}$$

Isso reflete a estrutura matricial da equação do modelo misto, onde os efeitos fixos e aleatórios são combinados linearmente com as matrizes de incidência X e Z, para prever as respostas observadas y, adicionando os erros e.

Tanto os efeitos aleatórios u quanto os resíduos e, são amostras de uma distribuição normal com média zero, sendo suas esperanças dadas por: $E(y) = X\beta$, $E(u) = 0$ e $E(e) = 0$. As suposições sobre a distribuição de y, u, e e a estrutura da variância e covariância (VCOV) são expressas na seguinte formulação:

$$\begin{bmatrix} y \\ u \\ e \end{bmatrix} \sim N \left(\begin{bmatrix} X\beta \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} ZGZ' + R & ZG & R \\ GZ' & G & 0 \\ R & 0 & R \end{bmatrix} \right) \tag{6}$$

A estrutura de variância dos resíduos é determinada pela matriz R, enquanto a estrutura de variância dos efeitos aleatórios é determinada pela matriz G. Os zeros presentes nessas matrizes indicam a ausência de covariância entre os efeitos aleatórios e os resíduos, o que implica que são independentes.

A matriz de variância e covariância de y, representada por V, pode ser expressa pela equação $V = ZGZ' + R$. Esta formulação não só depende da matriz de incidência Z, mas também das estruturas das matrizes R e G. A obtenção da matriz V pode ser expressa da seguinte forma:

$$V = Var(y) = Var(X\beta + Zu + e) = ZGZ' + R \tag{7}$$

Além disso, presume-se que V é não singular. Portanto, tem-se $E(y) = E(X\beta + Zu + e) = X\beta$, o que implica que $y \sim N(X\beta, ZGZ' + R)$.

Se cada observação contiver múltiplas medidas, a distribuição se torna multivariada. Se as medidas estiverem ordenadas dentro de cada observação no vetor y, as matrizes G e R assumem as seguintes formas:

$$G = A \otimes G_\circ, \quad R = I_n \otimes R_\circ \tag{8}$$

Onde A representa a matriz de correlação entre os efeitos aleatórios u das n observações, com dimensão n x n, e $G_\circ$ denota a matriz de variância e covariância entre os efeitos aleatórios nas q medidas que compõem uma observação, com dimensão (q x q). Similarmente, I_n é a matriz identidade de dimensão (n x n), e $R_\circ$ é a matriz de variância e covariância residual entre as q medidas que compõem uma observação, com dimensão (q x q).

1.1.1 Estimação e predição dos efeitos no modelo misto

Uma vez que os valores dos componentes de variância (G e R) ou suas estimativas são conhecidas, o próximo passo é estimar o vetor de efeitos fixos β e prever o vetor de efeitos aleatórios u. A estimativa e a previsão dos efeitos são realizadas resolvendo um sistema de equações lineares. Isso é feito maximizando a função de densidade de probabilidade conjunta de y e u. Para se proceder à maximização da função, pode-se usar a transformação por logaritmo. Ao derivar $L = \log[f(y,u)]$ em relação aos coeficientes β e u, e igualando tais derivadas a zero, obtêm-se as seguintes Equações de Modelos Mistos (MME) de Henderson:

$$\begin{bmatrix} \hat{\beta} \\ \hat{u} \end{bmatrix} = \begin{bmatrix} X'R^{-1}X & X'R^{-1}Z \\ Z'R^{-1}X & Z'R^{-1}Z + G^{-1} \end{bmatrix}^{-} \begin{bmatrix} X'R^{-1}y \\ Z'R^{-1}y \end{bmatrix} \tag{9}$$

Essas equações permitem obter o melhor estimador linear não viesado, conhecido como *Best Linear Unbiased Estimator* (BLUE), para os efeitos fixos ($\hat{\beta}$), assim como o melhor preditor linear não viesado, chamado *Best Linear Unbiased Predictor* (BLUP), para os efeitos aleatórios ($\hat{u}$).

Existem algumas propriedades da solução para os efeitos fixos e aleatórios em modelos de efeitos mistos, obtidos pelo método de Mínimos Quadrados Generalizados (MQG ou, em inglês, GLS). Para os efeitos fixos e ao considerar a primeira equação obtida pelas MME, onde o coeficiente $\hat{\beta}$ representa o BLUE de β, a obtenção é feita da seguinte maneira:

$$\hat{\beta} = (X'\hat{V}^{-1}X)^{-1}X'\hat{V}^{-1}y \tag{10}$$

A matriz de variância-covariância do estimador $\hat{\beta}$ é dada por $(X'\hat{V}^{-1}X)^{-1}$. No caso dos efeitos aleatórios e ao considerar a segunda equação obtida pelas MME, onde o coeficiente $\hat{u}$ representa o BLUP de u, a solução pode ser expressa da seguinte forma:

$$\hat{u} = GZ'(ZGZ' + R)^{-1}(y - X\hat{\beta}) \tag{11}$$

O termo "predição" refere-se aos fatores aleatórios, e o melhor preditor linear não viesado (BLUP) pode ser definido como o resultado da regressão dos efeitos de um fator aleatório u em função das observações y, ajustadas para os efeitos dos fatores fixos $X\beta$.