XXXIII Congresso de Iniciação Científica da Unicamp

Modelos de Regressão Logito e Probito com Efeitos Aleatórios para Dados Ordinais de Infecção por SCYLV em Cana-de-Açúcar

Palavras-Chave: Modelos-Cumulativos, SCYLV, Regressão-Ordinal, Genótipos, Efeitos-Aleatórios

Autores(as):
Carolina Tae Ishii Hara, IMECC, UNICAMP
Profa. Dra. Mariana Rodrigues Motta, IMECC, UNICAMP

INTRODUÇÃO:

Os modelos de regressão logito e probito para variáveis ordinais são amplamente utilizados em diversas áreas, como Ciências Sociais, para medir atitudes e opiniões, pesquisa de Marketing, para avaliar o grau de concordância em relação a afirmações, e Ciências Biológicas. Esses modelos possibilitam a análise de dados categóricos ordenados, incorporando efeitos fixos e aleatórios, sendo amplamente aplicados em estudos com resposta ordinal. A motivação deste estudo é a seleção de variedades de cana-de-açúcar resistentes ao Sugarcane yellow leaf virus (SCYLV), também conhecido como amarelinho. A doença, transmitida por afídeos, compromete a produtividade da cana e causa prejuízos econômicos ao setor agrícola. Modelos estatísticos serão empregados para prever a gravidade da infecção e classificar os genótipos mais resistentes, auxiliando programas de melhoramento genético. Assim, a estimação dos parâmetros foi realizada por meio da função clmm do pacote ordinal do R.

METODOLOGIA:

Conjunto de Dados

O conjunto de dados utilizado neste estudo foi obtido por meio de um experimento conduzido entre março de 2016 e abril de 2018 no Centro Avançado de Pesquisa Tecnológica no Agronegócio Canavieiro, localizado em Ribeirão Preto. O experimento foi estruturado com 97 genótipos distintos de cana-de-açúcar, organizados em três blocos completos casualizados. Como controle experimental, a cultivar SP71-6163 foi intercalada entre os genótipos avaliados.

As avaliações fenotípicas ocorreram em dois momentos distintos: na safra de cana-planta (junho de 2018) e na safra de cana-soca (julho de 2019). Em ambas as etapas, três avaliadores independentes atribuíram notas a cada planta com base na gravidade dos sintomas causados pelo vírus *Sugarcane yellow leaf virus* (SCYLV), causador da doença conhecida como "amarelinho".

As folhas analisadas foram as do topo visível do colmo (TVDLs), utilizando-se uma escala diagramática de severidade que varia de 1 a 4:

- **Nota 1**: folha verde, sem sintomas aparentes;
- **Nota 2**: leve amarelamento da nervura central e da lâmina foliar;
- **Nota 3**: amarelamento intenso da nervura central com amarelamento parcial da lâmina;
- **Nota 4**: amarelamento intenso tanto da nervura quanto da lâmina foliar.

Para efeitos deste trabalho, as avaliações dos sintomas foram realizadas separadamente para as fases de cana-planta e soca. Embora os dados tenham sido avaliados por três indivíduos, o efeito do avaliador não será incluído na modelagem estatística.

As variáveis consideradas para modelagem são:

- **Variável resposta**: nota ordinal de severidade da doença (1 a 4);
- **Covariáveis fixas**: ano de avaliação e bloco experimental;
- **Efeito aleatório**: genótipo da planta, representando sua constituição genética.

Modelagem Estatística

Neste estudo, utilizaremos modelos de regressão para variáveis categóricas ordinais com efeitos aleatórios, também conhecidos como **modelos cumulativos mistos**, tanto nas versões logito quanto probito. Esses modelos permitem modelar a probabilidade acumulada de uma resposta ordinal (nota da folha) em função de covariáveis explicativas, ao mesmo tempo em que incorporam variações associadas a fatores não observáveis, como os efeitos genéticos dos genótipos.

A formulação geral dos modelos é dada por:

- **Modelo Logito**:

$$\log\left(\frac{P(Y_{ik} \leq j)}{1 - P(Y_{ik} \leq j)}\right) = \alpha_j + \mathbf{x}_{ik}^{\top}\boldsymbol{\beta} + a_k$$

- **Modelo Probito**:

$$\Phi^{-1}(P(Y_{ik} \leq j)) = \alpha_j + \mathbf{x}_{ik}^{\top}\boldsymbol{\beta} + a_k$$

em que:

- Y_{Ik} é a nota atribuída à planta i do genótipo k;
- α_j representa os limiares cumulativos das categorias;
- x_{ik} corresponde ao vetor de covariáveis fixas (ano e bloco);
- β são os coeficientes associados às covariáveis;
- $ak \sim N(0, \sigma_2^{\alpha})$ é o efeito aleatório do genótipo k;
- $\Phi^{-1(.)}$ denota a inversa da função de distribuição normal padrão.

Estimação e Inferência

Como a distribuição marginal da resposta não possui forma fechada devido à presença de efeitos aleatórios, a estimação dos parâmetros será feita via **método da máxima verossimilhança**, utilizando aproximações numéricas para integração dos efeitos aleatórios. Será empregada a função clmm() do pacote **Ordinal** no ambiente R, que permite o ajuste de modelos lineares generalizados mistos com diferentes distribuições e funções de ligação.

Para a estimação dos efeitos aleatórios a_k, será utilizada a moda a posteriori condicional a partir dos dados observados e dos parâmetros estimados. A ordenação desses efeitos aleatórios permitirá identificar os genótipos mais resistentes à doença, contribuindo para programas de melhoramento genético.

A comparação entre os modelos ajustados (logito e probito) será realizada com base no **Critério de Informação Bayesiano (BIC)**.

RESULTADOS E DISCUSSÃO:

Os gráficos e tabelas apresentados a seguir, mostram os resultados obtidos até o momento presente da pesquisa.

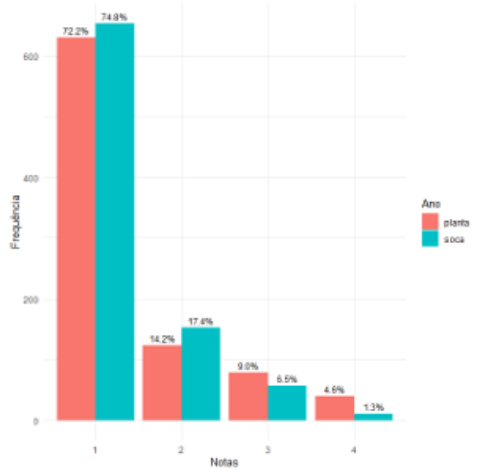

Figura 1: Distribuição das Notas por Soca e Ano.

A Figura 1 apresenta a distribuição das notas atribuídas ao nível da doença (SCYL) em relação às categorias Ano e Soca. Observa-se que as folhas classificadas como Soca tendem a receber mais frequentemente notas entre 1 e 2, enquanto as plantas da categoria Ano são predominantemente avaliadas com notas entre 3 e 4.

Ajuste Inicial com Dados Completos

Inicialmente, foram ajustados modelos cumulativos lineares (CLM), sem efeito aleatório, utilizando as funções de ligação logit e probit. O desempenho dos modelos foi avaliado por meio do Critério de Informação Bayesiano (BIC) e da significância dos efeitos fixos.

Tabela 1 - Resultados do Modelo CLM sem efeito Aleatório

Modelo	Link	BIC	Efeito de Ano (p-valor)	Efeito de Bloco (p-valor)
CLM	Logit	2900.7	0.080	0.095
CLM	Probit	2897.4	0.010	0.065

Observa-se que o modelo probit apresentou melhor ajuste (menor BIC) e efeito significativo do fator ano, indicando que as notas atribuídas variam de forma estatisticamente significativa entre os anos avaliados. O efeito de bloco se mostrou marginalmente significativo nos dois modelos.

Posteriormente, foram ajustados modelos cumulativos mistos (CLMM), incluindo o efeito aleatório de genótipos. Esses modelos capturam a variabilidade genética associada à resistência das plantas à doença.

Tabela 2 - Resultados do Modelo CLM com efeito Aleatório

Modelo	Link	Efeito Aleatório	BIC	Convergência	Significância de Ano e Bloco
CLMM	Logit	Genótipo	2376.2	Não convergiu	Não aplicável
CLMM	Probit	Genótipo	2385.6	Convergiu	Ambos significativos

Apesar do modelo logit apresentar menor BIC, ele não convergiu adequadamente. Já o modelo probit com efeito aleatório apresentou boa performance e convergência satisfatória, sendo o mais indicado para inferência.

Avaliação com Dados de Treinamento e Teste

A base de dados foi dividida em 80% para treinamento e 20% para teste. Os modelos foram ajustados ao conjunto de treinamento e avaliados em termos de acurácia no conjunto de teste.

Tabela 3 - Resultados dos Modelos Após Testes

Modelo	Link	Efeito Aleatório	BIC	Acurácia (%)	Convergência
CLM	Probit	Não	2331.86	73,6	Convergiu
CLM	Logit	Não	2328.02	73,6	Convergiu
CLMM	Logit	Genótipo	1971.58	74,8	Não convergiu
CLMM	Probit	Genótipo	1978.37	74,5	Convergiu

A inclusão do efeito aleatório resultou em modelos com melhor ajuste (menores valores de BIC) e ligeiro aumento na acurácia de predição. O modelo CLMM Probit combinou boa performance, maior estabilidade e convergência, sendo o modelo final escolhido para interpretação e seleção de genótipos resistentes.

Os resultados indicam que a severidade da doença do amarelinho é influenciada significativamente pelo **ano de avaliação** e pela **composição genética** das plantas. A modelagem por meio de modelos cumulativos mistos permite estimar a probabilidade de ocorrência de diferentes níveis de severidade, ao mesmo tempo em que considera a variabilidade entre os genótipos.

A inclusão do efeito aleatório de genótipos revelou-se essencial para capturar essa variabilidade genética, o que é fundamental para programas de melhoramento genético. Com base na predição dos efeitos aleatórios, torna-se possível classificar os genótipos de acordo com sua resistência à doença e, assim, orientar a seleção de variedades mais tolerantes ao SCYLV.

Além disso, a análise demonstra que o modelo com função de ligação probit foi o mais robusto entre os ajustados, oferecendo equilíbrio entre desempenho preditivo e estabilidade estatística.

CONCLUSÕES:

Os resultados obtidos neste estudo indicam que os modelos de regressão ordinal são ferramentas estatísticas eficazes para a análise da severidade da doença SCYLV em cana-de-açúcar. A comparação entre os modelos logito e probito, com e sem a inclusão de efeitos aleatórios, revelou que a consideração do efeito genotípico melhora o ajuste do modelo e a capacidade preditiva.

O modelo cumulativo misto com função de ligação probit apresentou melhor equilíbrio entre desempenho estatístico, estabilidade computacional e acurácia preditiva, sendo o mais indicado para fins de inferência. A inclusão do efeito aleatório de genótipos foi fundamental para capturar a variabilidade genética associada à resistência à doença, permitindo a identificação de materiais promissores para programas de melhoramento.

Dessa forma, conclui-se que a abordagem adotada é adequada para classificar genótipos quanto à severidade dos sintomas, contribuindo com ferramentas estatísticas robustas para a seleção de variedades mais tolerantes ao SCYLV.

BIBLIOGRAFIA

AGRESTI, Alan**.** ***Categorical Data Analysis***. Gainesville: University of Florida, 2002.

AGRESTI, Alan. ***Modeling Ordinal Categorical Data***. Department of Statistics, University of Florida, USA, 2013.

BURBANO, Roberto C. V. et al**. Screening of *Saccharum spp.* genotypes for sugarcane yellow leaf virus resistance by combining symptom phenotyping and highly precise virus titration**. *Crop Protection*, Elsevier, Londres, v. 15, n. 1, p. 47–61, 2021.

GONÇALVES, Marcos César. ***Cultura - Cana-de-açúcar (Saccharum sp.)***, Guia da Sanidade Vegetal. São Paulo: Instituto Biológico, 2016.

LEE, James. **Cumulative logit modelling for ordinal response variables: applications to biomedical research.** *CABIOS*, Londres, v. 6, p. 555–562, 1992.

LONG, J. S. ***Regression Models for Categorical and Limited Dependent Variables***. Thousand Oaks: Sage Publications, 1997.

OKURA, Roberta I. S.; ELIAN, Silvia N. ***Modelos de Regressão para Variáveis Categóricas Ordinais com Aplicações ao Problema de Classificação***. São Paulo: Instituto de Matemática e Estatística, USP, 2008.

PIMENTA, Ricardo J. G. et al. **Genome-wide approaches for the identification of markers and genes associated with sugarcane yellow leaf virus resistance**. *Nature*, Londres, v. 56, n. 4, p. 1030–1039, 2021.