V. 8, n.º 3 (setembre 1984) pp. 145-151

APLICACIÓ DEL CRITERI C_P DE MALLOWS A L'OBTENCIÓ DE LA MILLOR EQUACIÓ DE REGRESSIÓ D'UN PROBLEMA ENERGÈTIC

MOISÈS VALLS I COLOM

UNIVERSITAT POLITÉCNICA DE CATALUNYA

Un dels principals problemes que planteja la regressió és el de trobar la millor equació lineal a partir d'una certa llista de possibles variables regressores. L'article té com a objectiu comentar el mètode de generació de totes les regressions possibles amb posterior comparació mitjançant el criteri C_p de Mallows, il.lustrant el procediment amb l'aplicació del mètode a un cas real.

Keywords: MALLOWS C_p, ALL POSSIBLE REGRESSIONS, BEST REGRESSION EQUATION

1. INTRODUCCIÓ:

Generalment, en estudiar la regressió lineal es considera que el model que relaciona la resposta Y amb les variables independents Xj és ja conegut. A la pràctica en canvi, és molt habitual de disposar d'una llista de variables amb possible relació amb Y, sense saber però, quines són realment influents, en quina métrica i quines interaccions contenen. L'objectiu d'aquest article és precisament discutir un procediment per a seleccionar la millor equació de regressió a partir d'una llista de possibles regressors. Les variables incloses a l'esmentada llista pot ésser que siguin ja funcions d'unes altres variables originals i suposarem que constituiexen el conjunt complet de variables amb influència sobre la resposta Y.

Tradicionalment han existit dos criteris oposats per a seleccionar la millor equació de regressió:

a) Per una banda és desitjable que el model contingui tants regressors com sigui possible ja que així es reté la màxima informació i es poden fer previsions amb un biaix relativament petit.

b) Per altra banda, com que a partir d'un cert moment la variància de les prediccions augmenta en incrementar-se el nombre de regressors, és preferible que el model contingui poques variables amb la qual cosa també s'aconsegueix abaratir el cost d'obtenció de la informació.

El compromís entre aquests dos punts de vista constituiex la selecció de la millor equació de regressió i ha originat el desenvolupament d'un considerable nombre de criteris de comparació de les diferentes equacions candidates.

L'aparició dels ordinadors de gran potència i velocitat de càlcul ha fet possible el mètode que descriurem a continuació.

2. GENERACIÓ DE TOTES LES REGRESSIONS POSSIBLES.

Aquest procediment consisteix en ajustar totes les possibles equacions de regressió que contenen $X_o = 1$ i algún subconjunt de les variables $X_1, X_2, \ldots, X_r$. Atés que cada variable X_i pot entrar o no a l'equació, existiràn 2^r regressions possibles. Per tant, en el cas de r = 20 hi haurien $2^{20} = 1048576$ equacions a examinar.

- Moisés Valls i Colom - Escola Tècnica Superior d'Enginyers Industrials de Barcelona, Dep. Tècniques Cuantitatives de la Gestió - Universitat Politècnica de Catalunya - Av. Diagonal, 647 - Barcelona 08028.

Això condiciona la utilització del mètode a que el nombre total de variables no sigui molt elevat i obliga a posseir un algorisme eficient per a determinar la millor equació de regressió. A més a més, una vegada obtingudes les diferentes equacions necessitarem un criteri senzill que permeti de comparar-les entre elles.

El criteri que farem servir serà la Cp de Mallows /3/ i l'algorisme més utilitzat és el de Furnival i Wilson /1/, que només computa una fracció del total de regressions.

3 L'ESTADISTIC C_p DE MALLOWS.

Si es considera el model lineal de p paràmetre tres

$$Y_{pi}=B_o+B_1X_{1,i} + B_2X_{2,i} +\ldots+ B_{p-1}X_{p-1,i}+ \varepsilon_i \qquad i = 1,2,\ldots,N \quad (1)$$

que es pot escriure en forma matricial com

$$\underline{Y}_p = \underline{\underline{X}}_p \underline{B}_p + \underline{\varepsilon} \qquad (2)$$

$$\text{on } \underline{Y}_p = \begin{bmatrix} Y_{p1} \\ Y_{p2} \\ \cdot \\ \cdot \\ \cdot \\ Y_{pN} \end{bmatrix} \qquad \underline{\underline{X}}_p = \begin{bmatrix} 1 & X_{11} & \cdots & X_{p-1,1} \\ 1 & X_{12} & \cdots & X_{p-1,2} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ 1 & X_{1N} & \cdots & X_{p-1,N} \end{bmatrix}$$

$$\underline{B}_p = \begin{bmatrix} B_o \\ B_1 \\ \cdot \\ \cdot \\ \cdot \\ B_{p-1} \end{bmatrix} \qquad \underline{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \cdot \\ \cdot \\ \cdot \\ \varepsilon_N \end{bmatrix}$$

la millor previsió de $\underline{Y}_p$ segons el criteri dels mínims quadrats serà

$$\hat{\underline{Y}}_p = \underline{\underline{X}}_p \hat{\underline{B}}_p \qquad (3)$$

on $\hat{\underline{B}}_p = (\underline{\underline{X}}'_p \underline{\underline{X}}_p)^{-1} \underline{\underline{X}}'_p \underline{Y}$ és precisament l'estimador mínimquadràtic de $\underline{B}_p$.

Per a seleccionar la millor equació de regressió, Mallows /3/ proposa minimitzar

$$\Delta_p = \frac{1}{\sigma^2} E\left[\sum_{c=1}^{N} (\hat{Y}_{pi} - \theta i)^2\right] \qquad (4)$$

on $\hat{Y}_{pi}$ representa la previsió de Y_i amb l'equació (3) i $\theta_i=E(Y_i)$ la previsió de Y_i amb el veritable model.

Com que $E(\hat{Y}_{pi}-\theta_i)^2$ és precisament l'error quadràtic mitjà de $\hat{Y}_{pi}$, Δ_p mesurarà la discrepància total entre l'equació (3) i el veritable model.

Desenvolupant (4) resulta

$$\Delta_p = \frac{1}{\sigma^2}\left[\sum_{i=1}^{N} VAR\ (\hat{Y}_i) + \sum_{i=1}^{N} (\zeta_{pi}-\theta_i)^2\right] \qquad (5)$$

on $\zeta_{pi} = E(\hat{Y}_{pi})$ representa la previsió de Y_i amb el model (2).

Si es defineix $SSB_p = \sum_{i=1}^{N} (\zeta_{pi}-\theta_i)^2$, com que $\sum_{i=1}^{N} VAR(\hat{Y}_i) = p\sigma^2$ l'equació (5) quedarà reduida a

$$\Delta_p = p + \frac{SSB_p}{\sigma^2} \qquad (6)$$

que expressa la descomposició de la discrepància total Δ_p en un terme de variància p més un terme de biaix SSB_p/σ^2.

Per a estimar Δ_p, Mallows suggereix d'utilitzar

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - N \qquad (7)$$

on $\hat{\sigma}^2$ és la variància residual de l'equació que conté les r variables i RSS_p la suma de quadrats dels residus de l'equació (3).

Com que $E(RSS_p) = \sigma^2(N-p) + SSB_p$

es complirà $E(C_p)\simeq\Delta_p$ y per tant C_p serà aproximadament un estimador no esbiaixat de Δ_p.

Es podria construir un gràfic com el de la Figura 1 en el qual per a cada equació es representés el valor de C_p en ordenades i el nombre de paràmetres de l'equació p en abscises. Si el terme de biaix SSB_p fos despreciable $\Delta_p\simeq p$ i C_p prendria valors pròxims a p.

Per tant, els punts que estiguessin al voltant de la línia C_p=p representarien models amb poc biaix mentre que els que es trobessin per damunt de la línia tindrien una estimació del terme de biaix C_p-p. L'ordenada C_p mesuraria la discrepància entre les previsions de (3) i les del veritable model.

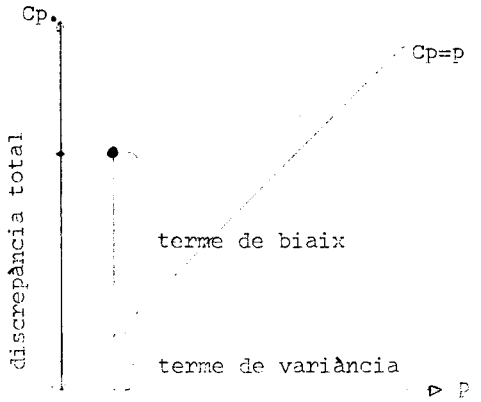

Fig. 1 Representació gràfica d'una equació de regresió.

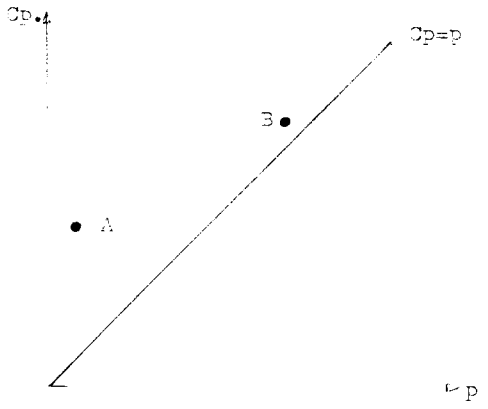

Fig. 2 Representació de dues equacions amb problemes de comparació.

La millor equació de regressió s'escolliria després d'inspeccionar el gràfic de totes les equacions representades. De totes formes l'elecció no sempre és evident. Per exemple, al gràfic de la Figura 2 s'elegirà el model A o B depenent de si es prefereix:

A) Una equació esbiaixada que no representa les actuals dades tan bé com B però amb una estimació de la discrepància total al veritable model menor.

B) Una equació amb més paràmetres que s'ajusta millor a les actuals dades però amb una discrepància total més gran

Per a fer previsions sembla millor eliminar alguns regressors i acceptar un cert biaix a canvi d'una menor discrepància total i un model més simple, la qual cosa equival a escollir el model que minimitzi C_p.

A més com que quan l'equació conté totes les variables, sempre es compleix

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + 2p - N = \frac{RSS_p}{RSS_p/(N-p)} + 2p - N = P$$

el model que s'obtindria minimitzant el terme de biaix seria sovint el que contingués les r variables. En general doncs, és preferible el model de C_p mínima al de mínim biaix.

Per altra banda, les r variables hauran d'haver estat acuradament elegides per poder estar raonablement segurs que el terme de biaix del model que les conté totes és despreciable tal com suposarem en estimar σ^2, ja que el conjunt de regressors escollits dependrà evidentment del conjunt total de possibles regressors.

4. EXEMPLE D'APLICACIO.

Una central generadora d'electricitat consumeix dos tipus de combustible (fuel oil i gas natural) produint una certa quantitat d'energia al llarg del dia (Energia generada) que en part s'utilitza per a les necessitats internes de la Central, quedant la resta per a satisfer la demanda en un moment determinat (Energia facturada).

Per a determinar la qualitat en el funcionament de la Central s'utilitza habitualment una mesura inversa de la del rendiment anomenada consum específic (ρ) que es defineix com la relació entre l'energia consumida i la facturada per la Central (Totes les dades estan mesurades en períodes de 24 hores):

$$\rho = \frac{\text{Energia Consumida}}{\text{Energia Facturada}} = \frac{9'76\times FUEL + 1000\times GAS}{\text{Energia Facturada}}$$

$$\left(\frac{\text{Tèrmies}}{\text{Mwh}}\right)$$

on FUEL i GAS són respectivament els Kg. de fuel oil i els milers de Tèrmies de gas consumits per la Central.

Per a mesurar la variabilitat de la potència al llarg del dia es defineix M com la diferència entre les potencies màxima i mínima del dia

M = Potència Màxima - Potencia Mínima (Mw)

Per últim i de cara a tenir en compte el tipus de combustible utilitzat, es defineix R com la relació entre l'energia consumida amb fuel i l'energia total consumida.

$$R = \frac{9'76\times FUEL}{9'76\times FUEL+1000\times GAS}$$

Si només s'utilitza fuel ⟶ R = 1
Si només s'utilitza gas ⟶ R = 0
Si es fan servir els dos combustibles ⟶ 0<R<1

Per a evitar les anomalies produides per les engegades i aturades de la Central, només s'utilitzaran els dies en què hagi funcionat les 24 hores, amb la qual cosa es redueix el nombre d'observacions de 1468 a 743.

La potencia mitjana facturada serà doncs

$$P = \frac{\text{Energia Facturada}}{24} \quad (Mw)$$

El principal objectiu de l'estudi és poder fer previsions del funcionament del sistema a partir de variables conegudes. Com a variable depenent prendrem el consum específic (ρ) i les independents les seleccionarem d'entre les variables P, P^2, P^3, R, R^2, M i M^2 que hem obtingut a partir d'una sèrie d'anàlisis prèvies com és ara, l'analisi exploratoria, els diagrames bivariants, l'anàlisi de la matriu de correlacions, la regressió pas a pas, etc. Com que hi ha 7 possibles regressors, per a seleccionar la millor equació de regressió s'hauran d'analitzar $2^7 = 128$ regressions, de les quals destaquen les incloses a la taula T1, que es troben representades a la Figura 3, i a les quals se les hi ha calculat 4 estadistics:

- C_p de Mallows
- R^2 Coeficient de determinació: Indica el tant per u de variabilitat explicada per la regressió.
- R^2 Coeficient de determinació ajustat. Es defineix com

$$R^2adj = 1 - (1 - R^2)\frac{N}{N-p}$$

Té l'avantatge respecte de R^2 de tenir en compte el nombre de regressors introduits al model

- S^2 variància residual: Mesura la variabilitat dels resodus.

El criteri clarament competidor al proposat de minimitzar C_p és el de maximitzar R^2adj. De fet Kennard /2/ va demostrar que hi ha una relació lineal entre ambdós estadistics. La principal avantatge de C_p estreba en que R^2adj. proporciona únicament un escalar per a cada equació mentre que el gràfic C_p versus p és molt més profitós ja que permet escollir el model més adequat depenent de l'objectiu de l'estudi.

Tots els criteris coincideixen en que la millor equació de regressió es:

$$\hat{\rho}=3363'03-11'0178P+0'0375292P^2-4'3555\times10^{-5}P^3 \quad (8)$$
$$(61'56)\quad (1'07)\quad (0'0059)\quad (1\times10^{-5})$$
$$+\ 98'6364R-131'492R^2+0'159811M$$
$$(33'61)\quad (33'84)\quad (0'036)$$

Entre parèntesis figuren les desviacions tipus dels coeficients.

Els valors que minimitzen el consum específic, tenint en compte que per raons tècniques la potència ha d'estar compresa entre 100 Mw. i 300 Mw., són:

$P^* = 300$ Mw.
$M^* = 0$ Mw. ja que evidentment la potència màxima sempre ha de ser superior a la mínima.
$R^* = 1$ que equival a utilitzar Fuel-oil com a combustible.

L'equació

$$\hat{\rho} = 3365'93-11'0068P+0'0376360P^2- \quad (9)$$
$$-4'4025\times10^{-5}P^3+100'216R-133'956R^2+$$
$$+7'0451\ 10^{-4}M^2$$

és quasi tan vàlida com l'anterior ja que tots

TAULA T 1

MALLOWS EQUACIONS DE REGRESSIO

P	Variables regressores	C_p	R^2	R^2adj	S^2
1	-	2280'61	0	0	14458'64
2	P	416'51	0'617586	0'617070	5536'65
3	P, P^2	102'06	0'722312	0'721562	4025'84
4	P, P^2, M	53'86	0'738926	0'737867	3790'10
5	P, P^2, M, R^2	29'19	0'747753	0'746386	3666'92
6	P, P^2, P^3, M, R^2	12'99	0'753777	0'752106	3584'20
6	P, P^2, P^3, M^2, R^2	13'68	0'753547	0'751875	3587'55
6	P, P^2, P^3, M, R	19'47	0'751633	0'749948	3615'41
6	P, P^2, P^3, M^2, R	20'51	0'751289	0'749602	3620'42
6	P, P^2, M, R, R^2	22'74	0'750551	0'748859	3631'16
6	P, P^2, P^3, R, R^2	23'51	0'750295	0'748601	3634'89
6	P, P^2, M^2, R, R^2	23'59	0'750268	0'748574	3635'28
7	P, P^2, P^3, M, R, R^2	6'38	0'756625	0'754641	3547'56
7	P, P^2, P^3, M^2, R, R^2	6'78	0'756495	0'754510	3549'45
7	P, P^2, P^3, M, M^2, R^2	14'62	0'753901	0'751895	3587'27
7	P, P^2, P^3, M, M^2, R	21'14	0'751742	0'749719	3618'74
7	P, P^2, M, M^2, R, R^2	24'29	0'750698	0'748666	3633'95
8	P, P^2, P^3, M, M^2, R, R^2	8	0'756752	0'754435	3550'53

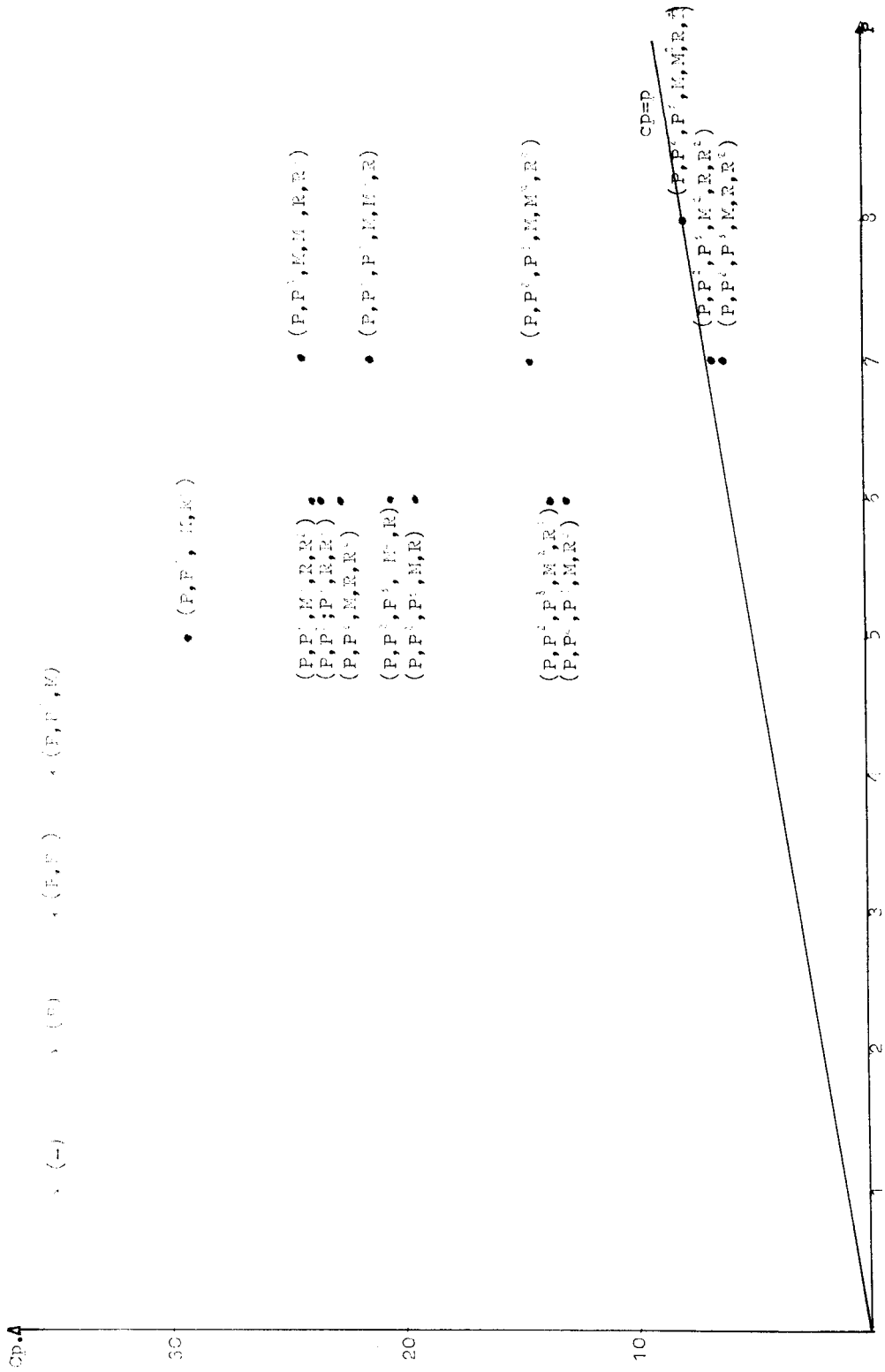

Figura 3: Representació gràfica de les equacions incloses a la taula T1

els estadístics són pràcticament iguals. Els valors dels regresors que minimitzen el consum específic resulten ser també els mateixos d'abans.

Per últim, cal esmentar que tots aquests resultats han estat obtinguts mitjançant el programa P9R del paquet de programes estadístics BMDP que utilitza l'algorisme de Furnival: Wilson /1/ per a trobar la millor rquació de regressió.

5. REFERÈNCIES.

/1/ FURNIVAL G.M. AND R.W. WILSON: "Regression by Lears and Bounds" - Technometrics, 16, 499-511 (1974).

/2/ KENNARD, R.W.: "A note on the C_p statistic" - Technometrics, 13, 899 (1971).

/3/ MALLOWS, C.L.: "Some comments on C_p" - Technometrics, 15, 661-675 (1973).