anales de psicología, 1992, 8(1-2), 139-147

El Algoritmo SWEEP y el Modelo Lineal General

Juan José López García[1]

Universidad de Murcia

Resumen. Interesados en la estimación y ajuste de complejos modelos lineales según el criterio mínimo cuadrático, en este artículo se analizan y critican algunos enfoques utilizados para la docencia de aquéllos. Destacando la perspectiva de la comparación de modelos como un enfoque adecuado para el modelado estadístico, se presenta una alternativa para la docencia de modelos lineales bajo este principio, según un algoritmo numérico frecuentemente implementado en aplicaciones computacionales: el algoritmo SWEEP.
Palabras Clave: algoritmos numéricos; algoritmo sweep; modelos lineales; docencia.

Abstract. In this work we analyze some views used for teaching about estimation and fitting of complex linear models based on ordinary least squared criterion. Introducing the model comparison approach as a fundamental approach, we enfatize an alternative to teach linear models with SWEEP algorithm.
Keywords: numerical algorithms; sweep algorithm; linear models; teaching.

Introducción

No cabe duda de que el caracter aplicado de las técnicas estadísticas al uso en ciencias del comportamiento y disciplinas afines requiere de una formación acorde con la funcionalidad de las mismas. En este sentido, los clásicos desarrollos matriciales del Modelo Lineal General presentan, frecuentemente, un alto grado de complejidad para el alumno, lo que ha derivado en la docencia de la técnicas más simples, casi limitadas a modelos univariantes, la aparición de métodos alternativos centrados en fórmulas elementales, sobre todo en modelos de diseño experimental y el uso del software estadístico como alternativa docente. El procedimiento de comparación de modelos (Judd & McClelland, 1989) ofrece un enfoque conceptual claro y simple del ajuste de modelos lineales, al margen de su naturaleza, concibiendo éste como los efectos de la inclusión o eliminación de variables que componen un modelo particular. Sin embargo, este enfoque descansa sobre los principios clásicos de estimación, además de precisar el ajuste de un número elevado de modelos, tantos como fuentes de variación se hayan especificado.

Un tratamiento diferente de la estimación según el criterio mínimo cuadrático lo ofrecen los algoritmos numéricos desarrollados para computación. Por lo general, estos algoritmos descansan sobre principios diferentes que en modo alguno se corresponden con el concepto de minimización seguido en el ajuste de modelos lineales. Sin embargo, uno de estos algoritmos denominado genéricamente SWEEP porque supone una modificación sucesiva, o barrido, de una matriz, presenta una correspondencia literal con el proceso de modelado estadístico dentro del enfoque de comparación de modelos. Con un substrato numérico esencialmente simple, el algoritmo SWEEP es conceptualmente claro, aplicable sobre cualquier derivación del Modelo Lineal General y muy preciso. Estas características pueden avalar su uso como instrumento para la docencia de modelos lineales independientemente de su complejidad, afirmación que intentaremos demostrar en las páginas siguientes.

Estimación y ajuste de modelos lineales

Como es sabido, la función de estimación mínimo cuadrática permite obtener los estimadores de los parámetros poblacionales que determinan la relación funcional entre el conjunto de variables independientes o predictores y el conjunto de variables dependientes o criterios, garantizando una diferencia cuadrática mínima entre valores observados y valores esperados. De esta forma, siendo $\mathbf{X}_{(n\times q)}$ la matriz de variables independientes e $\mathbf{Y}_{(n\times p)}$ la matriz de variables dependientes, la matriz de estimadores, $\mathbf{B}_{(q\times p)}$, resulta de (Rao, 1973; Graybill, 1976):

$$\mathbf{B} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \qquad (1)$$

[1] *Dirección:* Area de Metodología de las Ciencias del Comportamiento. Facultad de Psicología. Aptdo 4021, 30080 Murcia (Spain).
 ISSN: 0212-09728.

Esta función es aplicable al margen de la naturaleza de las variables independientes. En este caso, si alguno o todos los predictores son de naturaleza categórica, se hace necesario una adecuada representación de los sujetos en los niveles o categorías de pertenencia, mediante alguno de los sistemas de codificación al uso (ficticia, de efectos u ortogonal) o mediante el escalamiento de criterio, asignando a cada sujeto la esperanza matemática de la variable dependiente para el grupo de pertenencia del sujeto en la independiente. Esta opción, sin embargo, sólo es válida ante modelos univariantes.

Junto con la función de estimación, la descomposición de la suma cuadrática total, así como la hipótesis lineal general son las dos herramientas básicas para el ajuste de todo modelo (Timm, 1975). En este sentido, la suma cuadrática total indica el error que se comete al pronosticar los valores de las variables dependientes a partir de las esperanzas matemáticas de éstas, o en otras palabras, el error de pronóstico en ausencia de modelo:

$$SCT = \mathbf{Y}'\mathbf{Y} - n\bar{\mathbf{Y}}\bar{\mathbf{Y}} \tag{2}$$

que se fracciona en:

- la suma cuadrática del modelo, definida como la porción de la suma cuadrática total reducida como consecuencia de pronosticar los valores de las variables dependientes a partir de la relación funcional definida por el modelo postulado:

$$SCM = \mathbf{B}'\mathbf{X}'\mathbf{Y} - n\bar{\mathbf{Y}}\bar{\mathbf{Y}} \tag{3}$$

- y la suma cuadrática residual, como el error cometido al pronosticar las variables dependientes a partir del modelo propuesto:

$$SCR = \mathbf{Y}'\mathbf{Y} - \mathbf{B}'\mathbf{X}'\mathbf{Y} \tag{4}$$

Estas dos cantidades, ya escalares en caso de modelos univariantes, ya matriciales en caso de multivariantes, son la base para los criterios estadísticos sobre el ajuste del modelo.

De otra parte, la hipótesis lineal general es el instrumento adecuado para verificar la significación de las fuentes de variación incluidas en el modelo especificado, así como de aquellos contrastes definidos por el investigador. En el primer caso se trata de comprobar si la ausencia o nulidad de ciertas variables o fuentes de variación representadas por vectores de codificación, aumentan de forma significativa la suma cuadrática residual obtenida. En el segundo, si una combinación aditiva de variables, o una diferencia entre éstas, repercute significativamente sobre el residual del modelo propuesto. En cualquier caso, la hipótesis lineal general pasa por la especificación de un conjunto $\mathbf{L}$ $(z \times p)$ de z vectores que indican el propósito de la hipótesis formulada, siendo la suma de cuadrados asociada a la misma:

$$SCH = (\mathbf{LB}')(\mathbf{L}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{L}')^{-1}(\mathbf{LB}) \tag{5}$$

Esta suma de cuadrados, en referencia con el residual del modelo, es la base para los criterios estadísticos que confirman el rechazo o la aceptación de la hipótesis propuesta.

La función de estimación mínimo cuadrática, junto con la descomposición de sumas de cuadrados y la hipótesis lineal general son, pues, los tres elementos necesarios y suficientes para el ajuste y estimación de cualquier aplicación derivada del Modelo Lineal General. Sin embargo, la docencia de modelos lineales estadísticos recurre habitualmente al ordenador, en concreto a los paquetes de software estadístico, como estrategia instructiva que, sin duda, ofrece una herramienta más atractiva e interactiva para el alumno comparado con las complejas derivaciones matriciales. Ahora bien, ¿hasta qué punto es adecuado desplazar la docencia de técnicas estadísticas hacia el uso de software

estadístico? Esta cuestión ha encontrado recientemente una respuesta negativa, insistiendo en el peligro que supone formar investigadores potenciales con altas habilidades en el manejo de programas informatizados y escasos o nulos conocimientos sobre las técnicas utilizadas (Searle, 1989; Thisted, 1979; Chambers, 1980; Dallal, 1988). Desde un punto de vista más técnico, y relacionado con lo anterior, puede comprobarse una situación particular del paquete SYSTAT: la estimación de modelos ANCOVA para diseños mixtos incluye la interacción entre covariantes y factores intrasujeto, que en ningún caso es correcto ya que produce un sesgo de minimización sobre las fuentes de variación de la porción intrasujetos. Por lo demás, es imposible con este paquete estimar modelos ANCOVA cuando las covariantes son distintas para cada nivel o combinaciones de niveles intrasujeto, lo cual no indica que sean inestimables. Estas situaciones ocurren también en otros paquetes estadísticos, y no son superables sin una adecuada formación previa al uso del software estadístico.

Sobre la docencia de modelos lineales

Si admitimos la necesidad de formación en técnicas estadísticas antes de ofrecer a los alumnos la posibilidad de su explotación en paquetes estadísticos y, por otro lado, el uso de este software se presenta como una alternativa para la docencia de aquéllas, asumiendo siempre su complejidad docente cuando el interés es meramente aplicado, la formación previa, por sí misma, supone una tautología. Ante esta situación, y especialmente en la docencia de modelos de diseño experimental, es frecuente proveer al alumno de todo un recetario de fórmulas magistrales con las que puede ser autosuficiente ante un diseño particular. Esta alternativa, sin embargo, conlleva varios problemas:

a) desvirtuar la esencia del modelo, estableciendo una barrera ficticia entre éstos y los de regresión o covarianza;

b) limitar la capacidad de ajuste de modelos a los incluidos en su correcto y completo conjunto de sumatorios;

c) matematizar innecesariamente la comprensión del modelo lineal.

Estos riesgos precisan de la utilización de procedimientos más cercanos a la esencia del ajuste de aplicaciones lineales. En este sentido, la perspectiva de la comparación de modelos supone un puente intermedio como técnica para la docencia de complejos modelos lineales, ofreciendo una comprensión simple del ajuste y especialmente del contraste de hipótesis de nulidad (Maxwell & Delaney, 1990).

Dado que la partición de la suma cuadrática total es el eje de la estimación de cualquier modelo lineal, ésta supone una correspondencia exacta con la intención de todo modelo, a saber, una representación, matemática en este caso, de la realidad según la expresión (Judd & McClelland, 1989):

$$\begin{aligned} SC\,Total &= SC\,Modelo + SC\,residual \\ DATOS &= MODELO + ERROR \end{aligned}$$

Cuando el modelo contiene todas las fuentes de variación especificadas por el investigador, la denominación de *modelo saturado* indica la estimación del modelo global que, tomado como referencia, permite el ajuste y significación de las distintas fuentes de variación, planteando sucesivos *modelos restringidos*, o modelos con menos componentes que el saturado. Mediante este procedimiento, las diferencias observadas en los residuales coinciden, por generalización, con la suma cuadrática de las hipótesis sobre la nulidad de las fuentes de variación eliminadas.

Consideremos, por ejemplo, un modelo univariante con cuatro fuentes de variación (A,B,C,D) en su componente sistemático. El ajuste del modelo, mediante las ec. 2 a 4, determina si los coeficientes obtenidos a través de la ec. 1 presentan una relación funcional más allá de la esperable por mero

azar. Sin embargo, éste se centra en el modelo en su totalidad y, por tanto, cualquier referencia a las fuentes de variación contenidas en él precisa de tratamientos distintos. Para éllo, para verificar que cada fuente de variación se relaciona con el criterio de una forma significativa es preciso el ajuste de un contraste de hipótesis sobre ésta, incluyendo todos aquellos vectores de la matriz X que fueron necesarios para su representación. Imagínese, entonces, que la primera fuente de variación o variable independiente (A) es cualitativa, con cinco categorías; que el resto de variables son continuas y que se ha empleado codificación de efectos para representar los sujetos en la primera variable. En estas condiciones, siendo $H_0 : a_j = 0$, determinar si los cinco niveles de la variable categórica difieren con respecto al criterio precisa de la especificación de una matriz de contraste con cuatro vectores, los requeridos para representar los sujetos en la variable categórica:

$$\mathbf{B} = \begin{pmatrix} X_0 \\ a_1 \\ a_2 \\ a_3 \\ a_4 \\ b \\ c \\ d \end{pmatrix} \qquad (6)$$

$$\mathbf{L} = \begin{pmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix}$$

$$\mathbf{LB} = 0$$

Definida la matriz de contraste, con la ec. 5 se obtiene la suma cuadrática de la hipótesis formulada y de ésta el criterio estadístico para la significación de la primera variable. Ahora bien, si la hipótesis formulada refleja la nulidad de la primera variable, la suma cuadrática obtenida equivale al efecto de inclusión de aquélla. Por tanto, siendo SCR(A,B,C,D) ó SCM(A,B,C,D) la suma cuadrática residual y del modelo saturado respectivamente y SCR(B,C,D) o SCM(B,C,D) las equivalentes al modelo restringido a las tres últimas variables, la diferencia entre las SCM o SCR coinciden con la suma de cuadrados de la hipótesis previa:

$$\text{SCH}(A = 0) = \text{SCM}(A, B, C, D) - \text{SCM}(B, C, D)$$
$$\text{SCH}(A = 0) = \text{SCR}(B, C, D) - \text{SCR}(A, B, C, D)$$

Para las demás fuentes de variación el proceso resulta idéntico. Tan sólo es preciso especificar un modelo restringido eliminando la fuente de variación de interés y comprobar el efecto de esta eliminación. Es evidente que el procedimiento de comparación de modelos es conceptualmente sencillo, además de reflejar exactamente el propósito de ajuste. Estas ventajas, sin embargo, resultan un tanto mermadas por la operatividad del proceso. Para el modelo propuesto sería necesario estimar y ajustar un total de 5 modelos, el saturado y los cuatro restringidos para cada variable independiente. Resuelto el problema conceptual, la operatividad del ajuste, si cabe, es más laboriosa que haciendo uso de la hipótesis lineal general. El instrumento ideal para el ajuste de modelos sería un procedimiento interactivo con la capacidad de incluir y eliminar fuentes de variación sin la necesidad de especificar los modelos restringidos consecuentes. En otras palabras, ¿existe alguna alternativa para la estimación y ajuste de modelos lineales con la claridad conceptual de la comparación de modelos y la simplicidad concepto-operativa de la hipótesis lineal general? Afortunadamente contamos con un instrumento de estas características, dentro de los algoritmos numéricos para minimización cuadrática.

Con la aparición y acceso masivo a los sistemas digitales para el tratamiento de información han surgido a su vez algoritmos numéricos para la estimación mínimo cuadrática que intentan, de alguna u otra forma, evitar o simplificar el cálculo de inversas matriciales. En este sentido, algoritmos como el de eliminación de Gauss-Jordan, descomposición de Cholesky, QR o SWEEP han recibido una atención especial en los últimos años (Chambers, 1977; Kennedy & Gentle, 1980; Maindonald, 1984; Thisted, 1988; Heiberger, 1989). De entre éstos, el algoritmo SWEEP presenta especiales virtudes para la docencia de modelos lineales, como tratamos seguidamente.

El algoritmo SWEEP

Desarrollado, al parecer, por Beaton (1964), el algoritmo SWEEP se basa en un conjunto limitado de operaciones concretas que conducen de forma general a la obtención de la inversa de una matriz simétrica. Utilizando un sistema de pivotaje, la aplicación del algorimto sobre una matriz simétrica $\mathbf{M}_{(k\times k)}$ requiere de k aplicaciones sobre cada uno de los k elementos de la diagonal principal establecidos como elementos de referencia o pivotes, o en otras palabras, un barrido de la matriz mediante k iteraciones centradas en los elementos de la diagonal principal. Para la iésima iteración, siendo $p = M_{ii}$ el pivote de referencia, las operaciones se limitan a transformar todos los elementos de la matriz según:

$$M_{j,k} = \frac{M_{j,k}}{p}$$

para aquéllos que comparten fila o columna con el pivote;

$$M_{j,k} = M_{j,k} - \frac{M_{j,i} * M_{i,k}}{p}$$

para los situados en filas y columnas distintas a las del pivote, y

$$M_{i,i} = -\frac{1}{p}$$

para el propio pivote.

Los elementos transformados según estas operaciones sirven de base para la siguiente aplicación del algoritmo hasta completar las k iteraciones, con lo que se obtiene la inversa, cambiada de signo, de la matriz de partida.

No es, sin embargo, la obtención de una matriz inversa el potencial subyacente a este algoritmo. Antes bien, bajo una cierta definición de la matriz de partida, su uso es especialmente útil en minimización cuadrática. En concreto, definiendo la matriz **S** como la matriz **X** ampliada a los vectores de las variables dependientes y siendo **M** la matriz de productos cruzados de **S**, donde quedan contenidas las submatrices **X'X**, **X'Y**, **Y'X** e **Y'Y**:

$$\mathbf{S} = \begin{pmatrix} X_{10} & X_{11} & \dots & X_{1Q} & Y_{11} & \dots & Y_{1P} \\ X_{20} & X_{21} & \dots & X_{2Q} & Y_{21} & \dots & Y_{2P} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ X_{n0} & X_{n1} & \dots & X_{nQ} & Y_{n1} & \dots & Y_{nP} \end{pmatrix} \qquad (7)$$

$$\mathbf{M} = \mathbf{S'S} = \begin{pmatrix} \mathbf{X'X} & \mathbf{X'Y} \\ \mathbf{Y'X} & \mathbf{Y'Y} \end{pmatrix}$$

la aplicación del algoritmo sobre la submatriz **X'X**, que se traduce en la iteración de los vectores que definen la matriz **X**, conduce a la residualización de éstos sobre la matriz **Y'Y**, a la obtención de los

estimadores del modelo en las submatrices **X'Y** e **Y'X** y al cálculo de la inversa de **X'X**, cambiada de signo, en la misma submatriz:

$$\mathbf{M}_Q = \text{SWP}(\mathbf{M}, X_0, \ldots X_Q) = \begin{pmatrix} -(\mathbf{X'X})^{-1} & \mathbf{B} \\ \mathbf{B'} & \mathbf{Y'Y} - \mathbf{B'X'Y} \end{pmatrix} \quad (8)$$

Un análisis de las operaciones utilizadas en el algoritmo (López, 1987) permite comprobar que con la primera de éstas, tomando los elementos de partida como covarianzas, se obtienen estimadores mínimo cuadráticos; con la segunda se residualiza de las variables no iteradas el efecto de las iteradas, contribuyendo la tercera al cálculo de la inversa de las submatrices iteradas. La naturaleza de estas operaciones confiere un caracter general al algoritmo desde el momento en el que la iteración de una variable particular residualiza su efecto del resto de variables, ya independientes, ya dependendientes. De esta forma, el ajuste de un modelo pasa, necesariamente, por el análisis de la contribución aislada de cada uno de los elementos que forman su componente sistemático.

El algoritmo SWEEP presenta dos ventajas esenciales comparado con otros algoritmos (Heiberger, 1989): la reversibilidad y la generalización. SWEEP es un algoritmo reversible en tanto que una variable iterada, y por tanto parcializada del resto, puede ser re-iterada, reponiendo, a su vez, el efecto de su parcialización a las demás variables. Por otro lado, ante matrices simétricas singulares, es posible la obtención de inversas generalizadas, y por tanto, la minimización cuadrática por medio de éstas. Ambas características, junto con la aplicación normal del algoritmo se han traducido en tres operadores de éste (Goodnight, 1979, 1984; López, 1992):

- *SWP proactivo*, definido anteriormente y cuya función es la parcialización de variables iteradas.
- *SWP retroactivo*, modificado del anterior para permitir su reversibilidad. Operativamente tan sólo se diferencia del operador proactivo en la primera operación, que en este caso es cambiada de signo.
- *SWP generalizado*, que se aplica en el caso de que un pivote sea nulo, señal inequívoca de la singularidad de la matriz. En tal situación, todos los elementos correspondientes a su fila y columna se hacen cero. El resultado final es un conjunto de estimadores mínimo cuadráticos por medio de un tipo de inversa generalizada. Este método es utilizado por el paquete estadístico SAS y es especialmente útil para el ajuste de modelos de diseño experimental cuando se opta por sistemas de codificación ficticia, propensos a la especificación de matrices singulares.

SWEEP como método docente

El algoritmo SWEEP, dada la naturaleza de sus operadores, es un instrumento válido e interesante para la docencia de modelos lineales; la conjunción de los operadores proactivo y retroactivo proveen al alumno de los instrumentos necesarios para el ajuste y estimación mínimo cuadrática de cualquier modelo lineal, sea cual sea su complejidad. Puesto que las operaciones requeridas son simples, la aplicación manual del algoritmo siempre es posible, incluso con una calculadora de escritorio. Sin embargo, nuestra experiencia ha demostrado que el desarrollo de programas instructivos de software, en concreto, la programación de los operadores con libertad de aplicación por parte del usuario, favorecen la comprensión del ajuste de modelos lineales en un tiempo muy inferior al usual. Lograda esta comprensión, generalizar al cálculo matricial, como nociones de referencia, resulta más asequible para el alumno.

Retomando el ejemplo anterior, el componente sistemático del modelo estaba formado por una constante (variable falsa), cuatro vectores de codificación de efectos para la primera variable, de naturaleza categórica, y tres variables cuantitativas. Añadiendo a estos ocho vectores el correspondiente a la variable criterio se obtiene una matriz ampliada, que premultiplicada por sí misma es la matriz de partida del algoritmo. De esta matriz son elementos de referencia la fila y columna de

la variable dependiente y el elemento (y, y), o elemento de la diagonal principal correspondiente a aquélla. En este elemento siempre se reflejará el efecto de parcialización de las variables que componen el modelo, mientras que en los elementos de su fila o columna se reflejarán los estimadores de las variables iteradas. Según esto, la iteración de una constante, primera variable (falsa) del modelo, supone la parcialización sobre la variable dependiente de una porción común a todos los sujetos, y por común, no puede ser más que el efecto medio de la propia variable dependiente. De esta forma, la iteración de la constante conduce a la obtención de la suma cuadrática total como suma de productos cruzados de la variable dependiente corregidos de la media. Este dato es fundamental para el resto del procedimiento y conceptualmente supone el ajuste de un modelo caracterizado por la ausencia de modelo:

$$\begin{aligned} \text{MODELO} &: Y = X_0 \\ \text{SC ERROR} &= \text{SC}(Y) = \text{SWP}(\mathbf{M}, X_0) \\ \text{SC MODELO} &= 0 \end{aligned}$$

De la matriz resultante del paso anterior, la iteración de los cuatro vectores siguientes supone la inclusión al modelo de una variable categórica con cinco niveles, o la inclusión de la primera variable independiente. Una vez iterados, el efecto sobre la variable dependiente es la parcialización de dicha variable y consecuentemente el residual de un modelo limitado a ésta:

$$\begin{aligned} \text{MODELO} &: Y = X_0 + A \\ \text{SC ERROR} &= \text{SWP}(\mathbf{M}, X_0|A) \\ \text{SC MODELO} &= \text{SC}(Y) - \text{SC}(A) = \text{SWP}(\mathbf{M}, X_0) - \text{SWP}(\mathbf{M}, X_0|A) \end{aligned}$$

Nuevamente, de la matriz resultante la iteración de la segunda variable idependiente ofrece resultados similares, residualizando de Y el efecto de la inclusión al modelo de la variable en cuestión:

$$\begin{aligned} \text{MODELO} &: Y = X_0 + A + B \\ \text{SC ERROR} &= \text{SWP}(\mathbf{M}, X_0|A|B) \\ \text{SC MODELO} &= \text{SC}(Y) - \text{SC}(A, B) = \text{SWP}(\mathbf{M}, X_0) - \text{SWP}(\mathbf{M}, X_0|A|B) \end{aligned}$$

para la tercera variable:

$$\begin{aligned} \text{MODELO} &: Y = X_0 + A + B + C \\ \text{SC ERROR} &= \text{SWP}(\mathbf{M}, X_0|A|B|C) \\ \text{SC MODELO} &= \text{SC}(Y) - \text{SC}(A, B, C) = \text{SWP}(\mathbf{M}, X_0) - \text{SWP}(\mathbf{M}, X_0|A|B|C) \end{aligned}$$

y para la última:

$$\begin{aligned} \text{MODELO} &: Y = X_0 + A + B + C + D \\ \text{SC ERROR} &= \text{SWP}(\mathbf{M}, X_0|A|B|C|D) \\ \text{SC MODELO} &= \text{SC}(Y) - \text{SC}(A, B, C, D) = \text{SWP}(\mathbf{M}, X_0) - \text{SWP}(\mathbf{M}, X_0|A|B|C|D) \end{aligned}$$

con lo que se completa el modelo especificado, obteniendo los datos para su ajuste. Conviene destacar también que el ajuste de los distintos modelos según el orden impuesto a las variables independientes permite determinar la contribución de cada variable según este orden, lo que tradicionalmente se viene en denominar *sumas de cuadrados tipo I*, obtenidas por la diferencia de lo explicado por un modelo con respecto al modelo precedente. Por lo demás, el vector fila o columna de la variable dependiente contendrá en este momento los estimadores del modelo saturado.

En estas condiciones, la significación de cada fuente de variación tan sólo precisa del ajuste de un modelo restringido, eliminando del modelo la variable de interés. Esta operación no requiere el reinicio del procedimiento puesto que el operador retroactivo cumple con esta función. Aplicado sobre la primera variable, el resultado será el residual del modelo a falta de ésta, o lo que es lo mismo, el incremento en el residual del modelo saturado si es eliminada:

$$\begin{aligned} \text{SC}(A) &= \text{SC MODELO}(A, B, C, D) - \text{SC MODELO}(B, C, D) \\ \text{SC}(A) &= \text{SC ERROR}(B, C, D) - \text{SC ERROR}(A, B, C, D) \\ \text{SC}(A) &= \text{SWP}(\mathbf{M}, X_0|B|C|D) - \text{SWP}(\mathbf{M}, X_0|A|B|C|D) \end{aligned}$$

De esta forma, la utilización del operador retroactivo sobre cada variable independiente, y la comparación con el modelo restringido respectivo, conduce a las sumas cuadráticas para el ajuste de las fuentes de variación, o tradicionalmente sumas cuadráticas tipo III. Además, ciertas modificaciones sobre la matriz de trabajo permiten el contraste de cualquier hipótesis lineal para la comparación de efectos tanto en modelos univariantes como multivariantes (López, 1992).

Discusión

El proceso de estimación y ajuste de modelos lineales resulta conceptualmente comprensible desde la perspectiva de la comparación de modelos, mediante una secuencia de diferencias entre un modelo saturado y tantos modelos restringidos como fuentes de variación estén comprendidas en su componente sistemático. Operativamente, sin embargo, precisa del ajuste de múltiples modelos, en contra de la definición matricial clásica. En estas condiciones, el algoritmo SWEEP resulta especialmente útil, ofreciendo una herramienta de trabajo asequible para el alumno tanto a nivel conceptual como operacional. Si se entiende que la definición de modelos saturados implica la eliminación de variables independientes previamente incluídas, y que un modelo saturado engloba los efectos de parcialización sobre uno o varios criterios, el algoritmo SWEEP, frente a otros algoritmos numéricos y otras técnicas de estimación, ofrece una correspondencia literal con el proceso de modelado estadístico.

Desde otro punto de vista, el aplicado, las ventajas de este algoritmo son evidentes cuando se comprueba que la mayoría de los grandes paquetes estadísticos (SAS, BMDP, SPSS, SYSTAT) lo incluyen total o parcialmente. Sin embargo, la mejor ventaja es la de ofrecer un enfoque común para el ajuste de cualquier modelo lineal desde la perspectiva del propio modelo lineal e independientemente de las características de éste y del tipo de estimación efectuada (mínimos cuadrados ordinarios, ponderados o generalizados).

Referencias

Beaton, A.E. (1964). *The Use of Special Matrix Operators in Statistical Calculus.* Princenton, NJ: Educational Testing Service.

Chambers, J.M. (1977). *Computational Methods for Data Analysis.* New York, NY: John Wiley and Sons.

Chambers, J.M. (1980). Statistical computing: history and trends. *The American Statistician, 34,* 238-43.

Dallal, G.E. (1988). Statistical microcomputing: like it is. *The American Statistician, 42,* 212-16.

Dallal, G.E. (1990). Statistical computing packages: dare we abandon their teaching to others? *The American Statistician, 44,* 265-66.

Goodnight, J.H. (1979). A tutorial on the SWEEP operator. *The American Statistician, 33,* 149-58.

Goodnight, J.H. (1984). *The SWEEP operator: its importance in statistical computing.* SAS Technical Report R-106. Cary, NC: SAS Institute.

Graybill, F.A. (1976). *Theory and Application of the Linear Model.* Belmont, CA: Wadsworth.

Heiberger, R.M. (1989). *Computation for the Analysis of Designed Experiments.* New York, NY: Yohn Wiley and Sons.

Judd, C.M. & McClelland, G.H. (1989). *Data Analysis: A Model-comparison Approach.* San Diego, CA: Hartcourt, Brace and Jovanovich.

Kennedy, W.J. & Gentle, J.E. (1980). *Statistical Computing.* New York, NY: Marcel Dekker.

López, J.J. (1987). *Un algoritmo para la solución de los coeficientes de un modelo causal.* Tesis de licenciatura no publicada. Murcia, Universidad de Murcia.

López, J.J. (1992). *SWEEP: un algoritmo para la docencia e investigación con modelos lineales.* Tesis doctoral no publicada. Murcia, Universidad de Murcia.

Maindonald, J.H. (1984). *Statistical Computation.* New York, NY: Yohn Wiley and Sons.

Maxwell, S.E. & Delanay, H.D. (1990). *Designing Experiments and Analyzing Data: A Model Comparison Approach.* Belmont, CA: Wadsworth.

Rao, C.R. (1973). *Linear Statistical Inference and its Application.* New York, NY: John Wiley and Sons.

Searle, S.R. (1989). Statistical computing packages: some words of cautions. *The American Statistician, 43,* 189-90.

Thisted, R.A. (1979). Teaching statistical computing using computer packages. *The American Statistician, 33,* 27-35.

Thisted, R.A. (1988). *Elements of Statistical Computing.* New York, NY: Yohn Wiley and Sons.

Timm, N.H. (1975). *Multivariate Analysis with Applications in Education and Psychology.* Monterey, CA: Brooks/Cole.

(Original recibido: 6-7-1992)
(Original aceptado: 5-10-1992)