Seventh LACCEI Latin American and Caribbean Conference for Engineering and Technology (LACCEI'2009)
"Energy and Technology for the Americas: Education, Innovation, Technology and Practice"
June 2-5, 2009, San Cristóbal, Venezuela.

Selección de Variables en Regresión Componentes Principales

Ober Navarro
Universidad Simón Bolívar, Caracas, Venezuela, obernavarro@usb.ve

RESUMEN

La Regresión Lineal Múltiple (RLM) es un método de análisis de datos muy popular entre ingenieros y científicos. Esta investigación estudia el problema de la selección de variables en RLM pero, con aplicaciones en datos que presenten una marcada multicolinealidad. Se analizan las bases teóricas en que se basa el método de selección de variables propuesto por Mansfield, Webster y Gunst (MWG, 1977) el cual es de utilidad cuando la multicolinealidad es problema. Se implementa el método en un guión (script) del paquete estadístico R y se compara con los denominados procedimientos clásicos de selección de variables.

Palabras claves: Regresión Lineal Múltiple, Selección de Variables, Componentes Principales, Algoritmo computacional, PRESS.

ABSTRACT

The Lineal Multiple Regression (RLM) is a method of analysis of data very popular among engineers and scientifics. This investigation studies the problem of the selection of variables in RLM but, with applications in data that present a marked multicollinearity. The theoretical bases of the method of selection of variables proposed by Mansfield, Webster and Gunst (MWG, 1977) are analyzed in that which of utility when the multicollinearity is problem. The method is implemented in a script of the statistical package R and it acting is compared with the denominated classic procedures of selection of variables.

Keywords: Lineal Multiple Regression, Selection of Variables, Principals Components, Computational Algorithm, PRESS.

1. INTRODUCCIÓN

Cuando se tiene el modelo de Regresión lineal

$$Y = X\beta + \varepsilon \qquad (1)$$

donde $\varepsilon \sim N(0, I_n\sigma^2)$, X es de dimensiones *nx(k+1)* y β es *(k+1)x1*. Si se utiliza el procedimientos minimos cuadrados ordinarios (MCO) para estimar β. Es bien conocido que la ecuación de estimación viene dada por

$$\hat{\beta} = (X^{\perp}X)^{-1}X^{\perp}Y$$

En muchas situaciones se supone que la matriz de diseño $X^{\perp}X$ involucrada es no singular y por consiguiente el problema se resuelve satisfactoriamente, obteniendose estimadores lineales insesgados y de varianza mínima. Sin embargo, este panorama cambia radicalmente si esta matriz de diseño esta altamente condicionada o muy cercana a la singularidad, en este caso, los estimadores MCO no son confiables debido a que exhiben una elevada correlación y una alta inestabilidad en la varianza. Si este es el caso se dice que, existe *multicolinealidad*. Cuando la multicolinealidad es problema es menester utilizar procedimientos distintos a MCO. Se utilizan las técnicas de *regresión sesgadas*, tales como, la *Regresión Componentes Principales*, mediante el cual se obtienen estimadores sesgados. Pero que, al compararlos con los MCO resultan ser, aunque sesgados, más precisos.

Por otra parte, un problema que también es inherente al investigador que utiliza RLM es el denominado problema de la *selección de variables:* consiste en encontrar un subconjunto apropiado de variables explicativas para ser

usadas en la ecuación final de predicción. El objetivo es obtener un modelo parsimonioso, es decir, ajustar bien los datos pero usando la menor cantidad posible de variables explicativas o regresoras. Selección de variables y multicolinealidad son dos problemas que se pueden tratar de manera simulánea. Bajo este escenario se ubica el objetivo central de esta investigación: describir las bases matemáticas en que se basa el algoritmo para seleccionar variables propuesto por MWG, el cual es útil cuando la multicolinealidad existe en la data, codificarlo en un script del paquete estadístico R y finalmente comparar su desempeño con los *métodos clásicos de selección de variables*.

2. REGRESIÓN COMPONENTES PRINCIPALES

La Regresión Componentes Principales es una técnica de estimación para combatir la multicolinealidad. Si la matriz X es centrada y escalada se denota X^* el modelo (1) es reescrito como

$$Y = \beta_0 1 + X^* \beta^* + \varepsilon \qquad (2)$$

Un conjunto de variables denominadas *componentes principales* de la matriz de correlación $X^{*\perp} X^*$ se eliminan obteniendose un efecto sustancial en la reducción de la varianza. El método es superior en estimación y predicción a MCO. Los componentes principales son ortogonales mutuamente lo que permite restarle una cantidad de varianza a cada uno de ellos. Se consideran los autovectores asociados con los autovalores $(\lambda_1, \lambda_2, .., \lambda_k)$ de $X^{*\perp} X^*$. Sea $V = [v_1, v_2, .., v_k]$ una matriz ortogonal *kxk* donde la *j*-ésima columna el vector v_j de V está asociado al *j*-ésimo autovalor que verifica $VV^\perp = V^\perp V = I$ y el modelo (2) se escribe como

$$Y = \beta_0 1 + X^* V V^\perp \beta^* + \varepsilon \qquad \text{ó} \qquad Y = \beta_0 1 + W\gamma + \varepsilon \qquad (3)$$

Donde $W = X^* V$ es una matriz *nxk* y $\gamma = V^\perp \beta^*$ es un vector *kx1* de coeficientes $\gamma_1, \gamma_2, .., \gamma_k$. W representa a los *k* componentes principales y se verifica fácilmente que son ortogonales uno a uno:

$$W^\perp W = (XV)^\perp (XV) = V^\perp X^\perp XV = \Lambda$$

Donde $\Lambda = diag(\lambda_1, \lambda_2, .., \lambda_k)$ .Si la regresión es ejecutada via modelo (3) la varianza de los coficientes son los inversos de los autovalores, dada por

$$Var(\hat{\gamma}) = \frac{\sigma^2}{\lambda_j} \text{ para } j = 1, 2, .., k \qquad \text{y} \qquad \hat{\gamma} = \Lambda^{-1} W^\perp Y$$

Se observa que valores de λ_j cercanos a cero causan efectos notables, dado que, inflan el valor de la varianza. Por consiguiente, el camino más plausible de tomar es, prescindir de ellos. Si se supone que se eliminan *t* de ellos (*t* < *k*), entonces nos quedan *s* = *k*-*t* . Ahora el modelo (3) es escrito de la siguiente manera :

$$Y = \beta_0 1 + W_s \gamma_s + \varepsilon \qquad (4)$$

Los coeficientes de la regresión componentes principales para (4) en terminos de las variables centradas y escaladas viene dada por:

$$b^*_{cp} = V_s \hat{\gamma}_s \qquad (5) \qquad \text{y} \qquad \hat{\gamma}_s = \Lambda_s^{-1} W_s^\perp Y$$

3. MULTICOLINEALIDAD

La presencia de dependencia lineal entre las variables regresoras es llamada multicolinealidad. Matemáticamente, la dependencia lineal se entiende como el procedimiento para encontrar constantes c_j no todas cero tales que .

$$\sum_{j=1}^{k} c_j x_j$$

Donde los x_j son las columnas de X. Este problema usualmente es causado por las restricciones artificiales a que es sometida la muestra, o por el hecho de tener pocas observaciones y muchas variables regresoras.

3.1.-DETECTANDO LA MULTICOLINEALIDAD.

Existen muchas formas de para detectar la multicolinealidad, los métodos más utilizados por los investigadores son:

- Gráfico de Dispersión matricial: permite tener una idea acerca de la posible relación lineal entre las variables.
- Matriz de correlación de las variables regresoras: la existencia de algún valor alto próximo a ± 1 indica que existe una fuerte relación lineal entre las regresoras.
- Analisis de los elementos de la diagonal de la matriz inversa de correlación R^{-1}, ya que verifica que el *i*-ésimo elemento de esta matriz es

$$FIV(i) = \frac{1}{1 - r_i^2}$$

Por lo tanto si *FIV(i)* es un valor muy alto si existe multicolinealidad causada por la variable x_i.

- Calcular los autovalores de la matriz de correlación y verificar el valor de κ o índice de la matriz.

4. ELIMINACIÓN DE COMPONENTES PRINCIPALES

Los componentes principales de la matriz W contienen exactamente la misma información que la matriz centrada y escalada X^* excepto que los datos en esta nueva variable están completamente incorrelacionados, propiedad que permite ordenarlos en un rango de acuerdo a la magnitud de sus autovalores (Draper y Smith, 1981). Esta situación a conducido a proponer métodos para determinar cuantos componentes deben ser removidos del modelo, para lograr una sustancial reducción en la varianza de los parámetros del modelo. Entre estos métodos los más usados son:

- Descartar componentes asociados a autovalores muy pequeños. Usualmente los componentes son eliminados hasta que el resto de ellos explican algún porcentaje preseleccionado del total de la varianza (generalmente 85%), es decir se seleccionan los *s* componentes más grandes que verifican lo siguiente

$$\frac{\sum_{j=1}^{s} \lambda_j}{k} \geq 0.85$$

- La regla de Kaiser-Gutman: se escogen componentes asociados con autovalores de valor mayores que 1.0

5. MÉTODOS CLÁSICOS DE SELECCIÓN DE VARIABLES

Seleccionar un conjunto de variables regresoras en un modelo de regresión lineal múltiple consiste en dos aspectos: a) escoger un criterio de selección para evaluar el modelo. b) desarrollo del procedimiento computacional para construir el modelo.

5.1.- CRITERIOS DE SELECCIÓN DE VARIABLES.

La razón principal para suprimir una variable regresora de un modelo, es que se puede mejorar la presición de los parámetros estimados en las variables retenidas, aún si alguna de las variables suprimidas es importante. Esto también es cierto para las varianzas de la variable respuesta. Suprimiendo variables se introduce sesgo en los estimadores de los coeficientes de las variables retenidas y en la variable respuesta. Sin embargo si las variables suprimidas tienen pequeños efectos, el Cuadrado Medio del Error (CME) de los estimadores sesgados será menor que la varianza de los estimadores insesgados, esto se debe a que la cantidad de sesgo introducido es menor que la reducción de la varianza, de aquí la importancia y utilidad de los criterios de selección de variables. Entre los criterios más utilizados, es menester hacer referencia al coeficiente de determinación múltiple.

5.1.1.- EL COEFICIENTE DE DETERMINACIÓN MÚLTIPLE

Es la medida de adecuación por excelencia, ampliamente usada para medir la eficiencia del modelo. También es utilizada como criterio de selección de la siguiente manera. Se define por

$$R_p^2 = \frac{SCR(p)}{SCT} = 1 - \frac{SCE(p)}{SCT}$$

Donde *SCR(p)* y *SCE(p)* denotan la suma de cuadrados de regresión y la suma de cuadrados de error respectivamente para subconjuntos de *p* terminos. Existen valores de R_p^2 para cada valor de *p*. R_p^2 crece si *p* crece y es máximo cuando $p = k+1$. El analista usa este criterio para agregar regresoras hasta el punto donde donde una variable adicional no sea útil. Otros criterios similares son : C_p de Mallows y *AIC* de Akaike.

5.2 .-TODAS LAS REGRESIONES POSIBLES (TRP).

Una aproximación natural al problema de selección de variables consiste en utilizar algunos de los criterios descritos anteriormente y ver su reducción debido a la introducción en el modelo de alguna variable x_i. Por ejemplo puede suceder que la variable x_i que debiera ser incluida en el modelo, no origine una reducción significativa del *SCE* cuando es introducida despues de x_j . Si esto ocurre, es claro que, x_i no mostrará sus buenas condiciones como regresora mas que si es intoducida con x_j ausente. Una posible solución a esta incertidumbre sería, dado *p* regresoras, formar todos los posibles subconjuntos de regresoras y efectuar todas las regresiones posibles, reteniendo aquellas que de acuerdo al criterio que se halla adoptado muestre mejor eficiencia. El inconveniente es el gran volumen de cálculo que es preciso efectuar. Piénsese, por ejemplo, con *p* regresoras puede estimarse $2^p - 1$ regresiones distintas. Si $p=5$ se tendría 31 regresiones; si p=10 se tendría 1023 regresiones. Afortunadamente hoy día existen procedimientos para reducir y agilizar este cálculo.

5.3.- MÉTODOS DE SELECCIÓN ESCALONADA.

Los métodos de selección escalonada son técnicas de selección que secuencialmente agregan o suprimen una variable regresora sencilla de la ecuación de estimación. Dado que involucra una serie de pasos antes de obtener la ecuación final, y dado que, cada paso conduce automáticamente al próximo, implica que se necesita evaluar un número de ecuaciones mucho menor que $2^p - 1$ las requeridas en todas las regresiones posibles. Los métodos de selección escalonada básicamente son tres: a) eliminación progresiva (forward), b) eliminación regresiva (backward), c) el método por pasos (stepwise). Este ultimo es la combinación de los dos primeros más una modificación,coloquialmente se le llama el procedimiento de eliminación por pasos.

6. EL MÉTODO DE SELECCIÓN DE MWG

MWG consideran eliminar *r* variables regresoras del modelo (4). Particionan la matriz V_s en filas de la siguiente manera

$$V_s = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$$

Donde V_1 es de dimensiones *(k-r)xs* y V_2 de dimensiones *rxs* y cuyas filas corresponden a las *r* variables a ser eliminadas. MWG consideran un estimador alterno para γ_s dado por

$$\hat{\gamma}_s = A_s \Lambda_s^{-1} W_s^{\perp} Y = A_s \hat{\gamma}_s$$

El estimador análogo a (5) viene dado por

$$\tilde{b} = V_s \hat{\gamma}_s = \begin{bmatrix} V_1 A_s \hat{\gamma} \\ V_2 A_s \hat{\gamma} \end{bmatrix} = \begin{bmatrix} \tilde{b}_{k-r} \\ \tilde{b}_r \end{bmatrix} \qquad (6)$$

Donde las últimas filas de (6) corresponden a las variables que van a ser eliminadas. El uso de la estimación componentes principales con las *r* predictoras eliminadas conduce a la ecuación de predicción $\hat{Y} = \overline{Y} + \sum_{j=1}^{k-r} u_j \tilde{\gamma}$ (7) con $u_j = z^{\perp} v_j$. La suma de cuadrados de residuales para (7) es

$$SSR = SSR^* + u_r$$

Donde SSR^* es la suma de cuadrados de residuales para el modelo (4) y u_r viene dado por

$$u_r = \hat{\gamma}^{\perp} V_2^{\perp} \left(V_2 \Lambda_s^{-1} V_2^{\perp} \right)^{-1} V_2 \hat{\gamma} \qquad (8)$$

7. EL ALGORITMO

Paso1: el valor u_1 se calcula para cada una de las *k* variables mediante (8). Se denota el menor de estos por u_{1m} . Si se determina que este valor es suficientemente pequeño mediante un test de significancia se elimina la regresora correspondiente a ese valor y se va al paso 2.

Paso 2: Se calcula el valor u_2 para cada uno de los pares formados por la variable eliminada y las *k-1* restantes. Se denota el menor de estos por u_{2m}. Si se determina que la diferencia $\left(u_{2m} - u_{1m}\right)$ es suficientemente pequeña mediante un test de significancia, las dos variables involucradas se eliminan. De otro modo; se elimina la del paso 1.

Paso 3 : Se calcula el valor u_3 para cada una de las ternas formados por las dos variables eliminada y las *k-2* restantes. Se denota el menor de estos por u_{3m}. Si se determina que la diferencia $\left(u_{3m} - u_{2m}\right)$ es suficientemente pequeña mediante un test de significancia, las tres variables involucradas se eliminan. De otro modo; se eliminan las del paso 2.

Observese que en el paso *j+1*-ésimo se han eliminado *j* variables.

8. APLICACIONES A DATOS REALES

Para ilustrar el desempeño del algoritmo propuesto por MWG, se codificó en un script del software libre R. Este es un lenguaje y entorno para cálculos estadísticos y gráficos. Es un proyecto GNU similar al lenguaje y entorno S que fue desarrollado en los Laboratorios Bell por Jhon Chambers y colegas en los años 80. R proporciona una amplia variedad de técnicas estadísticas y que, ademas, es altamente extendible mediante la carga de paquetes (packages) que resuelven una gran cantidad de problemas relacionados con las Ciencias Estadísticas.

8.1 MATERIALES

Se usarán datos referidos a un experimento de una plantación de soya (Glicine max(L) merril). Los datos constan de 110 observaciones y 16 variables.

Variable	Descripción	Variable	Descripción
X1	Num de plantas	X9	Peso de vainas con 2 sem
X2	Num de vainas	X10	Vainas con 3 semillas
X3	Num de vainas por plantas	X11	Peso de vainas con 3 sem
X4	Vainas por semilla	X12	Peso total de las vainas
X5	Peso de vainas vanas	X13	Peso de 100 vainas
X6	Vainas con una semilla	X14	Peso de sem de 100 vaina
X7	Peso de vainas con 1 sem	X15	Peso de 100 semillas
X8	Vainas con 2 semillas	Y	Peso total de las vainas

8.2 ANALISIS DE LA MULTICOLINEALIDAD

a) Índice de la Matriz de Correlación: Los autovalores de la matriz de Correlación son los siguientes

6.83	3.05	1.35	1.12	1.10	0.54	0.43	0.22	0.11	0.09	0.063	0.038	0.02	0.012	0.010

Calculamos el índice de la matriz: $\kappa = \frac{6.83}{0.01} = 627.6$ este valor nos indica que se trata de una multicolinealidad calificada de muy fuerte.

b) Factores Infladores de la Varianza: de 15 FIV, 13 son mayores que 5 el cual es un indicativo de alta multicolinealidad.

16.6	22.3	6.5	5.8	5.3	7.5	7.6	42.2	39.7	42.2	47.7	42.96	2.0	4.5	3.0

8.3 COMPARACIÓN DE LOS MODELOS

MODELO	R^2	PRESS	MSEP
MWG	0.757	9997175	90883.4
Backward	0.7606	11675083	106137.118
Stepwise	0.7568	11172101	101564.55
TRP	0.7603	11675083	106137.118

En el análisis de los modelos sugeridos por los cuatro métodos basándose en el criterio del R^2 de estimación se observa que todos los modelos se ajustan satisfactoriamente a los datos, ya que pareciera que no hay diferencias significativas entre ellos. Sin embargo el modelo obtenido por el algoritmo de MWG presenta el mejor PRESS y MSEP lo que permite calificarlo como el mejor predictor y como modelo lineal válido.

9. CONCLUSIONES Y RECOMENDACIONES

9.1 CONCLUSIONES

- No existen diferencias significativas, desde el punto de vista estadístico, en los modelos sugeridos por los cuatro métodos.
- El modelo sugerido por el algoritmo de MWG presentó características de ser el mejor predictor de todos.
- Los métodos y criterios de selección de variables continúan siendo un campo abierto de investigación, desde las dos visiones: empírico y bayesiano.

9.2 RECOMENDACIONES

- En los problemas donde se intente aplicar técnicas de selección de variables, se recomienda hacer un análisis previo de multicolinealidad, a fin de utilizar el método más idóneo de selección de variables.
- Se recomienda la aplicación del algoritmo de MWG cuando la multicolinealidad sea muy severa.
- Se recomienda el uso del paquete estadístico R, dado que, es un software gratis en la Internet con licencia GNU altamente sofisticado para propósitos científicos.

REFERENCIAS

[1] Draper & Smith (1981) *Applied Regression Analysis*. Wiley. N.Y

[2] Golub G.H y C.F Van Loan.(1996) *Matrix Computations*. Jhons Hopkins University Press.

[3] Graybill, F (1976) *Theory and aplications of linear models*. Duxbury Press

[4] Greene Williams (2000) *Econometric Analysis*. Prentice Hall.

[5] Gunst, R.F & L.R Mason (1977) *Biased Estimation in Regression: An evaluation Using Mean Square Error*, JASA, **72**, 616-628

[6] Hocking R.R (1976) *The Analysis of Selection of Variables in linear Regression*. Biometrics,**32**,1-49

[7] J.J Faraway (2004). *Linear Models with R*, Chapman Hall.

[8] Mansfield E.R, Webster J.T y Gunst (1977) *An analytic variable selection procedure for principal components regressions*, Appl, Statist,. **26**,34-36

[9] Seber, G.A.F (1977) *Linear Regression Analysis*, Wiley, N.Y

[10] Searle S.R (1982) *Matrix Algebra Useful for Statistics*, Jhon Wiley and Sons.

[11] *www.r-cran-project.org*

Autorización y Renuncia

Los autores authorizan a LACCEI para publicar el escrito en los procedimientos de la conferencia. LACCEI o los editors no son responsables ni por el contenido ni por las implicaciones de lo que esta expresado en el escrito

Authorization and Disclaimer

Authors authorize LACCEI to publish the paper in the conference proceedings. Neither LACCEI nor the editors are responsible either for the content or for the implications of what is expressed in the paper.