Modelos Mistos: Predição, Estimação e Componentes de Variância

1 Representação matemática

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.

De acordo com Martins et al (1998), 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 (WEST; WELCH; GALECKI, 2022).

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 (MARTINS et al., 1998):

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

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.

De acordo com Henderson (1984), 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 $y \sim N(X\beta, ZGZ' + R)$, como demonstrado por Freitas (2013).

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 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 $L = \log\left[f\left(y, u\right)\right]$ em relação aos coeficientes β e u, e igualando tais derivadas a zero, obtêm-se as seguintes Equações de Modelos Mistos (MME), como demonstrada por Henderson (1986):

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

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 $X\beta$.

1.2 Estimação dos componentes de variância e Covariância

Como mencionado anteriormente, a matriz de variância e covariância V está relacionada à matriz Z (que representa os efeitos aleatórios conhecidos) e às matrizes G e R (que representam as covariâncias dos efeitos aleatórios e dos erros). Nesse contexto, V, G e R são desconhecidas, necessitando de estimativas obtidas por meio de algum método estatístico. Nesse contexto, diversos métodos de estimativa estão disponíveis, sendo os mais comuns a Máxima Verossimilhança (ML) e a Máxima Verossimilhança Restrita (REML). Contudo, o método mais frequentemente empregado para estimar as componentes de variância em modelos mistos é o REML. Este método é uma variante do ML que corrige o viés nas estimativas das componentes de variância, levando em conta os graus de liberdade utilizados para estimar os efeitos fixos (MARTINS et al., 1998; DUARTE; VENCOVSKY, 2001; FREITAS, 2013).

Segundo Marcelino (2000), o método REML busca maximizar a função de densidade de probabilidade das observações, levando em conta tanto os efeitos fixos quanto os componentes de variância dos efeitos aleatórios do modelo. Esse processo envolve a separação de cada observação em duas partes independentes: uma referente aos efeitos fixos e outra aos efeitos aleatórios. Dessa forma, a função de densidade de probabilidade das observações é expressa como a soma das funções de densidade de probabilidade de cada uma dessas partes.

O REML e o BLUP são métodos intimamente relacionados. Enquanto o BLUP assume que os componentes de variância são conhecidos, o REML estima esses componentes de forma iterativa, utilizando as estimativas BLUP dos efeitos aleatórios. Em outras palavras, o REML ajusta as estimativas dos componentes de variância com base nos efeitos aleatórios estimados pelo BLUP, proporcionando uma abordagem mais precisa e menos enviesada (WEST; WELCH; GALECKI, 2022).

Como demonstrado por Freitas (2013), Duarte e Vencovsky (2001), a estimativa dos parâmetros desconhecidos é realizada por meio da maximização de uma função em relação às matrizes G e R. No caso do método REML, o logaritmo da função de verossimilhança é representado da seguinte forma:

$$
\begin{aligned}
l_{REML}(G,R) = -\frac{1}{2}\log|V| - \frac{1}{2}\log\left|X'V^{-1}X\right| - \frac{n-p}{2}\log\left[y - X\left(X'V^{-1}X\right)X'V^{-1}y\right]' V^{-1} \\
\left[y - X\left(X'V^{-1}X\right)X'V^{-1}y\right] - \frac{n-p}{2}\log\left[1+\log\left(\frac{2\Pi}{n-p}\right)\right] \qquad (12)
\end{aligned}
$$

onde p é o posto da matriz X.

1.3 Referências

DUARTE, João Batista; VENCOVSKY, Roland. Estimação e predição por modelo linear misto com ênfase na ordenação de médias de tratamentos genéticos. **Scientia Agricola**, v. 58, p. 109-117, 2001.

FREITAS, Edjane Gonçalves de. **Uso de informações de parentesco e modelos mistos para avaliação e seleção de genótipos de cana-de-açúcar**. 2013. Tese de Doutorado. Universidade de São Paulo.

HENDERSON, C. R. Estimation of variances in animal model and reduced animal model for single traits and single records. **Journal of Dairy Science**, v. 69, n. 5, p. 1394-1402, 1986.

HENDERSON, C.R. **Applications of linear models in animal breeding**. Ontario: University of Guelph, p. 462, 1984.

MARCELINO, S. D. do R. **Métodos de estimação de componentes de variância em modelos mistos desbalanceados**. Scientia agrícola. 2000. Dissertação. Universidade de São Paulo.

MARTINS, Elias Nunes et al. **Modelo linear misto**. Editora UFV: Cadernos Didáticos. 1998.

WEST, Brady T.; WELCH, Kathleen B.; GALECKI, Andrzej T. **Linear mixed models: a practical guide using statistical software**. Chapman and Hall/CRC, 2022.