O USO DE VARIÁVEIS "DUMMY" EM SUPERFÍCIES DE RESPOSTA À ADUBAÇÃO (*)

Robério Ferreira dos Santos (**)

SINOPSE

Neste trabalho foi mostrado o uso mais simples de variáveis **"dummy"**, permitindo dois diferentes interceptos para duas localidades diferentes, bem como a generalização para o caso em que as variáveis **"dummy"** são utilizadas, permitindo diferentes interceptos e diferentes inclinações para diferentes localidades e diferentes períodos de tempo.

Os testes de igualdade de variância de Goldfeld-Quandt (para dois grupos de dados), são também apresentados.

SUMMARY

In this paper the use of dummy variables is shown allowing two different interpcepts for two different locations. After this, the use of dummy variables allowing different intercepts and the use of interaction terms allowing different slopes are shown for different locations and different growing seasons.

The use of the Goldfeld-Quandt test of difference of variances for two groups, and of the Bartlett test for two or more data groups is ilustrated.

(*) O Autor deseja agradecer as valiosas sugestões feitas pelos colegas Vitor Afonso Hoeflich, Levon Yeganiantz, Hernán R. Tejeda e Helio Tollini, na versão preliminar desse trabalho. Apresentado na XV Reunião Anual da SOBER, Viçosa 1977.

(**) Técnico da EMBRAPA—DDM (atualmente no GEIPOT - DATE).

O USO DE VARIÁVEIS "DUMMY" EM SUPERFÍCIES DE RESPOSTA À ADUBAÇÃO

Robério F. dos Santos

1. INTRODUÇÃO

Suponha que um pesquisador esteja especificando um modelo na forma $\tilde{Y} = \tilde{X}\tilde{\beta} + \tilde{\varepsilon}$, onde as variáveis podem ser medidas quantitativamente,

sendo:

$$\tilde{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix}, \qquad \tilde{X} = \begin{bmatrix} 1 & X_{21} & \cdots & X_{K1} \\ 1 & X_{22} & \cdots & X_{K2} \\ \cdot & & & \\ \cdot & & & \\ \cdot & & & \\ 1 & X_{2n} & \cdots & X_{kn} \end{bmatrix}$$

$$\tilde{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \cdot \\ \cdot \\ \cdot \\ \beta_K \end{bmatrix}, \qquad \tilde{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

Suponha que o pesquisador esteja especificando um modelo para obter uma superfície de resposta à adubação. Pode acontecer que ele tenha dados de resultados de experimentos referentes a diferentes períodos de tempo e/ou diferentes locais.

Claro que o pesquisador gostaria de incluir estes fatores, tempo e locais diferentes, uma vez que entre outras coisas, ele estaria agregando ao seu estudo fatores como diferenças climáticas e diferentes qualidades de solos. Ele poderia então testar a importância do clima e da qualidade do solo nas respostas à adubação. Estas diferenças podem ser examinadas rodando diferentes regressões para cada categoria em separado e testando as diferenças pelo uso, por exemplo, do teste de Chow ((1), (2)). Um procedimento muito simples é usar variáveis **"dummy"** para representar as diferentes categorias ((7), (8)).

Este trabalho é organizado da seguinte maneira. A seção 2 é dedicada ao uso de variáveis **"dummy"** quando o pesquisador pretende incluir, no modelo a ser especificado, as variáveis qualitativas, como diferentes localidades e diferentes períodos de tempo. A seção 3 apresenta duas restrições necessárias ao pesquisador quando da utilização de variáveis **"dummy"** e apresenta e discute dois testes de homocedasticidade: um para dois grupos de observações e outro para "n" grupos de observações.

2. USO DE VARIÁVEIS "DUMMY"

Estimativas não viesadas podem ser obtidas mesmo com a utilização de variáveis **"dummy"** para representar as diferentes categorias, já que as estimativas dos coeficientes das variáveis **"dummy"** amoldam-se a qualquer situação. Convém salientar que o uso de variáveis **"dummy"** refere-se somente às variáveis pré-determinadas. Se uma variável endógena for qualitativa (não pode ser medida em termos quantitativos), então o uso de regressão múltipla não será apropriado. Nestes casos, a técnica conhecida como análise discriminante múltipla será usada. Se a variável endógena assumir apenas os valores zero e um, então a técnica conhecida por análise **Probit** será relevante (4).

Considere que um pesquisador esteja especificando um modelo que visa a explicar resposta à adubação para determinado produto. Suponha que o modelo especificado foi:

$$Y = \alpha_1 + \alpha_2 X_2 + \ldots + \alpha_K X_K + \varepsilon$$

Suponha agora que o mesmo pesquisador tenha resultados de experimentos realizados em dois lugares diferentes, A e B. Ele poderia então definir duas variáveis, S_1 e S_2, onde S_1 tem valor um, se o local é A, e zero, se o local é B, e o modelo ficaria então especificado como:

$$Y = \alpha_1 + \alpha_2 X_2 + \ldots \alpha_K X_K + \beta_1 S_1 + \beta_2 S_2 + \varepsilon$$ ou, em termos matriciais, $Y = \tilde{X}\tilde{\psi} + \tilde{\varepsilon}$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \cdot \\ \cdot \\ \cdot \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & \cdots & X_{K1} & 1 & 0 \\ 1 & X_{22} & \cdots & X_{K2} & 1 & 0 \\ \cdot & \cdot & \cdots & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot \\ \cdot & \cdot & & \cdot & \cdot & \cdot \\ 1 & X_{2n} & \cdots & X_{Kn} & 0 & 1 \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \cdot \\ \cdot \\ \cdot \\ \alpha_K \\ \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_n \end{bmatrix}$$

Na matriz X pode-se verificar que a soma das duas últimas colunas iguala-se com a primeira, ou seja, o posto da matriz não é completo. Como conseqüência $(\tilde{X}' \tilde{X})^{-1}$ não existe e não podemos obter $\tilde{C} = (\tilde{X}'\tilde{X})^{-1}\tilde{X}'\tilde{Y}$, que é um estimador de $\tilde{\Psi}$ (7).

Para a obtenção de uma solução, é conveniente eliminar uma das variáveis **"dummy"**, suponha S_2. Com isto não ocorre nenhuma perda de informação, já que a variável **"dummy"** restante (S_1) continua identificando o lugar A, com o valor um, e o lugar B, com o valor zero. O intercepto da equação α_1 corresponde a testemunha para o lugar B e o coeficiente β_1 mostra a diferença existente entre as testemunhas das duas localidades.

Em algumas especificações a teoria que está por trás do modelo sugere que o uso de variáveis **"dummy"** deve permitir não somente diferentes interceptos, mas também diferentes inclinações nas relações entre as variáveis dependentes e independentes para as diferentes categorias.

Em termos do exemplo acima apresentado, o pesquisador pode esperar que, em razão de diferentes qualidades de solos nas localidades A e B, variações iguais nas duas localidades no nível de utilização da variável X_2 conduzam a diferentes respostas na variável dependente. Esta expectativa do pesquisador pode ser testada com a especificação do modelo:

$$Y = \alpha_1 + (\alpha_2 + \delta_1 S_1) X_2 + \ldots + \alpha_K X_K + \beta_1 S_1 + \varepsilon,$$

onde a variável **"dummy"** S_1 continua a assumir o valor um para a localidade A e o valor zero para a localidade B. A influência da localidade, na época da aplicação de

X_2, na variável Y pode ser testada com a aplicação de um teste de significância no coeficiente δ_1 ((6), (7)).

De modo análogo, a análise poderia ser expandida para abranger tanto o caso em que a expectativa do pesquisador de que as inclinações de todos X_k, k=2 ... k, fossem diferentes para as duas localidades, como no caso da existência de mais de duas localidades e, inclusive outras, categorias (como períodos de tempo diferentes). Convém sempre lembrar que, para a existência de uma solução, devem ser criadas tantas variáveis **"dummy"** quanto seja o número de categorias existentes menos um.

Para especificação de um modelo geral, admita-se que um pesquisador tenha dados de experimentos com adubação realizados em "n" lugares diferentes e em "t" períodos de tempo. Admita-se que ele esteja trabalhando com "K" variáveis independentes. O modelo ficaria então especificado como:

$$Y_{ij} = \alpha_1 + \left(\alpha_k + \sum_{i=1}^{n-1} \delta_i R_{ij} + \sum_{j=1}^{t-1} \lambda_j S_{ij}\right) X_{ij,k} + \sum_{i=1}^{n-1} \beta_i R_{ij} + \sum_{j=1}^{t-1} \Psi_j S_{ij} + \varepsilon_{ij}$$

onde:

Y_{ij} = variável dependente, i = 1,2,..., n e j = 1,2,... t;

$X_{ij,K}$ = variável independente, K = 2,3,...K;

$R_{ij} = W_L$, sendo, $W_L = 1$ para L = i, i = 1,2, ... n−1 e $W_L = 0$ para $L \neq i$;

$S_{ij} = Z_M$, sendo, $Z_M = 1$ para M = j, j = 1,2, ..., t−1, e $Z_M = 0$ para $M \neq j$;

ε_{ij} = termo estocástico.

3. RESTRIÇÕES AO USO DE VARIÁVEIS "DUMMY"

A primeira restrição com a qual o pesquisador se defronta quando do uso de variáveis **"dummy"** diz respeito ao número de graus de liberdade. Cada variável **"dummy"**, usada para permitir diferenciações de interceptos e/ou inclinações, re-

presenta um outro coeficiente a ser estimado. O pesquisador deve estar certo de que o número total de observações permanece maior que o número de parâmetros a serem estimados. Esta é uma condição mínima para a existência de um número positivo de graus de liberdade. No caso do úlitmo modelo especificado anteriormente, por exemplo, o número de graus de liberdade é igual a nt − ((n+t−1)K + (n+t−2)). Para que a condição acima seja satisfeita, é necessário que nt > (n+t−1)K + (n+t−2).

Ao dispor de dados de experimentos realizados em várias localidades diferentes, nem sempre o pesquisador pode criar variáveis **"dummy"**, ao especificar o seu modelo, visando a testar a influência da diferenciação da localidade na variável dependente. Ele tem uma restrição. Precisaria saber se os dados que ele pretende agregar têm pelo menos uma característica em comum, podendo ser esta característica o intercepto, a inclinação e/ou a variância.1/

Os testes de igualdade dos interceptos e das inclinações podem ser realizados após a especificação do modelo, bastando testar a significância dos coeficientes das variáveis **"dummy"**. O que precisaria ser testado a **"priori"** seriam, pois, as variâncias dos dados que se pretende agregar.

Se o pesquisador estiver trabalhando apenas com duas categorias, como localidades A e B por exemplo, ele poderia testar a igualdade das variâncias dos dois grupos de dados usando o teste Goldfeld-Quandt ((3), (7)). Para a aplicação deste teste primeiramente, para satisfazer a condição de independência, devem-se encontrar as variâncias dos grupos de resíduos utilizando regressões separadas para cada grupo de observações. A primeira regressão é baseada em N_A observações e a segunda em N_B observações. A soma dos quadrados dos resíduos para cada regressão é uma forma quadrática, ($(\tilde{e}'\tilde{e})_A = \tilde{Q}_A$) e ($(\tilde{e}'\tilde{e})_B = \tilde{Q}_B$). Cada rateio $\tilde{Q}/\sigma^2_\varepsilon$ tem uma distribuição qui-quadrado com (N_A-K) e (N_B-K) graus de liberdade, respectivamente2./ Estas duas variáveis qui-quadrado são independentes se Var-Covar $(\underset{n}{\varepsilon}) = \sigma^2_\varepsilon \tilde{I}$, onde $\tilde{I}$ é uma matriz identidade nXn. Portanto, quando Var $(\underset{n}{\varepsilon})$ é constante para

---

1/ No caso em que os grupos de dados são heterocedásticos surgem problemas no que se refere ao método de estimação a ser utilizado. Se o método dos mínimos quadrados for utilizado, então os estimadores obtidos não serão eficientes, ou seja, as variâncias dos estimadores obtidos pelo método dos mínimos quadrados são maiores do que as variâncias dos estimadores obtidos por algum outro processo de estimação que leve em conta a informação adicional sobre a diferença de variâncias. Ver a respeito, por exemplo, MURPHY (7), p. 298-307.

2/ Onde, $\tilde{e}$ é um valor de resíduos de mínimos quadrados; ϕ^2_ε é a variância do termo estocástico do modelo econométrico especificado e K o número de parâmentros do modelo.

todos "n", uma estatística distribuída de acordo com a distribuição "F" pode ser criada como:

$$F = \frac{(\tilde{e}'\tilde{e})_A/N_A - K}{(\tilde{e}'\tilde{e})_B/N_B - K}$$

Se as observações forem ordenadas de tal modo que o grupo com maior variância residual seja conhecido "a priori" e designado como grupo "A" no numerador, então a hipótese de que as variâncias dos dois grupos sejam iguais, testada contra a hipótese de que a variância do grupo A seja maior, será aceita quando $F < F(N_A - K, N_B - K)$, com o nível α% de significância.

Quando há W grupos de observações independentemente distribuídos, com $W > 2$ o teste conhecido pelo nome de Bartlett pode ser utilizado para testar a hipótese de que os grupos de observações foram tirados de populações com variâncias iguais (1).

Considerando N_t ($t = 1,2, \ldots, W$) o número de observações no t-ésimo grupo; Y_{ti} ($i = 1,2, \ldots, N_t$) a i-ésima observação do t-ésimo grupo, $Y_{ti} = Y_{ti} - Y_t$, sendo Y_t a média das observações do t-ésimo grupo;

$S_t^2 = \sum_{t=1}^{nt} y_{ti}^2/\phi_t$ a estimativa da variância do t-ésimo grupo, tendo $t = nt - 1$ graus de liberdade; $S^2 = \sum_{t-1}^{W} \phi_t S_t^2/\phi$ a média das variâncias estimadas, sendo $\phi = \Sigma_{\phi t}$ o número total de graus de liberdade, pode-se definir a estatística $M_1 = \phi \operatorname{Ln} S^2 - \sum_{t=1}^{W} \phi_t \operatorname{Ln} S_t^2$, utilizada por Bartlett para testar homogeneidade de variâncias de populações diferentes. Pode ser mostrado que quando a apropriada hipótese nula é verdadeira ($\sigma_1^2 = \sigma_2^2 = \ldots = \sigma_W^2$) e sob a pressuposição de populações normalmente distribuídas, nas grandes amostras M_1 é distribuída como uma distribuição qui-quadrado com W−1 graus de liberdade (X_{W-1}^2),(1). Bartlett mostrou que para pequenas amostras, com boa aproximação, M_1 é distribuída como $(1 + A)\, X^2_{W-1}$, onde A, uma constante de ajustamento que tende para zero para grandes valores de ϕ_t, é definida como:

$$A = (1/3(W-1)) \sum_{t=1}^{W} (1/\phi t - 1/\phi)^{3/}.$$

3/ Para um uso prático deste teste, ver OSTLE (7), pp. 136-8. Para restrições ao seu uso, ver BOX (1).

4. RESUMO E CONCLUSÕES

Ao especificar as superfícies de resposta à adubação, o pesquisador defronta-se, muitas vezes, com a necessidade de incluir na função os efeitos de, por exemplo, locais e períodos de tempo diferentes.

Este trabalho pretendeu mostrar o uso de variáveis **"dummy"** em situações semelhantes à sugerida acima. Iniciou-se com o uso mais simples de variáveis **"dummy"**, permitindo dois diferentes interceptos para duas localidades diferentes, generalizando-se depois para o caso em que as variáveis **"dummy"** são utilizadas, permitindo diferentes interceptos e diferentes inclinações para diferentes localidades e diferentes períodos de tempo.

Uma das exigências para o uso de variáveis **"dummy"** é que os grupos de dados que se pretende agrupar devem ter uma característica em comum, podendo ser esta característica o intercepto, a inclinação e/ou a variância. Os testes de diferenças de inclinações podem ser realizados depois da especificação dos modelos, usando variáveis **"dummy"**. Ficaria pois apenas o teste de igualdade de variâncias para ser realizado a **"priori"**. Neste trabalho, foram apresentados os testes de igualdade de variâncias de Goldfeld-Quandt (para dois grupos de dados) e o de Bartlett (para dois ou mais grupos de dados).

5. LITERATURA CITADA

1. BOX, G.E.P. **"Non-Normality and Tests on Variances"**. Biometrika, 40, 1953 p. 318-35

2. CHOW, Gregory C. **"Tests of Equality between Sets of Coefficients in two Linear Regressions"**. Econométrica, 28, julho 1960, p. 591-605

3. FISHER, F.M. **"Tests of Equality Between Sets of Coefficients in Two Linear Regressions"**: "An Expository Note". Econométrica, Março 1970, pp. 361–66.

4. GOLDBERGER, A.S. **Econometric Theory,** New York, John Wiley &Sons, INC, 1964, p. 250-51

5. GOLDFELD, S.M. &QUANDT, R.E. **"Some Tests for Homocedasticity"**. Journal of the American Statistical Association, June 1965, p. 539-47.

6. JOHNSTON, J. **Econometrics Methods.** New York. McGraw-Hill, 1963, 300 p.

7. MURPHY, James L. **Introductory Econometrics,** Homewood, Richard D. Irwin, 1973, 524 p.

8. OSTLE, Bernard. **Statistics in Research.** Ames, The Iowa State University Press, 1964, 585 p.

9. SEARLE, S.R. **Linear Models.** New York, John Willey & Sons, 1971. 532 p.

10. SUITS, Daniel B. **"Use of Dummy Variables in Regression Equations.** Journal of the American Statistical Association, 52, 1957. p. 548-51