SELECCION DE VARIABLES EN LA REGRESION LINEAL CON EL ALGORITMO RRQR RESTRINGIDO

María Victoria Mederos y Gladys Linares, Universidad de La Habana, Cuba
Jesús López Estrada, Universidad Nacional Autónoma de México (UNAM), México

RESUMEN
El polémico problema de la selección de variables ha dado lugar a diferentes procedimientos que tratan de buscar la ecuación de regresión que mejor ajusta los datos con el menor número de parámetros. En este trabajo se presenta un nuevo procedimiento que utiliza la descomposición RRQR con pivoteo restringido combinada con un criterio empírico de selección de modelos como es el Cp de Mallows. Dos aplicaciones ilustran las ventajas de este procedimiento.

MSc: 62G05

ABSTRACT
The polemical problem of variable selection has originated different procedures when seeking for the regression equation that best fits the data with minimun number of parameters. In this paper a new procedure for variable selection is proposed, which combines the restricted RRQR algorithm with the statistical criterium of Mallows for model selection. Two applications illustrate the advantages of this procedure.

1. INTRODUCCION

El problema de la selección de variables en la regresión lineal continúa siendo objeto de atención de muchos especialistas (Montgomery and Peck, 1982; Ronchetti and Staudte, 1994). El propósito general del mismo es establecer una ecuación de regresión lineal para una variable respuesta Y en términos de ciertas variables predictoras X_1, X_2, ..., X_k (o funciones de éstas), tratando de conciliar dos criterios contrapuestos: por una parte, incluir tantas variables Xs como sea posible para que la ecuación sea útil para propósitos predictivos, y por otra parte, incluir el menor número posible de variables Xs para disminuir los costos de obtención de información. De esta manera el problema se convierte en "buscar un balance entre simplicidad y ajuste" y es precisamente a lo que nos referimos al hablar del problema de "seleccionar la mejor ecuación de regresión".

Al tratar de dar solución a este problema se corren dos riesgos, por una parte, el de incluir variables irrelevantes, y por otra, el de omitir alguna relevante. Hay que tener en cuenta la dificultad esencial que constituye el desconocimiento de la varianza de las observaciones aleatorias, lo que implica la necesidad de juicios subjetivos. Así, puede decirse que no existe un procedimiento único para seleccionar la mejor ecuación de regresión y que se continúa investigando con el fin de brindar nuevos procedimientos para la solución de este problema. En los últimos años han surgido ideas con respecto a la inclusión de criterios numéricos en el ajuste (Miller, 1990; Thisted, 1988) que pudieran contribuir de alguna manera a seleccionar la mejor ecuación de regresión en problemas específicos.

Perseguimos, en este trabajo, tres objetivos fundamentales. El primero se refiere a la valoración del algoritmo RRQR con pivoteo restringido para su aplicación en la regresión lineal. El segundo, al estudio de los principales procedimientos estadísticos que se brindan en la literatura para la selección de la mejor ecuación de regresión. Y el tercero, proponer un nuevo procedimiento para la selección de variables en la regresión, combinando aspectos numéricos y criterios empíricos de selección de modelos.

Trataremos en el epígrafe 1 las descomposiciones QR y RRQR. En el 2, estableceremos el enlace con la regresión y resumiremos los principales procedimientos de selección de modelos que brinda la literatura y que son de uso más común. Un nuevo procedimiento se explica en el epígrafe 3 y finalmente en el 4 se brindan ejemplos que permiten una comparación del procedimiento propuesto con los utilizados en la práctica.

2. LAS DESCOMPOSICIONES QR Y RRQR

La ecuación de regresión lineal

$$y = X\beta + \varepsilon,$$

con matriz de diseño X_{mm}, vector de observaciones y_{mxl} y vector de errores aleatorios ε_{mxl}, se puede resolver aproximadamente por el método de los mínimos cuadrados usando las ecuaciones normales X´Xb = X´y. Estas permiten calcular las estimaciones b de β que minimiza

$$||y - X\beta||_2,$$

lo que puede hacerse siempre que X sea de rango completo y X´X sea bien condicionada. En la práctica, entre las variables que constituyen las columnas de X puede haber *cuasi* colinealidad, y el vector de parámetros que se calcule estará muy afectado de error.

Otra vía para calcular b por el método de los mínimos cuadrados es hacer una descomposición ortogonal de la matriz X, por ejemplo la conocida como QR, con Q ortogonal y R triangular superior. Por esta vía se minimiza

$$||Q´(y - X\beta)||_2 = ||y - X\beta)||_2,$$

y es más precisa por usar transformaciones ortogonales y no construir la matriz X´X, que en general es mal condicionada. El proceso usual de triangulización en el algoritmo QR se lleva a cabo escogiendo como columna pivote de X en cada paso a la de mayor norma euclidiana. En el pivoteo restringido que proponemos, se fijan como primeras columnas de X aquellas que corresponden a un número pequeño de variables, consideradas imprescindibles según la aplicación concreta, las cuales no serán elegibles para efectuar las permutaciones que requiere la transformación. Esta puede representarse matricialmente como XP = QR, donde P representa la permutación efectuada de las columnas.

Esta transformación es válida también cuando hay colinealidad, siendo entonces el rango r de X menor que el número n de columnas. En este caso, se puede considerar R particionada en bloques:

$$R \approx \begin{bmatrix} R_{11} & R_{12} \\ 0 & 0 \end{bmatrix}$$

y si $||R_{22}||_2 < \varepsilon$ entonces

$$R = \begin{bmatrix} R_{11} & R_{12} \\ 0 & R_{22} \end{bmatrix}$$

con R_{11} triangular superior rxr, de igual rango que X. Es conocido que en este caso, entre los valores singulares de X y $||R_{22}||_2$ existe la siguiente relación

$$\sigma_1(X) \geq \ldots \geq \sigma_r(X) \geq \varepsilon > ||R_{22}||_2 > \sigma_{r+1}(X) \geq \ldots \geq \sigma_n(X),$$

lo que define teóricamente el rango r asociado a la tolerancia ε. Pero siendo $\sigma_{r+1}(X) < \varepsilon$, no siempre se cumple que $||R_{22}||_2 < \varepsilon$ (Chan, 1987), lo que implica que la descomposición QR no siempre revela el rango. De aquí, la necesidad de introducir una corrección en la descomposición QR encaminada a que la matriz R sea reveladora del rango, lo que constituye el algoritmo RRQR de Chan y Foster. Este algoritmo proporciona entonces en todo caso el ε-rango numérico de la matriz X, así como la permutación P que da el orden de importancia de las columnas, es decir, de las variables para la regresión.

3. ENLACE CON LA REGRESION

En la selección de las variables para la regresión hay que tener en cuenta dos tipos de análisis que se complementan: el de la matriz X y el de su relación con el vector y.

En análisis de la X es el realizado a partir del algoritmo RRQR, que nos proporciona la información de *cuántas* son las variables más linealmente independientes de X de acuerdo al rango numérico, y *cuáles* son estas variables según la permutación obtenida.

Para el análisis relativo al vector y, lo que interesa determinar es *cuáles* son las columnas de X que mejor describen a y, es decir, *cuáles son las variables más correlacionadas con y*, que a la vez no estén muy

correlacionadas entre sí. Este análisis constituye precisamente el objetivo general de los procedimientos estadísticos de selección de variables. A continuación presentamos un cuadro resumen de los principales procedimientos que se brindan en la literatura.

PROCEDIMIENTOS	DESCRIPCION GENERAL
Todas las Regresiones Posibles	• Calcular todos los posibles conjuntos de regresión. • Valorar cada ecuación de regresión según algún criterio empírico de selección de modelos, como el coeficiente de determinación, el coeficiente de determinación ajustado, el Cp de Mallows u otro cualquiera.
Regresión Paso a Paso Inclusión ascendente Eliminación descendente PRESS	• Emplean reglas de parada basadas en niveles de significación para pruebas de hipótesis u otros criterios.
Regresión con raíces latentes	• Combina todas las regresiones, el análisis residual y técnica de validación.
Regresión por etapas	• Es una extensión de la Regresión con Componentes Principales para examinar ecuaciones alternativas y eliminar variables predictivas. • Ajustar la ecuación con la X más correlacionada con y, calcular los residuos y considerar los residuos como variable respuesta de una ecuación con la X (de las restantes) más correlacionada con esta nueva respuesta. Continuar hasta alcanzar un estado deseado.

Ninguno de los procedimientos anteriores da una solución completamente satisfactoria. Así, en el caso de "Todas las regresiones posibles" e inclusive con alguna de sus variantes como la conocida como "Mejor Subconjunto de Regresión" existe la opinión de que toma demasiado tiempo de computación y demasiado esfuerzo para examinar todas las ecuaciones de regresión. En los populares procedimientos computacionales como la Regresión Paso a Paso, de Inclusión Ascendente, de Eliminación Descendente y diferentes variaciones de esas ideas, se hace la crítica de que falta la capacidad del analista de datos para tomar decisiones. En PRESS también hay una enorme cantidad de computación y no se dan reglas precisas para elegir el mejor modelo. En la Regresión con raíces latentes se utilizan estimaciones sesgadas arbitrariamente, además de que la búsqueda de las raíces latentes es costosa desde el punto de vista computacional. En la Regresión por Etapas no hay una solución mínimo cuadrática y esto no satisface a muchos usuarios de la regresión.

En el epígrafe siguiente exponemos un nuevo procedimiento que combina criterios numéricos con criterios empíricos de selección de modelos.

4. NUEVO PROCEDIMIENTO PARA LA SELECCION DE VARIABLES EN LA REGRESION: *SELVAR*

Este procedimiento utiliza primeramente el algoritmo RRQR restringido propuesto por nosotros para la determinación del rango de la matriz X y el orden de importancia de las variables y después calcula las distintas ecuaciones de regresión por grupos de p variables aplicando el criterio Cp de Mallows, mediante el cual se brinda la información que permite la elección del mejor modelo. Aquí p toma los valores q+1, q+2,...,r, donde q es el número m de variables priorizadas y r es el rango. El número total N de modelos que deben ser analizados es $N = 2^{r-q} - 1$, siendo r - q el número de grupos. Así, la complejidad computacional resulta ser muy inferior a la de otros procedimientos para la selección de variables en la regresión.

Para la realización del procedimiento *SELVAR* hemos elaborado un programa MATLAB que consta de los pasos siguientes:

Paso 1. Centrar y escalar los datos y hallar la matriz de correlaciones. Resolver el modelo completo (si la casi deficiencia en rango lo permite) para determinar el estimador de la varianza de las observaciones aleatorias. Si no, determinarlo en el paso 4.

Paso 2. Pedir al usuario que declare cuántas y cuáles variables considera *imprescindibles;* por defecto, tomar la más correlacionada con la variable dependiente y.

Paso 3. Aplicar el algoritmo RRQR con pivoteo restringido a la matriz X, teniendo en cuenta la priorización definida en el paso anterior, y descartar las variables correspondientes a la deficiencia en rango:

$X \rightarrow [X_q, X_k]$, donde X_q representa las columns correspondientes a las q variables priorizadas y X_k, las correspondientes a las k variables que completan el rango r.

Paso 4. Calcular la ecuación de regresión con q + k = r variables, esto es, $y = f(X_q, X_k)$, y calcular la suma de cuadrados residuales RSS y el estadístico Cp de Mallows. Si el modelo completo no se pudo resolver, determinar aquí el estimador de la varianza.

Paso 5. Obtener las ecuaciones de regresión del grupo de p = q+1 variables, aumentando la primera de las no imprescindibles, y calcularle el Cp correspondiente.

Paso 6. Decidir cuál es la mejor ecuación del grupo p, según el criterio Cp.

Paso 7. Repetir los pasos 5 y 6 para p = q + j, j = 2,3,...,k-1.

Paso 8. Escoger la mejor ecuación de todas.

5. EJEMPLOS

Analizaremos dos ejemplos. Uno, los conocidos datos de Hald (Draper and Smith, 1981) sobre enfriamiento del cemento, que ha sido usado por innumerables investigadores. Otro, los datos obtenidos por el Instituto de Investigaciones del Arroz de Cuba, en un experimento realizado en la Granja Nueva Paz de la Empresa de Semillas, provincia de La Habana, en el año 1998, con el propósito de establecer ecuaciones de predicción del rendimiento agrícola de este grano en condiciones de cultivo.

Ejemplo1. Las variables en los datos sobre enfriamiento del cemento son cuatro compuestos de calcio, y la variable respuesta *y*, el calor desprendido por gramo de cemento. En la aplicación del algoritmo RRQR con pivoteo restringido, al no tener criterio de priorización para las variables, el procedimiento *SELVAR* fija X_4 por ser la de máxima correlación con *y*. Se obtiene rango tres y permutación (4; 3, 1, 2), lo que conduce a descartar la variable X_2.

Los grupos de ecuaciones considerados fueron:

Dos modelos de dos variables (X_4, X_3 y X_4, X_1) y uno de tres (X_4, X_3, X_1).

La siguiente tabla muestra los resultados obtenidos:

p	Ecuación	RSS	Cp	|Cp - p|	Mejor por grupo
2	$y_{est} = f(X_4; X_3)$	175.7	20.4	18.4	
	$y_{est} = f(X_4; X_1)$	74.7	3.5	1.5	*
3	$y_{est} = f(X_4; X_3; X_1)$	50.8	1.5	1.5	*

Como ya hemos dicho, este ejemplo ha sido objeto de amplio tratamiento en la literatura estadística (Draper and Smith, 1981) por lo que podemos resumir los resultados reportados sobre la aplicación de otros procedimientos diferentes de selección de modelos.

PROCEDIMIENTOS		Mejor ecuación de regresión	
		1er. lugar	2do. lugar
Todas las ecuaciones	R^2	$f(X_1, X_4)$	$f(X_1, X_2)$
	S^2	$f(X_1, X_2)$	$f(X_1, X_4)$
	Cp	$f(X_1, X_2)$	
Eliminación descendente		$f(X_1, X_2)$	
Paso a Paso (hacia adelante y hacia atrás)		$f(X_1, X_2)$	
PRESS		$f(X_1, X_2)$	
Raíces latentes		$f(X_1, X_2, X_4)$	
Regresión por etapas		$f(X_1, X_4)$	
RRQR restringido	Cp	$f(X_4, X_1)$	$f(X_4, X_3, X_1)$

Puede apreciarse que según nuestro procedimiento, por simplicidad se escogería como mejor ecuación de regresión $y_{est} = f(X_4; X_1)$, y en segundo lugar, $y_{est} = f(X_4; X_3, X_1)$, que no incluye la variable X_2.

Esto se debe a que X_2 está muy correlacionada con X_4, y fue descartada, habiendo sido X_4 la variable priorizada por su correlación máxima con *y*. De ahí que consideremos de mayor calidad una ecuación con variables X_4 y X_1 como mejor modelo, y no con X_2 y X_1.

Ejemplo 2. En el problema de la predicción del rendimiento agrícola del arroz, las variables predictoras son: número de hijos (X_1), altura de la planta (X_2), materia seca (X_3), contenido de nitrógeno, fósforo y potasio en la hoja y en el tallo (X_4 hasta X_9) y la variable respuesta *y*, el rendimiento en toneladas por hectárea.

Consideramos dos variantes, una primera, con priorización de las varaibles X_2 y X_3 teniendo en cuenta el criterio del especialista combinado con la correlación con *y*, y una segunda, donde se prioriza automáticamente X_7 (contenido de nitrógeno en el tallo) por ser la más correlacionada con *y*.

Primera variante:

De acuerdo con la tolerancia prefijada, el algoritmo RRQR da rango 5 para la matriz X y permutación (2, 3; 1, 6, 5, 8, 4, 9, 7), lo que conduce a descartar las variables X_8, X_4, X_9, y X_7. Aquí debe aclararse que X_1 no fue priorizada, a pesar de que junto con X_2 y X_3 son las de fácil medición sin necesidad de análisis químicos, porque su correlación con *y* era baja. Por otra parte, X_7 fue descartada al aparecer en la última posición de la permutación, y es lo que da lugar a analizar una segunda variante.

Los grupos de ecuaciones considerados fueron:

- Tres modelos de tres variables (X_2, X_3. X_1; X_2, X_3. X_6; X_2, X_3. X_1)

- Tres modelos de cuatro variables (X_2, X_3, X_1, X_6; X_2, X_3, X_1, X_5; X_2, X_3, X_6, X_5)

- Un modelo de cinco variables (X_2, X_3, X_1, X_6, X_5).

Los modelos son en este caso los siete mencionados anteriormente ya que r = 5 y q = 2. Si lo comparamos con los 2^n - 1 = 511 para n = 9, que requeriría el análisis de todos los modelos posibles, es indudable la reducción del esfuerzo computacional que se obtiene.

Resumimos en la siguiente tabla los resultados del procedimiento:

p	Ecuación	RSS	Cp	|Cp - p|	Mejor por grupo
3	$y_{est} = f(X_2; X_3; X_1)$	1.9	36.8	33.8	
	$y_{est} = f(X_2, X_3; X_6)$	1.2	22.1	19.1	*
	$y_{est} = f(X_2, X_3; X_5)$	1.8	34.6	31.6	
4	$y_{est} = f(X_2; X_3; X_1; X_6)$	1.2	23.2	19.2	
	$y_{est} = f(X_2; X_3; X_1; X_5)$	1.8	36.6	32.6	
	$y_{est} = f(X_2; X_3; X_6; X_5)$	1.1	21.7	17.7	*
5	$y_{est} = f(X_2, X_3; X_1; X_6, X_5)$	1.1	23.3	18.3	*

Segunda variante:

Al no definir variables imprescindibles, el procedimiento prioriza automáticamente a X_7, que es la de máxima correlación con *y*. Por aplicación del RRQR se obtiene rango 5 para X y permutación (7; 3, 2, 1, 5, 4, 6, 9, 8), luego se descartan las variables X_4, X_6, X_9 y X_8.

Los grupos de ecuaciones considerados fueron:

- Cuatro modelos de dos variables (X_7; X_3; X_7, X_2; X_7, X_1 y X_7, X_5)
- Seis modelos de tres (X_7, X_3, X_2; X_7, X_3, X_1; X_7, X_3, X_5; X_7, X_2, X_1; X_7, X_2, X_5 y X_7, X_1, X_5)
- Cuatro modelos de cuatro (X_7, X_3, X_2, X_1; X_7, X_3, X_2, X_5; X_7, X_3, X_1, X_5 y X_7, X_2, X_1, X_5).
- Un modelo de cinco (X_7, X_3, X_2, X_1, X_5).

Es decir, en total N = 2^{5-1} - 1 = 15 modelos de 511 posibles. A continuación resumimos los resultados obtenidos para el mejor modelo de cada grupo al aplicar el procedimiento.

p	Ecuación	RSS	Cp	|Cp - p|
2	$y_{est} = f(X_7; X_2)$	0.90	14.0	12.0
3	$y_{est} = f(X_7; X_3, X_2)$	0.61	8.3	5.3
4	$y_{est} = f(X_7; X_3, X_2, X_1)$	0.40	5.7	1.7
5	$y_{est} = f(X_7; X_3, X_2, X_1, X_5)$	0.37	6.9	1.9

Tomando en cuenta esta tabla y la que brinda la primera variante, el especialista dispone de la información

necesaria para hacer la elección final del modelo más conveniente, según la posibilidad o no de hacer análisis de laboratorio en la región de cultivo en cuestión.

6. CONCLUSIONES

El nuevo procedimiento *SELVAR* que presentamos reduce considerablemente el número de regresiones a realizar para la elección de la mejor ecuación, al fijar como tamaño máximo el dado por el rango calculado mediante el algoritmo RRQR restringido, y las variables correspondientes a dicho rango teniendo en cuenta criterios del especialista. Además permite la intervención del factor humano en las decisiones.

La experimentación con el uso de otros criterios de selección de modelos como pudiera ser el Cp robusto, así como estudios de simulación que permitan la comparación de criterios servirá para profundizar en las posibilidades para seleccionar la mejor ecuación de regresión.

REFERENCIAS

CHAN, T.F. (1987): "Rank Revealing QR Factorization", **Linear Albegra Appl**. 88/89, 67-82.

DRAPER, N.R. and H. SMITH (1981): "Applied Regression Analysis", John Wiley & Sons, N. York.

MATLAB (1993): User´s Guide, The Math Works Inc.

MILLER; A.J. (1990): "Subset Selection in Regression", Chapman & Hall, London.

MONTGOMERY, D.C. and E. PECK (1982): "Introduction to Linear Regression Analysis", John Wiley & Sons, New York.

RONCHETTI, E. and R.G. STAUDTE (1994): "A Robust Version of Mallows´s Cp", **Journal of the American Statistical Association**, 89(426), 550-559.

THISTED, R. (1988): "Elements of Statistical Computing", Chapman & Hall, London.