Universidade de Brasília
Departamento de Estatística

Modelos hierárquicos aplicados à recomendação de cultivares no contexto da ambientômica

César Augusto Galvão

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

César Augusto Galvão

Modelos hierárquicos aplicados à recomendação de cultivares no contexto da ambientômica

Orientador: Prof. Leandro T. Correia
Coorientador(a): Prof. Rafael T. Tassinari

Projeto apresentado para o Departamento de Estatística da Universidade de Brasília como parte dos requisitos necessários para obtenção do grau de Bacharel em Estatística.

Brasília
2024

Dedicatória

Agradecimentos

...

Resumo

...

Palavras-Chave: ...

Lista de Tabelas

Lista de Figuras

Sumário

1 Introdução

1.1 Motivação

A domesticação de espécies silvestres de plantas para a agricultura é uma prática antiga e passou por diversas revoluções até os dias atuais, em que a genética biométrica e o melhoramento de precisão protagonizam a criação de cultivares e seleção de características de interesse (RESENDE; BRONDANI; CHAVES, 2023). Além disso, pressões como crescimento populacional (HICKEY et al., 2019), redução de recursos naturais disponíveis, aquecimento global e uma variedade de consequências desses fatores (JORASCH, 2019) aumentam a necessidade de se produzir alimentos e outros recursos vegetais de forma incrementalmente eficiente. Uma das soluções para isso é justamente o melhoramento de precisão.

Neste contexto, o desenvolvimento e seleção de cultivares é associado a identificação de grupos ambientais (*Target Population of Environments* ou TPE), permitindo que se aproveite ao máximo a característica de interesse (CHENU, 2015). De fato, em posse da informação de que o ambiente em que a planta se desenvolve interfere em seu fenótipo (a característica de interesse, que é uma expressão gênica), cabe estudar a interação genótipos $\times$ ambientes ($G \times E$).

O estudo desse tipo de relação é potencializado com o uso de técnicas de Sistemas de Informações Geográficas – SIG, como sensoriamento remoto, entre outros (RESENDE; BRONDANI; CHAVES, 2023). A disponibilização pública de dados coletados via satélite com diversos graus de granularidade permite a inclusão de mais covariáveis ambientais como área cultivada, cobertura vegetal, temperatura, entre outros dados geofísicos[1].

A proposta de Resende et al. (2021), que será usada de estudo de caso, é expandir o uso de TPE para um estudo ômico do ambiente, daí *ambientômica*. Os autores propõem o uso de modelos hierárquicos, e o conceito de ambientipagem, resultante de agrupamentos ambientais, para predição de performance de genótipos não observados. Isto permite, por exemplo, recomendar o melhor genótipo de um determinado cultivar para uma região em que jamais foi cultivado e assim tornar a região produtiva.

[1] Por exemplo, o serviço Google Earth Engine disponibiliza seu catálogo em <https://developers.google.com/earth-engine/datasets/>

1.2 Objetivos

O objetivo geral deste trabalho de conclusão de curso é estudar o uso de modelos lineares hierárquicos (ou multinível) para recomendação de genótipos de um determinado cultivar em uma região delimitada e ambientipada, isto é, com dados sobre a maior quantidade de características ambientais possível. Pretende-se revisar metodologicamente o estudo de Resende et al. (2021), detalhando o processo de modelagem e sua adequação, bem como comparar computacionalmente variações do modelo utilizado.

Os seguintes objetivos específicos são propostos:

- Explorar a técnica de modelagem estatística via modelos lineares hierárquicos incluindo efeitos aleatórios;
- Explorar os conceitos necessários para aplicação do modelo ao contexto de melhoramento de plantas e ambientômica;
- Reproduzir a análise dos mesmos dados simulados feita no estudo de Resende et al. (2021);
- Comparar o desempenho do modelo original dos autores com um modelo que faça composição de marcadores ambientômicos utilizando análise fatorial.

Pretende-se ainda obter, para o relatório final, uma base de dados reais para a aplicação da análise descrita e comparar a metodologia original com a proposta neste relatório sem as possíveis complicações de utilização de dados simulados.

2 Referencial Teórico

2.1 Modelos Lineares de um nível

Modelos Lineares apresentam uma relação estocástica entre duas ou mais variáveis. Sua forma simples com efeitos fixos pode ser representada da forma

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k \, x_k + \varepsilon_i, \tag{2.1.1}$$

em que y_i refere-se à i-ésima observação de uma variável resposta, β_0 é o intercepto, β_k são os coeficientes associados às covariáveis x_k e ε_i é um erro estocástico associado à observação. Ao final do processo de modelagem, espera-se obter um modelo da forma

$$\hat{y}_i = \hat{\beta}_0 + \sum_{k=1}^{p} \hat{\beta}_k \, x_k, \tag{2.1.2}$$

em que $\hat{\beta}_k$ são estimadores, tipicamente obtidos pelo método de mínimos quadrados ordinários ou máxima verossimilhança, para β_k (KUTNER et al., 2005).

Este modelo de regressão, que também pode ser chamado de *modelo de efeitos fixos*, exige uma série de suposições a respeito da componente aleatória, que são avaliadas na etapa diagnóstica da modelagem como heteroscedasticidade, independência e distribuição Normal. Caso uma ou mais suposições não possam ser verificadas, se observe colinearidade entre as covariáveis do modelo, pontos de alavancagem ruins, ou outros comprometimentos do modelo, recorre-se a métodos de remediação, como transformações, redução de dimensionalidade, entre outros.

Transformações e outras remediações trazem complexidade à interpretação do modelo linear tradicional (no qual se supõe $\varepsilon_i \overset{iid}{\sim} N_1(0, \sigma^2)$). Além disso, introduzir outras representações de estruturas observacionais ou experimentais como dados categorizados, variável resposta discreta, níveis de agregação das observações, entre outros podem trazer complicações inferenciais (HOX; MOERBEEK; SCHOOT, 2017).

2.2 Modelos Lineares Hierárquicos

Frequentemente pesquisas em domínios variados do conhecimento estudam fenômenos em que as unidades de análise são agregadas em categorias (ADEWALE et al., 2007; MCMAHON; DIEZ, 2007). Em alguns casos, as unidades são aninhadas em um ou mais níveis superiores. Esses diferentes níveis de análise, indivíduos ou grupos, e suas características ou intervenções sobre níveis diferentes requerem diferentes formas de representação e técnicas de inferência que comportem adequadamente as estruturas de covariância envolvidas.

Modelos multinível substituem duas práticas comuns na utilização de regressões lineares: transformação de variáveis categóricas em variáveis binárias (*dummy*) e planificação do nível de análise, ou seja, utilização de medidas de grupos e indivíduos como descritores diretos da unidade de análise. A utilização de um modelo multinível permite a construção de estimadores que contornam essas estratégias e representam melhor indivíduos e grupos no contexto de suas características (HOX; MOERBEEK; SCHOOT, 2017; GELMAN; HILL, 2006). Esse tipo de modelo linear pode ser representado na forma

$$y_{ij} = \beta_{0j} + \sum_{k=1}^{p} \beta_{kj}\, x_{ik} + \varepsilon_{ij}, \tag{2.2.1}$$

em que y_{ij} é a variável resposta a nível de indivíduo, β_{0j} é o intercepto para o grupo $j = 1, 2, \ldots, J$ a que esse indivíduo pertence, β_{kj} são os coeficientes para cada covariável de nível individual $x_{ik}, k = 0, 1, \ldots, K$ e $\varepsilon_{ij} \overset{iid}{\sim} N(0, \sigma^2_{\varepsilon})$ é a componente aleatória para indivíduos.

No entanto, $\boldsymbol{\beta}$ é estimado a partir das covariáveis de nível superior. Se considerarmos apenas um nível, cada β_{kj} é expresso por

$$\beta_{kj} = \gamma_{k0} + \sum_{l=1}^{L} \gamma_{klj}\, z_{lj} + u_{kj}, \quad l = 1, 2, \ldots, L \tag{2.2.2}$$

em que γ_{k0} é o componente fixo, γ_{klj} é o coeficiente para cada covariável $z_{l(.)}$ de nível superior j e $u_{kj} \overset{iid}{\sim} N(0, \sigma^2_{u_{kj}})$ é a componente aleatória de cada $\beta_{k(.)}$ do grupo j. Uma propriedade deste tipo de modelo é que $E(\gamma_{klj}) = 0$, de modo que é possível depreender da equação (2.2.2) que $\beta_k \sim N(0, \sigma^2_{u_k})$.

Se for considerado o caso simplificado de apenas uma covariável de cada nível e substituirmos (2.2.2) em (), obtém-se

$$y_{ij} = \gamma_{00} + \gamma_{01} z_{1j} + \gamma_{10} x_{1ij} + \gamma_{11} z_{1j} x_{1i} + u_{1j} x_{1ij} + \varepsilon_{ij} + u_{0j}. \tag{2.2.3}$$

É imediato da equação (2.2.3) que:

- Existe um intercepto geral – γ_{00};
- Existem efeitos que agem exclusivamente sobre variáveis de um nível hierárquico específico – $\gamma_{01} z_{1j}$ e $\gamma_{10} x_{1ij}$;
- Existe um efeito de mediação do comportamento do grupo sobre a unidade de observação – $\gamma_{11} z_{1j} x_{1i}$;
- Existe uma componente de variância do grupo que incide sobre o comportamento da unidade – $u_{1j} x_{1ij}$; e
- Existem componentes de variância entre unidades e entre grupos – ε_{ij} e u_{0j} respectivamente.

As componentes de variância deste modelo são obtidas a partir do modelo ajustado apenas com os interceptos (HOX; MOERBEEK; SCHOOT, 2017) de ε_{ij} e u_{0j}, de modo que se pode calcular a proporção de variância no segundo nível da hierarquia, entre agrupamentos. Essa estatística pode ser interpretada como uma correlação entre indivíduos de um mesmo grupo, presumidamente mais similares entre si quando comparados a outro grupo. Essa medida é chamada de correlação intraclasse e, para o caso de apenas dois níveis, é dada por

$$\rho = \frac{\sigma^2_{u_0}}{\sigma^2_{u_0} + \sigma^2_{\varepsilon}}. \tag{2.2.4}$$

3 Metodologia

3.1 Software

Para todo o relatório preliminar foi utilizada a linguagem R versão 4.3.2 e os seguintes pacotes:

- Tidyverse 2.0.0;
- scales 1.3.0;
- sommer 4.3.3;
- lme4 1.1-35.1

O código que foi utilizado para a análise deste relatório parcial está disponível no repositório pessoal do autor[2].

3.2 Conjunto de dados

Para a análise preliminar, foram utilizados os dados desenvolvidos via simulações em Resende et al. (2021), que conta com três subconjuntos: (1) dados fenotípicos, (2) dados de covariáveis ambientais e (3) dados de parentalidade. O princípio desses dados é simular o desempenho de uma grande variedade de genótipos, alguns dos quais pertencem à mesma linhagem familiar, e seus desempenhos em uma superfície dividida em pixels.

O subconjunto de dados fenotípicos (características observáveis do organismo) contém dados de produção (*yield*) para cada genótipo (variedade de uma espécie) em cada célula experimental (*trial*).

A base de dados para marcadores ambientais contém 103 variáveis, 100 das quais são marcadores ambientais hipotéticos. As outras três contém uma variável indicadora de qual ensaio experimental ocorreu em uma determinada célula – se ocorreu – e duas variáveis indicadoras de localização, latitude e longitude. Essas são ilustrativas, discretas e sem unidades interpretáveis, pois se está trabalhando com um mapa simulado de dimensão 100×100.

Por fim, o subconjunto de dados de parentalidade, ou *pedigree*, contém apenas três variáveis: genótipo, que pode assumir 100 valores diferentes, *dam* e *sire*, termos tradicionais da

[2] https://github.com/cesar-galvao/TCC-Modelos-Multinivel/tree/main/dados%20simulatos%20resende%202021

área de melhoramento para se referir a filiação a uma fêmea e a um macho, respectivamente. Para esse conjunto de dados, 20 genótipos compõem uma geração de genitores e os demais compõem a geração filiada seguinte. Além disso, são consideradas apenas plantas alógamas, isto é, que dependem de fecundação de outra planta da mesma espécie e não podem se autofecundar.

3.3 Análise exploratória

Para avaliar a distribuição da variável resposta por célula experimental e por famílias, foram construídos intervalos para $\gamma = 0,95$ de confiança utilizando a expressão

$$IC(Y_j; \gamma = 0,95) = \bar{Y}_j \pm Z_{0,975} DP_{Y_j}, \tag{3.3.1}$$

em que $\bar{Y}_j$ é a média da variável resposta para um determinado agrupamento, Z é uma variável aleatória com distribuição Normal padrão avaliada em um determinado quantil e DP_{Y_j} é o desvio padrão da variável resposta do grupo avaliado em torno de sua média. A utilização de intervalos de confiança foi priorizada devido à grande quantidade de hipóteses sendo testadas simultaneamente e ao tamanho da amostra por vezes muito grande.

Para análise do ambiente, foram construídos gráficos para ilustrar a ocorrência de experimentos e distribuição de variáveis ambientais na região avaliada.

3.4 Geração de marcadores ambientômicos

A geração de marcadores ambientômicos é parte essencial do processo de análise ambientômica proposto por (RESENDE et al., 2021). É neste momento em que se obtém os marcadores a partir das covariáveis ambientais em um procedimento que é, essencialmente, uma forma de redução de dimensionalidade. Os marcadores são gerados da seguite forma:

- São selecionados entre 2 e 10 genótipos;
- A base de dados é filtrada para os genótipos selecionados e os dados resultantes são divididos entre partição de treino e partição de validação, com 50% dos dados para cada;
- É ajustado um modelo linear múltiplo com a partição de treino;
- Calcula-se a correlação entre o ajuste do modelo para a partição de validação e os dados da resposta nessa mesma partição. Essa correlação é registrada como um valor para o

vetor rgg. Os valores desses vetores são tomados como uma medida de qualidade de ajuste;

- É ajustado um modelo linear para todos os casos do item 2, sem particionamento;
- Usa-se o modelo do item 5 para realizar predições para todas as células disponíveis, incluindo aquelas em que não há experimento – esses valores são considerados marcadores ambientais;
- São selecionados como covariáveis finais os marcadores ambientais com rgg maiores que $0,5$.

3.5 Modelagem

O processo de modelagem considera como variável resposta a mesma Y representando produtividade e como covariáveis os marcadores ambientômicos gerados a partir da metodologia de bootstrap sugerida por (RESENDE et al., 2021). Dessa forma, é ajustado um modelo com intercepto e coeficientes aleatórios com apenas dois níveis: unidades experimentais como o nível mais baixo e genótipos aos quais essas pertencem como o nível mais alto.

Enquanto o modelo pode ser representado pela equação (2.2), a syntaxe do pacote lme4 para este tipo de modelo segue a forma `Yield  1 + (covar_1 + ... + covar_p | Genotype, data = dados)`. É importante frisar que quando o modelo é expresso conforme esta sintaxe, não há interação entre covariáveis.

O modelo apenas com interceptos também foi ajustado para obtenção da correlação intraclasse, expresso da seguinte forma: `Yield` $\sim$ `1 + (1 | Genotype, data = dados)`.

Para avaliação do modelo, foi utilizada apenas análise gráfica de intervalos de confiança para os estimadores, considerando sua distribuição teórica.

Referências

ADEWALE, A. J. et al. Understanding hierarchical linear models: applications in nursing research. *Nursing Research*, LWW, v. 56, n. 4, p. S40–S46, 2007. 12

CHENU, K. Characterizing the crop environment – nature, significance and applications. *Crop physiology*, Elsevier, p. 321–348, 2015. 9

GELMAN, A.; HILL, J. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. [S.l.]: Cambridge University Press, 2006. (Analytical Methods for Social Research). 12

HICKEY, L. T. et al. Breeding crops to feed 10 billion. *Nature biotechnology*, Nature Publishing Group US New York, v. 37, n. 7, p. 744–754, 2019. 9

HOX, J.; MOERBEEK, M.; SCHOOT, R. Van de. *Multilevel analysis: Techniques and applications*. [S.l.]: Routledge, 2017. 11, 12, 13

JORASCH, P. The global need for plant breeding innovation. *Transgenic Research*, Springer, v. 28, n. Suppl 2, p. 81–86, 2019. 9

KUTNER, M. H. et al. *Applied linear statistical models*. [S.l.]: McGraw-hill, 2005. 11

MCMAHON, S. M.; DIEZ, J. M. Scales of association: hierarchical linear models and the measurement of ecological systems. *Ecology letters*, Wiley Online Library, v. 10, n. 6, p. 437–452, 2007. 12

RESENDE, R. T.; BRONDANI, C.; CHAVES, L. J. O melhoramento na era de agricultura de precisão. In: RESENDE, R. T.; BRONDANI, C. (Ed.). *Melhoramento de Precisão*. Santo Antônio de Goiás, GO: Embrapa Arroz e Feijão, 2023. cap. 1, p. 13–40. 9

RESENDE, R. T. et al. Enviromics in breeding: applications and perspectives on envirotypic-assisted selection. *Theoretical and Applied Genetics*, Springer, v. 134, p. 95–112, 2021. 9, 10, 14, 15, 16