UM MODELO LINEAR GERAL
MULTIVARIADO NÃO–PARAMÉTRICO

Belmer Garcia Negrillo

RELATÓRIO TÉCNICO Nº 21/90

Resumo: Hattmansperger e McKean (1977, 1983, 1984) apresentaram um modelo linear geral univariado, baseado em postos (G LM–R) Pirie e Rauch (1984) provam que são mais eficientes que o modelo linear geral (GLM), quando os erros não tem distribuição normal, neste trabalho generalizamos o GLM–R multivariado, para várias funções escores.

Abstract: Hattmansperger and McKean (1977, 1983, 1984) show a univariate general linear model based on ranks (GLM–R), Pirie and Rauch (1984) show that they are more efficient that the general linear model (GLM), when the errors are not normaly distribuited, this paper generalizes that GLM–R to the multivariate case for several score functions.

Instituto de Matemática, Estatística e Ciência da Computação
Universidade Estadual de Campinas
13.081, Campinas, S.P.
BRASIL

Maio – 1990

Um Modelo Linear Geral Multivariado Não–Paramétrico

Belmer Garcia Negrillo

IMECC/UNICAMP

Resumo: Hattmansperger e McKean (1977, 1983, 1984) apresentaram um modelo linear geral univariado, baseado em postos (G LM–R) Pirie e Rauch (1984) provam que são mais eficientes que o modelo linear geral (GLM), quando os erros não tem distribuição normal, neste trabalho generalizamos o GLM–R multivariado, para várias funções escores.

1.- Introdução:

Consideremos o modelo linear geral multivariado

$$X_i = B_0 + B(C_i - \overline{C}) + e_i \qquad i = 1 \ldots N \tag{1}$$

onde $B = (B_1, \ldots, B_q)$ é uma matriz pxq e B_o é um p–vetor, de parâmetros desconhecidos

$C_i = (C_{i1}, \ldots, C_{iq})'$ são q–vetores especificados

e_i são v. a. i. i. d. com função de distribuição F continua.

Nosso maior interesse é estimar e testar hipóteses acerca de B, para isso, primeiramente vamos definir a estatística linear de postos, que apresentaremos a seguir.

$$Seja \quad d_i = N^{-1/2}(C_i - \overline{C}) = (d_{i1}, \ldots, d_{iq})', \qquad i = 1, \ldots, N \qquad (2)$$

e

$$D = \sum_{i=1}^{N} d_i d_i' = \left(\left(\sum_{i=1}^{N} d_{ik} d_{ik'}\right)\right), \quad k, k' = 1, \ldots, q$$

Assim o posto de D é q.

Por outro lado temos que $X_i = (X_{i1}, \ldots, X_{ip})'$, $i = 1, \ldots N$, e se R_{ij} é o posto de X_{ij} em $(X_{1j}, \ldots, X_{Nj})$ para $i = 1, \ldots, N$ e $j = 1, \ldots, p$ e utilizando as funções de seleção de escores φ_1 e φ_2 (Negrillo(1989)), selecionamos para cada variável a função escore ϕ_j, obtendo o conjunto de escores

$$a_j(i) = \phi_j(i/N + 1) \qquad (3)$$

para $j = 1, \ldots, p$
A estatística linear de postos para a j–ésima variável e o k–ésimo coeficiente é dada por

$$S_{jk} = \sum_{i=1}^{N} d_{ik}\ a_j\ (R_{ij}) \qquad j = 1 \ldots p, \quad k = 1, \ldots, q \qquad (4)$$

com a matriz

$$S = (S_1, \ldots, S_q) = ((S_{jk})) \;\; j = 1, \ldots, p, \quad k = 1, \ldots, q$$

2.- Teste das hipóteses $H_0 : B = 0$ vs $H_1 : B \neq 0$.

O teste proposto é baseado em S e para $p > 1$ a distribuição de S, ainda que $B = 0$, depende da distribuição F. Para resolver esta dificuldade é adotado o princípio básico de permutações de postos (Sen and Puri(1971)), que diz o seguinte:

Se $R = ((R_{ij}))$ $i = 1, \ldots, N, j = 1, \ldots, p$ é a matriz de postos e se arranjarmos as colunas de R tal que a primeira linha seja $(1, 2, \ldots, N)$, que denotaremos, esta nova matriz de postos por R^* assim, temos que

$$P(R = r\ /\ \sum(R^*),\ H_0) = 1/N! \qquad \forall\, r \in \sum(R^*) \qquad (5)$$

Teorema 2.1

Se u é a medida de probabilidade gerada por (5), temos que

$$E(S_{jk}/u) = \frac{1}{N} \sum_{i=1}^{N} d_{ik} \sum_{i=1}^{N} a_j(R_{ij}) = 0 \qquad (6)$$

desde que $\sum_{i=1}^{N} d_{ik} = 0$

e

$$cov(S_{jk}, S_{j'k'}/u) = \frac{1}{N-1} \sum_{i=1}^{N} d_{ik}\, d_{ik'} \sum_{i=1}^{N} (a_j(R_{ij}) - \overline{a_j})(a_{j'}(R_{ij'}) - \overline{a_{j'}}) \quad (7)$$

Prova: ver: Negrillo (1989) teorema 1.3.1.

Se a matriz S é reescrita em forma de vetor, isto é:

$$S = (S_{11}, \ldots, S_{p1}; \ldots ; S_{1q}, \ldots, S_{pq})' \quad \text{temos de (6) e (7)}$$

que

$$E(S/u) = 0$$

e

$$E(SS'/u) = V \otimes D = M = ((m_{jk,j'k'})) \quad j, j' = j, \ldots, p, \quad k, k' = 1, \ldots, q$$

onde

$\otimes$ é o produto de Kronecker de duas matrizes,

M é uma matriz $pq\ x\ pq$, com posto igual a [posto V] $q = rq$,

Se $r < p$, podemos usar a inversa generalizada ou trabalhar com um subconjunto de r variáveis,

$$V = ((v_{jj'})) \quad j, j' = 1, \ldots p \quad \text{sendo}$$

$$v_{jj'} = (N-1)^{-1} \left\{ \sum_{i=1}^{N} a_j(R_{ij}) a_{j'}(R_{ij'}) - N \overline{a}_j \overline{a}_{j'} \right\}$$

e

$$\overline{a}_j = N^{-1} \sum_{i=1}^{N} a_j(R_{ij})$$

De Sen and Puri (1971), Ruschendorf (1976), Ruymgaart and Zuijlen (1978), Lea and Puri (1986), e Harel (1988), temos que se $B = 0$ e algumas suposições (depende do autor)

$$S \xrightarrow{D} N_{pq}(0, \mathcal{D} \otimes v) \tag{8}$$

onde $\mathcal{D} = \lim_{N \to \infty} D$ e $V \xrightarrow{P} v$ quando $n \to \infty$.

Se $M^{-1} = V^{-1} \otimes D^{-1} = ((m^{jk,j'k'}))$, $jj' = 1, \ldots, p$, $k, k' = 1, \ldots, q$ a estatística para testar

$$H_0 : B = 0 \quad vs \quad H_1 : B \neq 0$$

é dada por

$$Q = \sum_{j=1}^{p} \sum_{j'=1}^{p} \sum_{k=1}^{q} \sum_{k'=1}^{q} m^{jk,j'k'} S_{jk} S_{j'k'} \tag{9}$$

Se M é de posto completo, sob H_0, a distribuição de Q é assintoticamente X^2_{pq}.

3.- Caracterização do problema de locação para c amostras multivariadas

Se $X_{k1}, \ldots, X_{kn_k}$ são n_k p-vetores aleatórios $i.i.d$ com função de distribuição $F_k(x)$ continua, para $k = 1, 2, \ldots, c$ $(c \geq 2)$ podemos escrever

$$F_k(x) = F(x - \theta_k) \quad k = 1 \ldots c$$

Se $N = \sum_{k=1}^{c} n_k$, fazendo $\theta_k = \alpha + B_k$ com $\sum_{k=1}^{c} (\frac{n_k}{N}) B_k = 0$, assim, somente $c - 1$ dos B_k são linearmente independentes e a hipótese nula $F_1 = \ldots = F_c$ implica $B_2 = \ldots = B_c = 0$ portanto por reparametrização podemos escrever

$$F_k(x) = F(x - \alpha - C_{k2}B_2 - \ldots - C_{kc}B_c), \quad k = 1 \ldots c$$

onde as constantes C_{kr} satisfazem a condição

$$\sum_{k=1}^{c} n_k C_{kr} = 0 \quad r = 2 \ldots c$$

obtendo assim

$$X_k = \alpha + C_{k2}B_2 + \ldots + C_{kc}B_c + e_k$$

O modelo linear é relacionado ao modelo original de classificação simples como se segue

$$\begin{array}{rclcl} \theta_1 & = & \alpha & & \\ \theta_2 & = & \alpha + \ B_2 & & B_2 = \theta_2 - \theta_1 \\ \vdots & & & & \vdots \\ \theta_c & = & \alpha + \ B_c & & B_c = \theta_c - \theta_1 \end{array}$$

em termos de modelo linear podemos testar

$H_0 : B_2 = \ldots = B_c = 0 \quad vs \quad H_1$: existe pelo menos um $B_k \neq 0 \ \ k = 2 \ldots c$ ou $H_0 : B = 0 \ \ vs \ \ H_1 : B \neq 0$

4)- Regressão simples multivariada

Consideramos o modelo de regressão simples ($q = 1$)

$$X_i = \alpha + BC_i + e_i \quad i = 1 \ldots N$$

onde

$\alpha = (\alpha_1, \ldots, \alpha_p)'$ e $B = (B_1, \ldots, B_p)'$ são parametros desconhecidos
C_i são constantes conhecidas
e_i são $v.a.i.i.d$ com função de distribuição F continua.

Para obter R–estimadores de α e B, como no caso univariado, vamos determinar as estatísticas lineares de postos, que apresentamos a seguir:

Se $R_{ij}(a_j, b_j)$ é o posto de $X_{ij} - a_j - b_jC_i$ em $X_{1j} - a_j - b_jC_1, \ldots, X_{Nj} - a_j - b_jC_N$ e

$R^{+}_{ij}(a_j, b_j)$ o posto de $|X_{ij} - a_j - b_jC_i|$ em $|X_{1j} - a_j - b_jC_1|, \ldots, |X_{Nj} - a_j - b_jC_N|$ para $i = 1, \ldots N \quad j = 1, \ldots p$ então as estatísticas lineares de postos são definidas para cada variável como

$$S_j(a_j, b_j) \ = \ N^{-1}\sum_{i=1}^{N}(C_i - \overline{c})a_j(R_{ij}(a_j, b_j))$$

$$T_j(a_j, b_j) = N^{-1} \sum_{i=1}^{N} \text{sinal } (x_{ij} - a_j - b_j C_i) a_j^*(R_{ij}^*(a_j, b_j))$$

onde $\bar{c} = N^{-1} \sum_{i=1}^{N} C_i \qquad a_j(i) = \phi_j(i/N+1)$

e $a_j^*(i) = \phi_j(\frac{1}{2}(1 + \frac{i}{N+1}))$

A função escore $a_j(.)$ e a estatística $S_j(.,.)$ são utilizados para obter R–estimadores para B e $a_j^*(.)$ e $T_j(.,.)$ são utilizados para obter estimadores de α.

O R–stimador de B_j é o valor $\widehat{B}_j$ tal que

$$S_j(a, \widehat{B}_j) = 0 \qquad \text{independente de } a$$

O R–estimador de α_j é o valor $\widehat{\alpha}_j$ tal que

$$T_j(\widehat{\alpha}_j, \widehat{B}_j) = 0 \quad \text{para } \widehat{B}_j \text{ fixo.}$$

Para determinar $\widehat{B}_j$, para $j = 1, \ldots p$ podemos usar um processo interarivo. Se para um dado b_r para $r = 1, 2, \ldots$ e $a = 0$

$$S_j(0, b_r) \neq 0 \quad \text{escolhemos} \quad \begin{array}{l} b_{r+1} > br \quad \text{se} \quad S_j(0, b_r) > 0 \quad \text{ou} \\ b_{r+1} < br \quad \text{se} \quad S_j(0, b_r) < 0 \end{array}$$

Se $S_j(0, b) = 0$ para $b' < b < b''$ onde

$$\begin{aligned} b' &= \inf\{b : S_j(0, b) < 0\} \\ b'' &= \sup\{b : S_j(0, b) > 0\} \end{aligned}$$

então $\widehat{B}_j = \frac{1}{2}(b' + b'')$

Similarmente, para determinar $\widehat{\alpha}_j$, para $j = 1, \ldots, p$, podemos usar um processo iterativo se para $\widehat{B}_j$ fixo e para algum $a_r; r = 1, 2, \ldots$, temos que

$$T_j(a_r, \widehat{B}_j) \neq 0 \quad \text{escolhemos} \quad \begin{array}{l} a_{r+1} > a_r \quad \text{se} \quad T_j(a_r, \widehat{B}_j) > 0 \\ a_{r+1} < a_r \quad \text{se} \quad T_j(a_r, \widehat{B}_j) < 0 \end{array}$$

Se $T_j(a, \widehat{B}_j) = 0$ para $a' < a < a''$ onde

$$\begin{aligned} a' &= \inf\{a : T_j(a, \widehat{B}_j) < 0\} \\ a'' &= \sup\{a : T_j(a, \widehat{B}_j) > 0\} \end{aligned}$$

então $\widehat{\alpha}_j = \frac{1}{2}(a' + a'')$

Teste para as hipóteses $H_0 : B = 0 \ \ vs \ \ H_1 : B \neq 0$.

Desde que o posto de $X_{ij} - a - bc_i$, para a e b fixo, é o mesmo que de X_{ij} já que os C_i são constantes conhecidas, assim podemos definir R_{ij} como o posto de X_{ij} em $X_{1j}, \ldots, X_{Nj}$ para $j = 1, \ldots, p$ independente de a e b.

De (4) a estatistica linear de postos para $q = 1$ é dada por

$$S_j = \sum_{i=1}^{N} d_i \, a_j(R_{ij}) \qquad j = 1 \ldots p$$

onde

$$d_i = N^{-1/2}(C_i - \overline{C})$$

Seja $S = (S_1, \ldots, S_p)$, então de (8) temos que a estatística do teste é dada por

$$\mathcal{L} = (N \, S' \, V^{-} \, S)/D$$

onde $D = \sum_{i=1}^{N} d_i^2$

Se V é de posto completo, sob H_0, $\mathcal{L}$ tem distribuição assintótica χ^2_p.

REFERÊNCIAS

1 Harel M. (1988), Weak convergence of multidimensional rank statistics under φ-maximing conditions, Journal of Statistical Planning and inference, Vol. 20, pg 41-63

2 Hettmansperger, T. P. and McKean, J. W. (1977). A Robust Alternative Based on Ranks to Least Squares in Analyzing Linear Model, Technometrics, Vol. 19, nº3, pg 275-284

3 Hettmansperger, T. P. and McKean, J. W. (1983). A Geometric Interpretation of Inference Based on Rank in the Linear Model, Journal of the American Statistical Association, Vol. 78, nº384, pg 885-893

4 Hettmansperger, T. 8 (1984) Statistical Inference Based on Ranks, John Wiley e Sons

5 Negrillo, B. G. (1989) Um modelo Linear Geral Não Paramétrico, Relatório Técnico nº 34/89

6 Pirie, W. R. and Rauch, H. L. (1984) Simulated Efficiencies of Tests and Estimators from General Linear Models Analysis based on Ranks: The two-way Layout with Interation, Statist. Comput. Simul., vol. 20, pg 197-204

7 Puri, L. M. and Sen, K. P. (1971). Nonparametric Methods in Multivariate Analysis, John Wiley e Sons

8 Ruschendorf L. (1976), Asymtotic Distribuitions of Multivariate Rank Order Statistics, The Annals of Statistics, Vol. 4, nº5, pg 912-913

9 Ruymgaart, F. H. and Van Zuijlen C. A. (1978), Asymtotic normality of MultivariateLinear Rank Statistics in the Non-I.I.D. Case, The Annals of Statistics, Vol.6, nº3, pg 588-602

RELATÓRIOS TÉCNICOS — 1990

01/90 **Harmonic Maps Into Periodic Flag Manifolds and Into Loop Groups** — *Caio J. C. Negreiros.*

02/90 **On Jacobi Expansions** — *E. Capelas de Oliveira.*

03/90 **On a Superlinear Sturm–Liouville Equation and a Related Bouncing Problem** — *D. G. Figueiredo* and *B. Ruf.*

04/90 **F– Quotients and Envelope of F–Holomorphy** — *Luiza A. Moraes, Otília W. Paques* and *M. Carmelina F. Zaine.*

05/90 **S-Rationally Convex Domains and The Approximation of Silva-Holomorphic Functions by S–Rational Functions** — *Otília W. Paques* and *M. Carmelina F. Zaine.*

06/90 **Linearization of Holomorphic Mappings On Locally Convex Spaces** — *Jorge Mujica* and *Leopoldo Nachbin.*

07/90 **On Kummer Expansions** — *E. Capelas de Oliveira.*

08/90 **On the Convergence of SOR and JOR Type Methods for Convex Linear Complementarity Problems** — *Alvaro R. De Pierro* and *Alfredo N. Iusem.*

09/90 **A Curvilinear Search Using Tridiagonal Secant Updates for Unconstrained Optimization** — *J. E. Dennis Jr., N. Echebest, M. T. Guardarucci, J. M. Martínez, H. D. Scolnik* and *C. Vacchino.*

10/90 **The Hypebolic Model of the Mean × Standard Deviation "Plane"** — *Sueli I. R. Costa* and *Sandra A. Santos.*

11/90 **A Condition for Positivity of Curvature** — *A. Derdzinski* and *A. Rigas.*

12/90 **On Generating Functions** — *E. Capelas de Oliveira.*

13/90 **An Introduction to the Conceptual Difficulties in the Foundations of Quantum Mechanics a Personal View** — *V. Buonomano.*

14/90 **Quasi-Invariance of product measures Under Lie Group Perturbations: Fisher Information And L^2-Differentiability** — *Mauro S. de F. Marques* and *Luiz San Martin.*

15/90 **On Cyclic Quartic Extensions with Normal Basis** — *Miguel Ferrero, Antonio Paques* and *Andrzej Solecki.*

16/90 **Semilinear Elliptic Equations with the Primitive of the Nonlinearity Away from the Spectrum** — *Djairo G. de Figueiredo* and *Olimpio H. Miyagaki.*

17/90 **On a Conjugate Orbit of G_2** — *Lucas M. Chaves* and *A. Rigas.*

18/90 **Convergence Properties of Iterative Methods for Symmetric Positive Semidefinite Linear Complementarity Problems** — *Álvaro R. de Pierro* and *Alfredo N. Iusem.*

19/90 **The Status of the Principle of Relativity** — *W. A. Rodrigues Jr.* and *Q. A. Gomes de Souza.*

20/90 **Geração de Gerenciadores de Sistemas Reativos** — *Antonio G. Figueiredo Filho* e *Hans K. E. Liesenberg.*