Fol
11209

APLICAÇÃO DA ANÁLISE DISCRIMINANTE LINEAR NA IDENTIFICAÇÃO DOS COMPONENTES DO RENDIMENTO QUE LIMITAM A PRODUTIVIDADE DO FEIJÃO MACASSAR

Carlos Alberto V. Oliveira*
Evaristo Eduardo de Miranda*
José Wilton de Queiroz**

1- I N T R O D U Ç Ã O

Com vistas a determinar o que limita a produtividade das culturas a pesquisa agropecuária no Trópico Semi-Árido vem buscando conhecer o nível e a variabilidade dos rendimentos culturais nas propriedades rurais. Todavia esse conhecimento necessário não é suficiente sem a identificação e a classificação dos componentes do rendimento que mais contribuem em sua variabilidade.

Este trabalho, usando os dados obtidos pelo acompanhamento de 87 pequenas propriedades na região de Ouricuri (MIRANDA, 1), classificou pelo uso da função discriminante, os componentes do rendimento cultural que em ordem de prioridade mais contribuiram para a baixa produtividade do feijão macassar (Vigna unguiculata Walp.) na região estudada.

Esta classificação, além de contribuir para um melhor conhecimento da complexidade do meio rural, pode proporcionar subsídios para o planejamento da pesquisa agropecuária no Trópico Semi-Árido, indicando os pontos que devem ser prioritariamente abordados.

---

* Pesquisadores do CPATSA

** Professor Asistente da UFRN

2- METODOLOGIA

2.1) Do Local

Por razões logísticas e de diversidade de situações ecológicas, o trabalho foi realizado com base nos dados obtidos na região de Ouricuri-PE (mapa 1), onde o CPATSA/EMBRAPA possui atualmente cerca de 12 projetos de pesquisa através do Programa Nacional de Pesquisa 027 - "Avaliação dos Recursos Naturais e Sócio-Econômicos do Trópico Semi-Árido" (MIRANDA *et al.*, 1980). O trabalho foi desenvolvido com o conjunto de culturas alimentares existentes na região, mas neste trabalho, nos limitaremos ao feijão macassar (*Vigna unguiculata* Walp.).

2.2) Da Amostra

Já que o pesquisador não pode transportar ou reproduzir, a nível de campo experimental, a totalidade da realidade que lhe interessa, neste caso, ele é obrigado a examinar amostras limitadas do espaço rural.

No caso da escolha dos produtores a serem estudados, postulou-se que a produtividade e a produção da cultura do feijão estavam diretamente ligados aos sistemas de cultivo e de produção praticados e que estes sistemas variaram de um agricultor a outro em função de sua situação sócio-econômica e agroecológica. Essas situações foram levantadas e caracterizadas através de um trabalho de campo, envolvendo cerca de 100 variáveis que foram objeto de sínteses numéricas e gráficas.

Dessa síntese, definiu-se uma amostra de cerca de 87 unidades de produção, cujo conjunto de campos e parcelas foram acompanhados semanalmente por pesquisadores e técnicos do CPATSA desde o plantio até a colheita. Esse acompanhamento inclui uma série de observações qualitativas e quantitativas vinculadas ao clima, à planta cultivada, às adventícias, aos predadores, ao solo, às técnicas culturais praticadas pelo agricultor, etc. Esse conjunto de observações periódicas foi completado por informações obtidas junto ao produtor sobre os antecedentes e precedentes culturais de cada campo assim como sobre aspectos sócio-econômicos de suas estruturas de produção. No total obteve-se uma matriz de cerca de 50 variáveis (MIRANDA 1981).

2.3) Do Tratamento dos dados

A média dos rendimentos culturais do feijão, em quilos por hectare, na totalidade dos campos observados foi de :

259,5 Kg/ha $\pm$ 22,3 (214,8 ; 304,2)

Para analisar-se, numa primeira aproximação, os rendimentos do feijão, utilizou-se a equação lógica do rendimento onde:

$$\text{Redt Kg/ha} = (\text{Número de plantas/ha}) \times (\text{Número de vagens/planta}) \times (\text{Número de grãos/vagem}) \times (\text{Peso médio de um grão}) \quad (2.3.1)$$

Os componentes do rendimento constituindo o segundo membro da equação (2.3.1) são variáveis doravante denotadas por x_1, x_2, x_3 e x_4, segundo a ordem em que se apresentam. Tal equação, simples e multiplicativa, realiza-se a nível de campo ao longo do tempo. Assim, cada estado ou fase do ciclo vegetativo da planta determina o valor de cada uma das variáveis. É o que chamaremos de elaboração do rendimento (veja FIGURA - 2).

Para efeito de análise, a região estudada foi dividida em duas populações. A primeira, constituindo-se de campos com produtividade inferior a 300 Kg/ha,

denominada "população de baixa produtividade", e a segunda, constituindo-se de campos com produtividade superior a 300 Kg/ha, denominada "população de média produtividade". Tomando-se uma amostra aleatória de cada população arrolou-se 55 campos na primeira e 32 campos na segunda.

2.4) Do Modelo

A análise discriminante consiste em determinar uma regra que nos permita classificar uma unidade amostral (campo ou propriedade), à partir do vetor de características observado (componentes de produção), em uma das populações consideradas levando em conta uma minimização do risco que, se tem em proceder uma classificação errônea. Além disso, ela nos permite hierarquizar as variáveis componentes segundo a contribuição de cada uma no processo de classificação (Singh, 4).

A análise discriminante assume que:

1º - O vetor x das características tem distribuição normal multivariada.

2º - A matriz de variância-covariância da 1ª população é igual à matriz de variância-covariância da 2ª população.

3º - As duas populações diferem quanto à seus vetores de médias (Morrison, 1).

Satisfeitas as três condições acima, no sentido de se construir uma regra de classificação, parece intuitivo determinar uma combinação linear das carac-

terísticas observadas

$$Z = \alpha_1 x_1 + \alpha_2 x_2 + \ldots + \alpha_p x_p \qquad (2.4.1)$$

chamada "função discriminante". Em seguida, classificar x na primeira população se $Z \leq C$ ou classificar x na segunda população se $Z > C$, sendo $\alpha_1, \alpha_2, \ldots \alpha_p$ e C constantes reais.

Os valores numéricos para os αi são obtidos ao solucionar-se o sistema de equações lineares

$$\begin{aligned} \alpha_1 \ \sigma i^2 + \alpha_2 \ \sigma_1 \ 2 + \ \ldots \ + \alpha_k \ \sigma_1 \ k &= \mu_1 \ 1 - \mu_1 \ 2 \\ \vdots \qquad\qquad \vdots & \\ \alpha_1 \ \sigma_k \ 1 + \alpha_2 \ \sigma_k \ 2 + \ \ldots \ + \alpha_k \ \sigma_k^2 &= \mu_k \ 1 - \mu_k \ 2 \end{aligned} \qquad (2.4.2)$$

solucionando αi desta maneira maximizaremos o quadrado da diferença entre as médias das observações transformados pela unidade de sua variância. Se o quadrado de sua diferença é um máximo, também o será a diferença por unidade de dispersão.

Em síntese, a equação (2.4.1) pode equivalentemente ser escrita como

$$Z = \alpha' x \qquad (2.4.3)$$

e a regra de classificação sendo a seguinte:

$$\begin{cases} x \text{ pertencente à } 1^{a} \text{ população se } Z \leq C \\ x \text{ pertencente à } 2^{a} \text{ população se } Z > C \end{cases} \qquad (2.4.4)$$

Adotando a regra acima podemos ocasionalmente cometer dois tipos de erro:

a) Classificar x na 2^{a} população quando, na realidade, x pertence à 1^{a}.

b) Classificar x na 1^{a} população quando, realmente, x pertence à 2^{a}.

Estes erros ocorrem com probabilidades simbolizadas por P (2/1) e P (1/2), respectivamente, (veja FIGURA 2).

O valor crítico C é determinado de modo a minimizar a soma P (2/1) + P (1/2) enquanto que o vetor α é calculado de modo a maximizar o poder de discriminação da função (2.4.3).

Sob as três hipóteses assumidas neste parágrafo, a função discriminante (2.4.3) será uma variável aleatória com distribuição normal univariada cuja média é

$$\gamma_1 = \alpha' \mu^{(1)}, \quad \text{se } x \in \text{à } 1^{a} \text{ população}$$

ou

$$\gamma_2 = \alpha' \mu^{(2)}, \quad \text{se } x \in \text{à } 2^{a} \text{ população}$$

Sua variância é dada por

$$\sigma_z^2 = \alpha' \Sigma \alpha ,$$

onde Σ é a matriz de variância-covariância comum às duas populações. Assim, a função discriminante transfere a informação contida nas duas populações multivariadas para duas populações univariadas (veja FIGURA - 2).

Mediante um exame da FIGURA - 2, pode-se perceber, intuitivamente, que o valor adequado de C, no sentido de minimizar a soma P (2/1) + P (1/2), será a média entre γ_1 e γ_2, isto é,

$$C = \frac{\gamma_1 + \gamma_2}{2} \qquad (2.4.5)$$

como os vetores de médias populacionais $\mu^{(1)}$ e $\mu^{(2)}$ são desconhecidos, torna-se impossível o cálculo exato de C pela fórmula (2.4.5) contudo, uma estimativa não viciada para este valor crítico pode ser obtida por

$$\hat{C} = \frac{\bar{Z}_1 + \bar{Z}_2}{2},$$

onde $\bar{Z}_1 = \alpha' \bar{x}^{(1)}$ e $\bar{Z}_2 = \alpha' \bar{x}^{(2)}$

Testar o poder discriminante da função (2.4.3) e testar as hipóteses Ho: $\gamma_1 = \gamma_2$ versus H_1: $\gamma_1 \neq \gamma_2$ são coisas equivalentes. Na prática, a matriz Σ frequentemente desconhecida, é estimada pela matriz de variância-covariância amostral, S, calculada com base na amostra combinada. Uma estatística para o teste das hipótese acima será

$$t^2(\alpha) = \frac{N_1 N_2}{N_1 + N_2} \frac{[\alpha' (\bar{x}^{(1)} - \bar{x}^{(2)})]^2}{\alpha' S \alpha} \qquad (2.4.6)$$

Dizemos que (2.4.3) tem um grande poder discriminante se Ho for regeitado a um nível de significancia, mostrando que γ_1 e γ_2 estão suficientemente afastados. Naturalmente este poder é máximo quando o vetor α maximiza a $t^2(\alpha)$.

Antes de encontrar a expressão de α que maximiza $t^2(\alpha)$ deve-se notar que (2.4.6) não é afetada por mudanças na escala de α, portanto, podemos resumir o problema em maximizar

$$[\alpha' (\bar{x}^{(1)} - \bar{x}^{(2)})]^2 \qquad (2.4.7)$$

sujeito à restrição

$$\alpha' S \alpha = 1 \qquad (2.4.8)$$

Usando multiplicador de Lagrange mostra-se que o máximo de (2.4.7) com respeito à α sujeito à restrições (2.4.8) é a "distância de Mahalanobis, dada por

$$D^2 = (\bar{x}^{(1)} - \bar{x}^{(2)})' S^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}) \qquad (2.4.9)$$

que mede o afastamento entre duas populações, enquanto que α é dado pela expressão

$$\alpha = S^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}) \qquad (2.4.10)$$

Vale salientar que, embora determinada a expressão do vetor α com o intuito de maximizar o poder discriminante da função (5.2), este esforço seria em vão caso a 3ª hipótese não se verificasse. Portanto, antes de levarmos à efeito a análise discriminante, necessitamos testar as hipóteses:

$$Ho: \mu^{(1)} = \mu^{(2)} \quad \text{versus} \quad H_A: \mu^{(1)} \neq \mu^{(2)} \qquad (2.4.11)$$

A estatística de teste, também denominada "estatística T^2 de Hotelling", baseia-se na distância de Mahalanbis e sua expressão é

$$T^2 = \frac{N_1 N_2}{N_1 + N_2} (\bar{x}^{(1)} - \bar{x}^{(2)})' S^{-1} (\bar{x}^{(1)} - \bar{x}^{(2)}) \qquad (2.4.12)$$

A quantidade de $F = \dfrac{N_1 + N_2 - p - 1}{(N_1 + N_2 - 2)\, p} T^2 \qquad (2.4.13)$

tem distribuição F - Fisher-Snedecor com p e $N_1 + N_2 - p - 1$ graus de liberdade. A um nível de significância pré-estabelecido, rejeita-se a hipótese Ho em (2.4.11) se o valor de F calculado segundo (2.4.13) for maior que o valor tabelado. A rejeição de Ho revela um afastamento significativo entre as duas populações, tornando possível o uso da análise discriminante. Porém, não estabelece a contribuição de cada característica para este afastamento, o qual poderia ser causado por apenas algumas dessas características podendo as demais serem descartadas da análise. Um exame neste sentido pode ser feito através da construção de "intervalos de confiança simultâneos" para as componentes do vetor diferença $\mu^{(1)} - \mu^{(2)}$. Intervalos desse tipo, com coeficiente de confiança conjunto 1-B,

são dados por

$$(\bar{x}_i^{(1)} - \bar{x}^{(2)}) \pm \sqrt{S^2 i \left(\frac{N_1 + N_2}{N_1 N_2} - \frac{P(N_1 + N_2 - 2)}{N_1 + N_2 - P - 1} F_{p_1 (N_1 + N_2 - P - 1), B}\right)} \qquad (2.4.14)$$

onde S_i^2 é a variância conjunta da i-ésima característica (elemento Sii da matriz S). As características descartáveis são aquelas cujos intervalos possuem o zero como elemento.

3- Resultados e Discussão

Conforme os dados referentes às populações em estudo, a estatística F calculada por (2.4.12) e (2.4.13) resultou em 112,70. O valor tabelado correspondente a 4 e 82 graus de liberdade e ao nível de significância de 5% é 2,50, logo, menor que o valor calculado. Por conseguinte, a hipótese de igualdade entre os vetores de médias das populações foi rejeitada indicando que as características consideradas são úteis para a classificação das propriedades nos grupos de baixa e média produtividade. Entretanto, calculados os "intervalos de confiança simultâneos" conforme a expressão (2.4.14) e um coeficiente de confiança conjunto de 95% comprovou-se a irrelevante participação das variáveis x_1 (Nº plantas/ha) e x_3 (Nº grãos p/vagem) no afastamento de dois grupos e, consequentemente, no processo de classificação (veja TABELA - 3.1).

Considerando então as variáveis x_2 e x_4, a função discriminante ajustada foi:

$$Z = 1,30\ x_2 + 4,1\ x_4 \qquad (3.1)$$

e $\hat{C} = 17,68$

logo, classificaremos a i-ésima propriedade no grupo de baixa produtividade se $Z_i \leq 17,68$ e no grupo de média produtividade se $Z_i > 17,68$.

Obviamente, o i representa o incremento na função discriminante Z quando au-

TABELA - 3.1. Intervalos de confiança simultâneos para diferenças de médias. Coeficiente de confiança conjunto: 95%

VARIÁVEIS	INTERVALOS	
Nº planta/ha	- 36,3450 a 91,9850	(*)
Nº vagem p/planta	8,1700 a 14,4420	
Nº grãos p/vagem	- 1,1085 a 0,9725	(*)
Peso médio/grão	0,0017 a 0,0183	

(*) Diferença não significativa entre as médias. O zero pertence ao intervalo.

mentamos a variável x_i em uma unidade. Se as variáveis tivessem o mesmo desvio-padrão, o que não é o caso, a importância de cada uma no mecansimo de classificação seria proporcional ao valor absoluto de seu coeficiente. Todavia, é intuitivo perceber que entre duas variáveis com coeficiente de mesmo valor absoluto, a mais importante será aquela de maior variabilidade. Assim, uma maneira de hierarquiza-las é ordena-las segundo os valores absolutos produtos de seus coeficientes por seus respectivos desvios-padrões (veja TABELA - 6.2).

TABELA - 3.2. Hierarquização das variáveis segundo o Produto de seus desvios-padrões por seus coeficientes.

VARIÁVEIS	DESVIOS-PADRÕES (Si)	COEFICIENTES	|Si αi|
Nº vagem p/planta	8,74	1,3	11,3
Peso médio/grão	0,03	4,1	0,123

Em termos percentuais, a contribuição de uma variável x_i naturalmente expressa por $|Si\ \alpha|$ pode ser assim calculada:

$$a_i = 100 \frac{|Si \alpha i|}{\Sigma |Si\alpha i|}$$

De acordo com a última coluna da TABELA - 6.2 o nº de vagens p/planta possui um poder de classificação em torno de 98% enquanto o peso médio/grão possui um poder de 2%, aproximadamente.

7- Conclusão

A produtividade extremamente baixa do feijão na região de Ouricuri, quando analisada através da equação do rendimento, indica que um aumento de qualquer componente do rendimento se traduzira sem dúvida num aumento de produção. Todavia a identificação do número de vagens por planta, através da função discriminante, como o componente do rendimento que mais pesa na variabilidade atual da produtividade, permite de indicar uma meta prioritária para a intensificação dessa agricultura