APLICACION DEL METODO ABREVIADO DE DOOLITTLE A LA REGRESION LINEAL MULTIPLE

JESUS ALBERTO MELLADO BOSQUE

T E S I S

PRESENTADA COMO REQUISITO PARCIAL
PARA OBTENER EL GRADO DE

MAESTRO EN CIENCIAS
EN ESTADISTICA EXPERIMENTAL

Universidad Autonoma Agraria
Antonio Narro

PROGRAMA DE GRADUADOS
Buenavista, Saltillo, Coah.

DICIEMBRE DE 1994

Tesis elaborada bajo la supervisión del comité particular de asesoría y aprobada como requisito parcial, para optar grado de

MAESTRO EN CIENCIAS EN
ESTADISTICA EXPERIMENTAL

Asesor principal: _____
M.C. Víctor Cantú Hernández

Asesor: _____
M.C. Félix de Jesús Sánchez Pérez

Asesor: _____
M.C. Mario Cantú Sifuentes

Asesor: _____
M.C. Raúl Cesar Gozález Rivera

_____
Dr. Jesús Fuentes Rodríguez
Subdirector de Postgrado

Buenavista, Saltillo, Coah. Diciembre 1994

AGRADECIMIENTOS

Deseo agradecer a los maestros del Departamento de Estadística y Cálculo de la UAAAN por haben compartido conmigo su tiempo y sus conocimientos, especialmente a:

M.C. Felix de Jesús Sánchez Pérez

M.C. Mario Cantú Sifuentes

Dr. Rolando Cavazos Cadena

M.C. Regino Morones Reza

M.C. Emilio Padrón Corral

A todos ellos, Gracias.

COMPENDIO

APLICACION DEL METODO ABREVIADO DE DOOLITTLE A LA REGRESION LINEAL MULTIPLE

POR

JESUS ALBERTO MELLADO BOSQUE

MAESTRIA

ESTADISTICA EXPERIMENTAL

UNIVERSIDAD AUTONOMA AGRARIA ANTONIO NARRO

BUENAVISTA, SALTILLO, COAH. DICIEMBRE 1994

M.C. Víctor Cantú Hernández - Asesor

Palabras claves: Regresión múltiple, Análisis, Varianza

El método numérico desarrollado por Doolittle, ha sido ajustado a la obtención de la regresión múltiple a través de presente siglo. El método Abreviado de Doolittle (ABDO) tiene muchas ventajas respecto a otros métodos, ya que no se requieren operaciones de álgebra de matrices.

El método parte del principio de mínimos cuadrados, que obtiene el mejor estimador de los coeficientes de la regresión con la solución de la ecuación $X'Xb = X'Y$, donde X e Y son las matrices de las variables independientes y dependiente respectivamente, y b es el vector de coeficientes.

El método propone no buscar la inversa de la matriz $X'X$, sino incorporarla al vector $X'Y$ y por medio de un método numérico, de n etapas (donde n es el número de variables independientes), encontrar los coeficientes de la ecuación de la regresión.

Cada etapa es la elaboración consecutiva de modelos en donde se van incorporando cada una de las variables independientes. En cada etapa, se encuentra también los coeficientes secuenciales y la varianza que representa cada modelo respecto al anterior.

El cálculo del coeficiente de la variable x_j, es la regresión de esa variable, corregida por la regresión de las variables anteriores, respecto a la variable dependiente, así, en cada etapa se obtiene el coeficiente de una variable y el ajuste de las restantes.

El método, aparte de ser muy accesible a su elaboración manual, es un excelente instrumento para el comportamiento de un conjunto de datos.

ABSTRACT

ABBREVIATED DOOLITTLE METHOD APLICATION TO LINEAR MULTIPLE REGRESSION ANALISYS

BY

JESUS ALBERTO MELLADO BOSQUE

MASTER DEGREE

EXPERIMENTAL STATISTICS

UNIVERSIDAD AUTONOMA AGRARIA ANTONIO NARRO

BUENAVISTA, SALTILLO, COAH. DECEMBER, 1994

M.C. Víctor Cantú Hernández - Adviser

Key Words: Multiple regresion, Analysis, Variance

The numerical method developed by Doolittle, has been adjusted througout this century to get multiple regresion analysis. The abbreviated Doolittle method (ABDO) has many advantage over other methods, because it does not need the use of matrix algebra.

The method theory begins with the least squares principles, and the solution of the ecuation $X'Xb = X'Y$ provides the coeficient best estimation, where X and Y are matrix of explanatoy and response variables, b is a coeficient vector.

With this method is not necessary to obtain the inverse of the X'X matrix, instead, the matrix is joined to the XY vector to form the first stage format. The method consists of n stages (n is the number of explanatoy varibles), in each stage, one ecuation coeficient is found.

Each stage represent the incorporation of a new variable through the model, so that, the model begins with one variable and ends up with all of them. Each stage find the secuential coeficient and variance of that model over de last one.

The X_j coeficient represents the regression of that variable, corrected by the previos varibles in the regression equation, with the response variable. In each stage it is obtained one variable coeficient and the correction of the others.

The ABDO method is easy to use mannualy and is an excelent way to understand data.

INDICE DE CONTENIDO

Página

INDICE DE FIGURAS ix

INTRODUCCION. 1

REVISION DE LITERATURA. 3

DEFINICION DEL MODELO 6

DEFINICION DEL MODELO. 6

SECUENCIA DE ELABORACION 8

METODO ABREVIADO DE DOOLITTLE 12

PRINCIPIO METODOLOGICO 12

METODO DE MINIMOS CUADRADOS. 14

APLICACION DEL METODO. 14

GENERALIZACION 19

ANALISIS DE VARIANZA. 22

CONSIDERACIONES MATEMATICAS. 22

DEFINICION DE LA VARIANZA DE LA REGRESION. . 24

OBTENCION DE LA VARIANZA 26

SELECCION MODELOS CANDIDATOS 29

COEFICIENTES SECUENCIALES Y PARCIALES 32

MODELO DE MEDIA Y PENDIENTE. 32

INTERPRETACION DE COEFICIENTES SECUENCIALES .34

INTERPRETACION DE COEFICIENTE PARCIALES . . .38

VECTORES ORTOGONALES 40

OBTENCION DE LOS VECTORES ORTOGONALES 40

OBTENCION DE INTERVALOS DE CONFIANZA. 44

CONCLUSIONES 50

RESUMEN 52

LITERATURA CITADA 53

INDICE DE FIGURAS

Página

Figura 3.1 Modelo que solamente contempla la variable dependiente 9

Figura 3.2 Modelo con una variable independiente. 10

Figura 3.3 Desarrollo de modelos consecutivos 11

CAPITULO 1

INTRODUCCION

El objetivo general de la estadística, sin lugar a dudas, es elaborar modelos para poder entender los fenómenos de la naturaleza. En cada fenómeno existen cientos de factores (variables) que influyen para obtener los resultados que observan, desafortunadamente, es prácticamente imposible para el ser humano evaluar todas las variables que influyen en algún resultado de interés, es por eso que los investigadores se tienen que conformar con unas cuantas variables, y en muchos de los casos con una sola.

El análisis multivariado es importante en cualquier investigación, ya sea en el formato de varias variables independientes y una dependiente o en una independiente y muchas dependientes, ya que en términos generales, ninguna variable es totalmente indepeniente de las otras. El resultado del análisis conjunto no es la sumatoria de los resultados de los análisis individuales de cada pareja de variables, si así fuera, el trabajo del investigador sería mucho más sencillo.

En el caso específico de la búsqueda de un modelo para variables independientes y una dependiente, el análisis de regresión lineal múltiple se intuye naturalmente, ya que es el que explica ampliamente el comportamiento de los datos y permite un análisis de varianza cojunto.

Por muchos años, las pruebas multivariadas fueron poco usadas por sus complicaciones metodológicas, la complicación de usar el álgebra de matrices siempre es un buen pretexto para simplificar el análisis al modelo bivariado. Actualmente, gracias a las computadoras, los análisis complicados se han vuelto accesibles y han retomado importancia; desafortunadamente, ha quedado un hueco entre el procedimiento manual y el automático, un procedimiento no rudimentario que explique ampliamente el comportamiento de un grupo de datos.

Al revisar la propuesta del método Doolittle, no queda duda que es un eslabón entre el procedimiento manual y el automatizado, conjuntando los beneficios de ambas alternativas, es por eso que el presente trabajo, recopila, evalúa y presenta el método Doolittle y sus ajustes en la aplicación de la regresión lineal múltiple, como una valiosa alternativa para un completo análisis de regresión lineal múltiple.

CAPITULO II

REVISION DE LITERATURA

A finales del siglo pasado, por circuntancias del tiempo, era necesario diseñar procedimientos matemáticos de fácil operación, que redujeran los riesgos por errores y que optimizaran su tiempo aplicación. Doolittle, un experto en álgebra de matrices que trabajó para el departamento de Costas y Geodesia de los Estados Unidos, diseñó más de una docena de procedimientos para la solución de problemas en diferentes disciplinas.

Al desarrollarse la teoría sobre el análisis de regresión múltiple, no se vió acompañado de un método práctico para su aplicación, no fue hasta mediados de siglo cuando Anderson y Bancroft (1952) ajustaron el método propuesto por Doolittle logrando una buena aceptación.

Unos años después, Dudley (1958) publicaba un programa para computadora (de las primeras que existieron en el mundo), que utilizaba el método de Doolittle para obtener resultados de alta confiabilidad en la prueba de regresión múltiple.

Rohde y Harves (1965) aprovechan el método de Doolittle como medio para la explicación del método de mínimos cuadrados, en sus conclusiones, describen que el método utilizado es altamente descriptivo y útil para fines didácticos.

Desafortunadamente, hoy en día no abundan los libros que traten la parte metodológica de las pruebas estadísticas, por ejemplo, Searle (1980), en el capítulo 3 de su libro explica ampliamente la obtención de la regresión múltiple, pero no propone un método específico, la realización de la prueba es casi un seguimiento de los pasos teóricos.

Igualmente, Graybill (1976) explica en su libro la obtención, netamente teórica, de la realización de la prueba, sin siquiera mencionar la utilización de un método que facilite el manejo de matrices.

En la Universidad de Carolina del Norte, en la década de los setentas, un par de doctores en estadística, llamados David M. Allen y Foster B. Cady (4), utilizaron el método de solución de ecuaciones propuesto por Doolittle en la obtención de la regresión múltiple, con la ventaja de obtener del mismo método información adicional de gran utilidad.

Un problema importante al que se han enfrentado los expertos en estadística, es la definición de métodos que no requieran obtener la inversa de una matriz, ya que la mayoría de las veces no son de rango completo, lo que ocasiona encontrar inversas generalizadas en lugar de la inversa normal. Por ejemplo, Searle tiene que dedicarle un capítulo entero de su libro al caso en donde la matriz en cuestión no tiene inversa. Una ventaja importante del método Doolittle, es que puede trabajar con matrices de rango incompleto, ya que no es necesario invertir las matrices.

Dado que las matrices usadas para la regresión múltiple son simétricas, el método original de Doolittle fue modificado para el manejo de este tipo de matrices y ha sido renombrado Método Abreviado de Doolittle (ABDO).

Actualmente el uso de las computadoras han relegado los procedimientos manuales de obtención de regresiones múltiples, pero si se desea realizar el procedimiento de forma manual, el método abreviado de Doolittle ha demostrado ser el mejor en cuanto a rapidez y exactitud, con la ventaja de que se puede llevar a cabo utilizando una pequeña calculadora de escritorio aunque se tengan más de diez variables independientes. Según Goodnigth (1979), en su mayoría las computadoras actualmente usan el método "sweep operator" que metodológicamente es el más efectivo, aunque demasiado largo para hacerse manualmente y para entenderlo.

CAPITULO III

DEFINICION DEL MODELO LINEAL

Definición del Modelo

El objetivo de la regresión lineal múltiple, es encontrar la línea que mejor interprete el comportamiento de varias variables independientes respecto a una variable dependiente. Para efectos del presente trabajo la variable dependiente será representada por "Y" y cada una de sus observaciones se denotarán de la siguiente manera.

$$y_i \quad i=1..n$$

Y las variables independientes por

$$x_{ij} \quad i=1..n; \quad j=1..k$$

Donde i es el número de observación y j el número de la variable. El modelo lineal que se pretende contruir es el siguiente:

$$y_i = B_o + B_1 x_{ij} + ... B_j x_{ij} + e_i$$

Nótese que el primer parámetro B_o no está asigando a una variable y denota la intersección de la línea resultante al eje de la variable "y".

El objetivo es encontrar los coeficientes Bj de tal forma el error sea menor posible.

Para entender la estructura del modelo y su representación matricial se muestra el siguiente ejemplo: Supongamos que se tienen seis observaciones y cuatro variables independientes, entonces la representación matricial es la siguiente.

$$Y = \begin{matrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \\ y_6 \end{matrix} \quad X = \begin{matrix} x_{10} & x_{11} & x_{12} & x_{13} & x_{14} \\ x_{20} & x_{21} & x_{22} & x_{23} & x_{24} \\ x_{30} & x_{31} & x_{32} & x_{33} & x_{34} \\ x_{40} & x_{41} & x_{42} & x_{43} & x_{44} \\ x_{50} & x_{51} & x_{52} & x_{53} & x_{54} \\ x_{60} & x_{61} & x_{62} & x_{63} & x_{64} \end{matrix} \quad B = \begin{matrix} B_o \\ B_1 \\ B_2 \\ B_3 \\ B_4 \end{matrix} \quad E = \begin{matrix} e_1 \\ e_2 \\ e_3 \\ e_4 \\ e_5 \\ e_6 \end{matrix}$$

Como el primer coeficiente debe ser una constante, la matriz X se puede expresar tambien de la siguiente manera:

$$X = \begin{matrix} 1 & x_{11} & x_{12} & x_{13} & x_{14} \\ 1 & x_{21} & x_{22} & x_{23} & x_{24} \\ 1 & x_{30} & x_{31} & x_{32} & x_{33} \\ 1 & x_{41} & x_{42} & x_{43} & x_{44} \\ 1 & x_{51} & x_{52} & x_{53} & x_{54} \\ 1 & x_{61} & x_{62} & x_{63} & x_{64} \end{matrix}$$

El modelo, entonces, en notación matricial es el siguiente:

$$\underset{(n \times 1)}{Y} = \underset{(n \times p)(p \times 1)}{XB} + \underset{(n \times 1)}{E}$$

En la parte inferior se muestran las dimensiones de cada una de las matrices.

Secuencia de Elaboración

Es importante mostrar la secuencia de elaboración de una regresión lineal múltiple, ya que de eso depende el entender después el análisis de varianza correspondiente.

Supongamos que se tienen dos variables independientes (X_1 y X_2) y una variable dependiente. El modelo más sencillo que se puede utilizar con estos datos es el que no contempla ninguna variable independiente, el modelo sería:

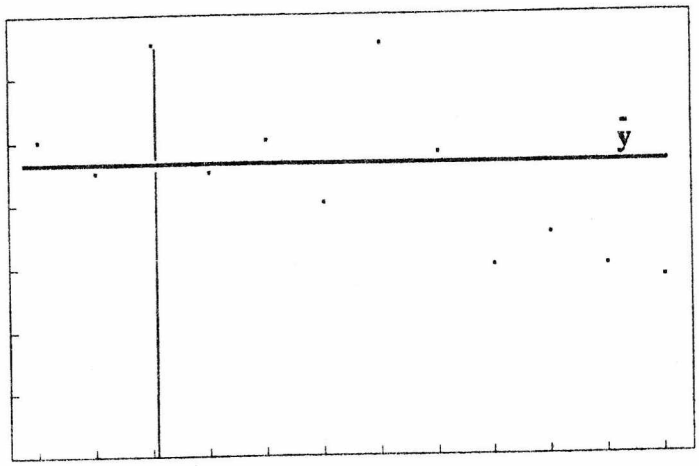
Modelo 0: $$Y_i = B_o + E_i$$

Si se aplica el método de mínimos cuadrados se podrá verificar que el mejor estimador de B_o es la media de Y

$$Y_i = \bar{y} + E_i$$

La figura del ajuste del modelo 0 es la siguiente:

Figura 3.1 Modelo que solamente contempla la variable dependiente



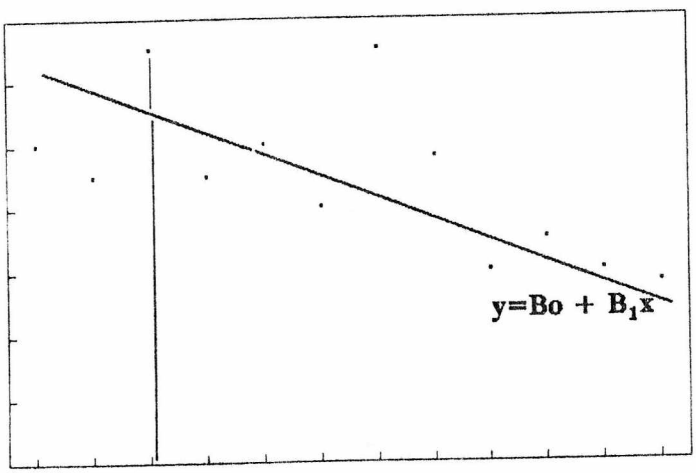
En la figura 3.1, la distancia que existe entre el eje horizontal y el punto 3 ha sido dividido en dos segmentos, el que va del eje horizontal a la línea del modelo 0 y el que va de la línea del modelo al punto en cuestión.

Obviamente este modelo es muy rudimentario, si se obtuviera la suma de cuadrados del error seguramente sería grande. El siguiente modelo a obtener contempla la utilización de la primera variable independiente.

Modelo 1: $Y_i = B_o + B_1X_1 + E_i$

Con este modelo se obtiene una línea recta que tiene una pendiente y intersección con el eje Y. La figura correspondiente es la siguiente:

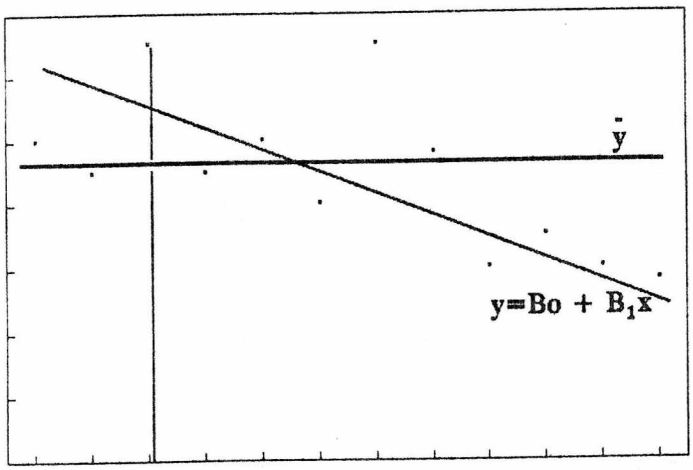
Figura 3.2 Modelo con una variable independiente



En esta figura, también se ha dividido la distancia que existe del eje horizontal al punto 3 en dos segmentos, el que va del eje a la línea y el que va de la línea al punto. como se puede inferir, en este modelo las distancias de la línea al punto (el error) debe ser menor que en el modelo 0.

Ahora, al sobreponer ambas gráficas se obtiene la figura 3.3.

Figura 3.3 Desarrollo de modelos consecutivos



Como se puede apreciar, ahora la distancia que existe entre el eje horizontal al punto 3 ha sido dividida en tres segmentos, el que va del eje horizontal a la línea del modelo 0, el que va de la línea del modelo 0 a la línea del modelo 1, y por último, el segmento que va de la línea del modelo 1 al punto.

La mejor manera de desarrollar una regresión lineal múltiple es esta última, ya que las distancias, que después se convertirán en sumas de cuadrados, van mostrando el desarrollo del análisis y van proporcionando una información valiosa para entender el comportamiento del modelo. Así, se iniciará con el primer modelo con una sola variable, luego se irán agregando variables hasta tener una línea n-dimensional, con un comportamiento explicable.

CAPITULO IV

METODO ABREVIADO DE DOOLITTLE

Principio Metodológico

Si se desea resolver el siguiente sistema de ecuaciones:

$$
\begin{aligned}
3a + 2b + c &= 1 \\
2a + b &= 0 \\
a + 2c &= 2
\end{aligned}
$$

Doolittle propone reexpresar los coeficientes de las ecuaciones mediante una matriz a la que se le ha agregado una igualdad.

$$
\begin{array}{cccc}
3 & 2 & 1 & = 1 \\
2 & 1 & 0 & = 0 \\
1 & 0 & 2 & = 2
\end{array}
$$

Como se puede apreciar, que la matriz principal es simétrica. Para resolver las ecuaciones se deberán aplicar las reglas de modificaciones de renglones, hasta lograr que en la diagonal se tengan números "1" y en el triángulo inferior ceros.

El primer renglón se divide entre 3, para lograr la unidad, luego se multiplica el primer renglón por un escalar, que al sumar al segundo renglón, el primer elemento se convierta en cero, y así sucesivamente hasta lograr modificar la primera columna. Después de las tres primeras modificaciones el resultado debe ser el siguiente:

1	2/3	1/3	=	1/3
0	-1/3	-2/3	=	-2/3
0	-2/3	5/3	=	5/3

El primer renglón ya no será modificado, porque ya tomó la forma necesaria. Ahora se modificará el segundo renglón para tener un "1" en la segunda columna, además, se harán los pasos necesarios para llegar al resultado final.

1	2/3	1/3	=	1/3
0	1	2	=	2
0	0	1	=	1

Con la estructura anterior ya es fácil despejar los valores de las variables, **a**,**b** y **c**.

Este método de solución de ecuaciones simultáneas, evidentemente sencillo, es el fundamento para el método propuesto por Doolittle.

El Método de Mínimos Cuadrados.

El método de mínimos cuadrados no será tratado en el presente trabajo, ya que es fácil consultarlo en los libros de Searle (2) en su capítulo tres y en Graybill (3) en el capítulo diez.

Lo que hay que entender es que los coeficientes de la ecuación resultante, que minimizan la distancia de la línea de regresión a los puntos originales (el error), son los que componen el vector "b" y que cumplen con la siguiente condición:

$$X'Xb = X'Y$$

Aplicación del Método

Como ya se había mencionado, el modelo de la regresión múltiple es el siguiente:

$$Y = XB + e$$

En donde B es el vector de coeficientes que se desea encontrar. Para lograrlo, se aplica el método de mínimos cuadrados, que especifica que:

$$X'Xb = X'Y$$

La B de coeficientes se ha cambiado por la b, que corresponde a los estimadores de los coeficientes.

El procedimiento lógico para resolver la ecuación de mínimos cuadrados, es despejar la b, encontrando la inversa de la matriz simétrica X'X y sustituyendo en la siguiente fórmula.

$$b = (X'X)^{-1}X'Y$$

Un problema que puede surgir al querer encontrar la inversa de la matriz X'X, es que tal inversa no exista, por ser una matriz de rango incompleto, entonces se tiene que aplicar la metodología de la inversa generalizada.

Doolittle propone no buscar la inversa, sino encontrar los estimadores b a partir de la metodología tradicional para ecuaciones simultáneas, con la facilidad de que X'X siempre es simétrica.

La ecuación de mínimos cuadrados puede expresarse como un sistema de ecuaciones simultáneas ya que el lado derecho es sólo un vector. X'Y tienen dimensiones (n x p) y (p x 1), el resultado es (n x p).

La solución del sistema de ecuaciones puede resolverse por la simple aplicación de operaciones de

renglón en las matrices, pero si se siguen pasos ordenados se puede obtener más que los simples estimadores de b, se obtiene el análisis de varianza, coeficientes parciales y otros datos de interés.

Para describir los pasos a desarrollar en el método se presenta el siguiente ejemplo, ya que es la mejor forma de entenderlo.

Sean la matriz X'X y la X'Y como se muestran a continuación.

$$X'X = \begin{matrix} 3 & -1 & 1 \\ & 5 & -1 \\ & & 3 \end{matrix} \qquad X'Y = \begin{matrix} 10 \\ -10 \\ 12 \end{matrix}$$

Dado que la matriz X'X es inversa, siempre se deberá representar sólo con la diagonal y con el triángulo superior, ya que facilita las operaciones.

El paso 0 es acomodar las matrices una después de la otra

$$\begin{matrix} 3 & -1 & 1 & 10 \\ & 5 & -1 & -10 \\ & & 3 & 12 \end{matrix}$$

Paso 1: Reescribir el primer renglón.

Paso 2: Dividir el primer renglón entre el primer elemento

de la matriz para lograr la unidad y reescribirlo.

La matriz se debe reescribir de la siguiente manera:

	3	-1	1	10
		5	-1	-10
			3	12
	---	---	---	---
Paso 1	3	-1	1	10
Paso 2	1	-1/3	1/3	10/3

Paso 3 : Se tomará como pivote el valor que está a la derecha del ^{1} obtenido en el paso anterior, el pivote en este paso es el -1/3, que deberá ser multiplicado por el renglón del paso 1 y será restado por el renglón 2 de la estructura original.

Paso 4 : Dividir el renglón del paso 3 entre el primer valor para obtener un ^{1} en la primera posición.

Los resultados se escriben de la siguiente manera:

	3	-1	1	10
		5	-1	-10
			3	12
	---	---	---	---
Paso 1	3	-1	1	10
Paso 2	1	-1/3	1/3	10/3
	---	---	---	---
Paso 3		14/3	-2/3	-20/3
Paso 4		1	-1/7	-10/7

Paso 5: Se toma como pivote el -1/7 y se multiplica por el renglón del paso 3, es decir (-1/7)(-2/3), luego se multiplica el pivote del paso 2, que es (1/3) y se multiplica por el correspondiente en el paso 1, es decir

(1/3)(1), y por último se restan ambas cantidades obtenidas al tercer renglón de la estructura original.

$$3 - (-1/7)(-2/3) - (1/3)(1) = 18/7$$

El segundo elemento se obtiene multiplicando los pivotes por los valores correspondientes a las columnas de los pasos anteriores y restándolo al correspondiente del tercer renglón original.

$$12 - (-1/7)(-20/3) - (1/3)(10) = 57/7$$

Paso 6 : Dividir el renglón del paso 5 entre el primer valor para obtener un ^{1}. El resultado debe ser como sigue:

	3	-1	1	10
		5	-1	-10
			3	12
	---	---	---	---
Paso 1	3	-1	1	10
Paso 2	1	-1/3	1/3	10/3
	---	---	---	---
Paso 3		14/3	-2/3	-20/3
Paso 4		1	-1/7	-10/7
	---	---	---	---
Paso 5			18/7	54/7
Paso 6			1	3

Con la estructura lograda se puede saber que:

$B_2 = 3$ $B_1 = -10/7$ $B_o = 10/3$

Como se puede apreciar, se tendrán 2n número de

pasos, y mientras se avanza en el desarrollo de los pasos, estos se vuelven más complejos, sin embargo, la cantidad de números que se tienen que recalcular se van reduciendo en número, de tal forma que en el penúltimo paso, que es el más largo, solamente se realiza sobre dos números. Además, es importante hacer notar que los pasos impares tienen cierto grado de complejidad, ya que los pares sólo es una división.

Generalización

Para explicar la aplicación general del método, se presenta el siguiente ejemplo: Supongamos que se tienen cuatro variables independientes y una variable dependiente. Al obtener X'X y X'Y, y al expresarlas juntas se tienen los siguientes datos:

	1	2	3	4	5	6
1.-	25	388	1,257	54.3	10.31	23.15
2.-		11,650	20,975	620.7	125.88	198.11
3.-			85,051	2,322.4	568.92	1,340.49
4.-				925.13	106.86	92.91
5.-					16.15	15.27
6.-						30.41

Los datos serán identificados por pares ordenados, el primer valor representará el renglón y el segundo la columna, por ejemplo, el dato (2,3) será el 20,975.

El dato del renglón 6, que es la sumatoria de los

cuadrados de la variable dependiente, se ha agregado al diseño que se mostró en la sección anterior para obtener cierta información del modelo.

El método aplicado debe quedar como se muestra a continuación

		1	2	3	4	5	6
	1.-	25	388	1,257	54.3	10.31	23.15
	2.-		11,650	20,975	620.7	125.88	198.11
	3.-			85,051	2,322.4	568.92	1,340.49
	4.-				925.13	106.86	92.91
	5.-					16.15	15.27
	6.-						30.41
Etapa 0							
Paso 1	7.-	25	388	1,257	54.3	10.31	23.15
Paso 2	8.-	1	19.4	62.85	2.71	0.51	1.157
Etapa 1							
Paso 3	9.-		4,122.8	-3,410.8	-432.72	-73.13	49.00
Paso 4	10.-		1	-0.82	-0.1	-0.01	0.011
Etapa 2							
Paso 5	11.-			3,226.7	-1,448.3	-139.56	-73.949
Paso 6	12.-			1	-0.44	-0.04	-0.022
Etapa 3							
Paso 7	13.-				82.19	8.5	2.015
Paso 9	14.-				1	0.10	0.024
Etapa 4							
Paso 10	15.-					2.61	0.797
Paso 11	16.-					1	0.305
Etapa 5							
Paso 12	17.-						1.052
Paso 13	18.-						1

Como se puede apreciar, el método es muy sensible a los decimales, es importante incluir la mayor cantidad de dígitos significativos en las operaciones; en el ejemplo anterior se han utilizado pocos decimales por razones de espacio, pero en los trabajos reales la exactitud debe de aumentar.

Los pasos impares tienen cierto grado de complejidad, los pasos pares son muy sencillos, sólo hay que dividir por el primer elemento.

Una vez desarrollados todos los pasos, se deben encontrar los coeficientes mediante el despeje de las ecuaciones resultantes, así por ejemplo:

$$B_4 = 0.305$$

$$B_3 + (0.1)(0.305) = 0.024$$

$$B_3 = -0.0065$$

Y así sucesivamente.

CAPITULO V

ANALISIS DE VARIANZA

Consideraciones Matemáticas

La matriz que se obtiene al multiplicar la transpuesta X consigo misma, es el conjunto de la suma de productos de cada una de las variables, y en la diagonal principal, se tienen la suma de cuadrados de cada una de las variables. En el caso de existir dos variables independientes la matriz resultante es la siguiente (la variable x_o es la unidad):

$$\begin{matrix} \sum x_o^2 & \sum x_o x_1 & \sum x_o x_2 \\ & \sum x_1^2 & \sum x_1 x_2 \\ & & \sum x_2^2 \end{matrix}$$

Como la sumatoria del cuadrado de X_o es n, la matriz se puede reescribir haciendo el cambio. Además, si se agrega la columna que corresponde al lado derecho de la ecuación, se configura como sigue.

$$\begin{matrix} n & \sum x_o x_1 & \sum x_o x_2 & \sum x_o y \\ & \sum x_1^2 & \sum x_1 x_2 & \sum x_1 y \\ & & \sum x_2^2 & \sum x_2 y \end{matrix}$$

Se revisa ahora la etapa 0, que corresponde a los pasos 1 y 2. El paso número 1 es la transcripción del primer renglón de los datos originales, no se aplica ningún procedimiento, pero al dividir todo el renglón entre n, se obtienen las medias de todas las variables. Hay que recordar que la sumatoria de cualquier variable que ha sido multiplicada por la variable cero, sigue siendo la misma variable, ya que la variable cero son todos la unidad. al desarrollar la primera etapa se tienen los siguientes datos:

paso 1:	n	$\sum x_1$	$\sum x_2$	$\sum y$
paso 2:	1	$\bar{x}_1$	$\bar{x}_2$	$\bar{y}$

En la etapa 1, donde se modificará el segundo renglón de los datos originales, se empieza con el paso 3 y con la transcripción del primer dato, que será considerado como la suma de cuadrados **no corregida**, la resta que exige el método será el factor de corrección.

$$\sum x_1^2 - x_1(\sum \bar{x}_1) = \sum (x_1 - \bar{x}_1)$$

Si se generaliza el procedimiento, se tiene que existe un renglón de sumatorias no corregida, luego se les aplica la corrección y el renglón tres es la suma de cuadrados corregida.

No corregida $\sum x_1^2 \quad \sum x_1 x_2 \quad \sum x_1 y$

Factor de corrección $\bar{x}_1(\sum x_1) \quad \bar{x}_1(\sum x_2) \quad \bar{x}_1(\sum y)$

Paso 3 $\sum (x_1 - \bar{x}_1)^2 \sum (x_1 - \bar{x}_1) x_2 \sum (x_1 - \bar{x}_1) y$

En el paso 1 no existe factor de corrección, para el paso 3 habrá un solo factor de corrección, en el paso 5 habrá dos y así sucesivamente, con lo que se puede concluir, que cada una de las columnas será la corrección consecutiva de las sumas de cuadrados de cada variable, según se vaya ajustando el modelo a uno nuevo. Además, en la última columna se tendrán los coeficientes parciales de la regresión, ya que se tendrán las siguientes ecuaciones.

$$b_j = \frac{\sum (x_{ij} - \hat{x}_{ij}) y_j}{\sum (x_{ij} - \hat{x}_{ij})^2}$$

Donde x es la corrección hasta el último modelo calculado.

Definición de la Varianza de la Regresión

La varianza de una regresión múltiple puede ser tan descriptiva como se desee, puede ser desde la más sencilla, en donde la suma de cuadrados de la variable dependiente y se descompone en dos sumas, la del cero a la línea de

regresión y en la de la línea de regresión al punto.

$$y = (\hat{y}) + (y - \hat{y})$$

Donde y es el valor estimado mediante la regresión. O se puede llegar a la descomposición más completa de la varianza, especificando la suma de cuadrados que se genera en cada uno de los modelos.

Como se había visto en la sección 2.2, la mejor forma de realizar una regresión es ir aumentando una variable a la vez e ir complementando el modelo hasta llegar a la forma final. La secuencia de modelos es como sigue:

modelo 0 : $y_i = B_o + e_i$
modelo 1 : $y_i = B_o + B_1x_1 + e_i$
modelo 2 : $y_i = B_o + B_1x_1 + B_2x_2 + e_i$
modelo 3 : $y_i = B_o + B_1x_1 + B_2x_2 + B_3x_3 + e_i$
.....

Para tener el desglose más completo de la varianza, es necesario especificar la varianza de cada uno de los modelos respecto a los anteriores, una estructura que se le puede dar al análisis de varianza es como sigue:

Fuente de Variación	Suma de Cuadrados	Grados de Libertad
Total	30.41	20
Modelo 0	26.79	1
Error	3.62	19
Modelo 1	0.58	1
Error	3.04	18
Modelo 2	1.69	1
Error	1.35	17
Modelo 3	0.04	1
Error	1.31	16

Como puede verse, cada una de las sumas de cuadrados es la descomposición del error del modelo anterior, lo que explica la secuencia de encontrar la varianza de cada modelo dados los anteriores.

Obtención de la Varianza

La varianza generada del modelo 0 (del origen a la media de la variable dependiente) es la multiplicación de los dos elementos de la derecha de la etapa 0, la varianza generada por la distancia del modelo 0 al modelo 1 se obtendrá al multiplicar los dos elementos de la derecha de la etapa 1 y así sucesivamente, la varianza del modelo M respecto al modelo anterior, será la multiplicación de los dos elementos de la derecha de la etapa M.

Siguiendo con el ejemplo iniciado en las secciones anteriores, cuyo segmento se muestra a continuación, se

pueden obtener las varianzas para el modelo 4 dado el 3, al multiplicar ambos números del lado derecho de la etapa 4.

Etapa 4----------		
Paso 10 15.-	2.61	0.797
Paso 11 16.-	1	0.305

Varianza modelo 4 = 0.797 x 0.305 = 0.243

Una vez que se han obtenido cada una de las sumas de cuadrados generados por cada modelo respecto a los anteriores, la forma correcta de agruparlos es como sigue:

Fuente de variación	Suma de cuadrados
Total (sin corregir)	30.4187
Modelo 0 (corrección)	26.7961
Total (corregido)	3.6226
Modelo	2.5699
Modelo 1 ¡ 0	0.5824
Modelo 2 ¡ 0,1	1.6947
Modelo 3 ¡ 0,1,2	0.0495
Modelo 4 ¡ 0.1.2.3	0.2433
Residuo	1.052

Es importante hacer notar que una vez obtenida la suma de cuadrados del modelo ésta no variará en los pasos siguientes, también la suma de cuadrados del residuo será contante para ese modelo, obviamente se modifica en el aumento de cada una de las etapas; las varianzas de cada uno de los modelos respecto a los anteriores, variará si se llega a cambiar el orden en que se han acomodado las variables. Ya que existen diferentes correlaciones entre

cada una de las variables y la variable dependiente.

Para poder hacer un análsis de varianza completo, es necesario completar la tabla correspondiente, incluyendo, si se quiere, solamente los modelos importantes, como se muestra a continuación.

fv	gl	sc	cm	fc	ft
Modelo 1 ¡ 0	1	0.5824	0.5824	8.308	(0.5)4.54
Modelo 2 ¡ 0,1	1	1.6947	1.6947	24.175	(0.1)8.68
Modelo 3 ¡ 0,1,2	1	0.0495	0.0495	0.706	
Modelo 4 ¡ 0,1,2,3	1	0.2433	0.2433	3.470	
Residuo	15	1.052	0.0701		

Cuando un modelo tiene una suma de cuadrados grande, significa que tiene gran influencia sobre la variable dependiente o que la variable que se incluye la explica mejor, así, se puede definir el estadístico F (la división de los cuadrados medios de cada modelo entre los cuadrados medios del error), como lo menciona Snedecor (1981) en su libro.

Si el estadístico Fc (F calculada) excede el valor de la distribución Ft (F tabular) con la significancia de 0.05 o 0.01, se puede concluir que ese modelo (o la variable que se incluye en ese modelo) explica el comportamiento de la variable dependiente.

Para mayor comprensión del estadístico F, la definición de la hipótesis nula e hipótesis alterna para la prueba, se puede revisar a Snedecor (1981).

En el presente ejemplo, las variables uno y dos tienen gran influencia, la variable 1 al 0.05 y la variable 2 al 0.01; probablemente, y para simplificar el modelo, las variables subsecuentes pueden ser eliminadas, ya que colaboran muy poco en la elaboración de un modelo que explique un comportamiento.

Selección de Modelos Candidatos

A una variable se le ha llamado x_1, a la siguiente x_2 y así sucesivamente, también en ese orden se han ido incorporando al modelo, primero la variable uno, depués la dos, etc. El valor de los coeficientes de la ecuación de predicción final no se modifican si se cambia de orden las variables, pero los demás valores sí, por ejemplo, los coeficientes secuenciales y las varianzas de cada modelo, es por eso que es necesario definir el orden en que las variables se pueden incorporar.

En la metodología Forward (stepwise) se propone ir agregando las variables según su coeficiente de correlación parcial, se inicia con la variable que tiene mayor relación con la variable dependiente y se van agregando poco a poco.

Se define también un parámetro para dejar de incorporar variables, aquellas que colaboren con menos de un valor a la explicación del modelo, ya no se incorporan.

En el método backward se propone crear el modelo completo y luego ir retirando las variables que menos colaboran con la explicación. En este modelo se tiene la ventaja de que ya se conoce, en parte, la varianza de cada uno de los modelos, además de que no existe un criterio para interrumpir el proceso de eliminación de variables.

Existen otras metodologías, como el de sweep operator, que va acomodando las variables según su comportamiento. Una forma de armar una regresión sería como sigue:

	Modelo
Agrega x_1	x_1
Agrega x_2	x_1, x_2
Quita x_1	x_2
Agrega x_3	x_2, x_3
Agrega x_1	x_1, x_2, x_3
Quita x_2	x_1, x_3
Quita x_1	x_3

Como puede observarse, la secuencia de agregado y eliminación de variables permite la identificación de todos los modelos posibles, desafortunadamente, es un proceso muy complicado que requiere de ayuda computacional, ya que se tendrán 2^n modelos si se manejan "n" variables. Si son diez las variables, existen 1,024 modelos.

Desafortunadamente, el método ABDO no provee un mecanismo para seleccionar el orden de incorporación de las variables, se pueden calcular las correlaciones parciales para seguir con el stepwise o bien, una vez elaborado el método ABDO recalcularlo con el acomodo adecuado de las variables según su varianza en el ensayo previo.

CAPITULO VI

COEFICIENTES SECUENCIALES Y PARCIALES DE LA REGRESION

Modelo de Media y Pendiente

El modelo representado por

$$Y = B_o X_o + B_1 X_1 + B_2 X_2 + \ldots.$$

Puede ser nombrado como "intersección y pendiente", ya que el primer coeficiente indica la intersección del hiperplano resultante con el eje de las variable Y, y los siguientes coeficientes representan las pendientes que tiene el hiperplano resultante respecto al eje de cada una de las variables independientes.

Existe otro modelo que es similar al de "intersección y pendiente", que puede ser llamado "media y pendiente"; su representación es la siguiente:

$$Y = \bar{Y} + B_1(X_1 - \bar{X}_1) + B_2(X_2 - \bar{X}_2) + \ldots$$

En este modelo, las variables han sido corregidas antes de ser incorporadas al modelo, lo que ocasiona que la constante (o el coeficiente de la variable X_o) sea la media de la variable dependiente. Por esta razón el nuevo modelo cambia de nombre, se sustituye "intersección" por "media". A partir de la segunda etapa los coeficientes de las siguientes variables son iguales, ya que las pendientes no se alteran al restar a cada variable sus media, lo único que cambia el primer coeficiente o constante. En el siguiente ejemplo se muestra la matriz inicial y el desarrollo del método ABDO en cierto experimento realizado con dos variables independientes

	17	1,244	400.9	703
		109,328	30,228.0	49,557
			9,555.6	16,363
Etapa 0				
Paso 1	17	1,244	400.9	703
Paso 2	1	73.17	23.58	41.35
Etapa 1				
Paso 3		18,297.05	892.45	-1,885.72
Paso 4		1	0.04	-0.10
Etapa 2				
Paso 5			57.95	-123.2
Paso 6			1	-2.12

Si en este ejemplo ajustan las variables independientes para lograr el modelo "media y pendiente", el proceso es el siguiente:

	17	0	0	703
		109,328	30,228.0	49,557
			9,555.6	16,363
Etapa 0				
Paso 1	17		0	0	703
Paso 2	1		0	0	41.35
Etapa 1				
Paso 3			18,297.05	892.45	-1,885.72
Paso 4			1	0.04	-0.10
Etapa 2				
Paso 5				57.95	-123.2
Paso 6				1	-2.12

Como puede apreciarse, los coeficientes de todas las variables se mantienen, ya que al corregir las variables independientes no cambian de pendiente, aún cuando en la etapa 0 se tienen varios ceros que alteran el inicio del procedimiento.

Interpretación de los Coeficientes Secuenciales

En el siguiente ejemplo, se pretende obtener una regresión lineal simple (una variable dependiente y una dpependiente) de ciertos datos cuya matriz y desarrollo del método ABDO es como sigue:

	17	1,244	703
		109,328	49,557
Etapa 0			
Paso 1	17	1,244	703
Paso 2	1	73.17	41.35
Etapa 1			
Paso 3		18,297.05	-1,885.72
Paso 4		1	-0.10

El coeficiente de la variable independiente es -0.10 y el valor de la constante es 48.66.

Ahora, si se agrega una segunda variable; la aplicación del método en la etapa cero y uno es de la forma siguiente:

	17	1,244	400.9	703
		109,328	30,228.0	49,557
			9,555.6	16,363
Etapa 0				
Paso 1	17	1,244	400.9	703
Paso 2	1	73.17	23.58	41.35
Etapa 1				
Paso 3		18,297.05	892.45	-1,885.72
Paso 4		1	0.04	-0.10

Nótese que en el primer ejemplo, el coeficiente de la variable ^{1} es -0.10, y en el segundo ejemplo, el coeficiente de la variable ^{1} (en el lado derecho del paso 4) sigue siendo -0.10; al momento de pasar al segundo modelo el coeficiente de la variable 1 cambiará a 0.0006.

Como puede apreciarse, al encontrar los valores de cada etapa se está encontrando los coeficientes de cada variable como si no existieran los subsecuentes, ésto porque las columnas en donde se ubican las variables que restan de incluir en cada etapa no se involucran en el lado derecho. Los coeficientes que se van obteniendo de cada variable se llamarán coeficientes secuenciales.

Como puede observarse, el ejemplo que se ha utilizado en la sección 5.1 es el mismo que en la presente sección, de tal forma que se puede hacer una comparación con los resultados de la regresión simple que se obtiene de la regresión simple:

Modelo intersección y pendiente $y = 48.66 - 0.10\ x_1$

Modelo media y pendiente $y = 41.35 - 0.10(x_1 - \bar{x}_1)$

$y = 41.35 - 0.10(x_1 - 73.18)$

En ambos modelos, el coeficiente de la variable independiente, es un coeficiente de una variable que ha sido corregida, ya sea como un paso previo al método o durante el método mismo. Al extrapolar lo antes expuesto a la segunda variable independiente, se podrá apreciar que tiene un coeficiente de -2.12, ese coeficiente es producto de una regresión simple, con la condicionante que la variable manejada, x_2, ya ha sido corregida dos veces, una en la etapa cero y otra en la etapa uno.

En cada etapa las variables sufriendo correcciones, ésto se puede ver con la variable dos del ejemplo anterior, los valores que va tomando son los siguientes:

Etapa 0 $\dfrac{\sum \bar{x}_2}{n}$

Etapa 1 $$\frac{\sum (x_1 - \bar{x}_1)\, x_2}{\sum (x_1 - x_1)^2}$$

Si se realiza una regresión simple respecto a los valores que tienen en la columna de la variable dos, de la etapa uno, se tendrá el coeficiente correspondiente. El valor corregido de cada variable, representa la fracción de los valores de las variables que no se ajustan al modelo anterior, en este caso, el valor de 0.04 de la penúltima columna en el paso cuatro, es la parte de la variable uno que no se correlaciona con la variable uno.

Como consecuencia de la conclusión anterior, cada coeficiente secuencial, es la relación entre la variable dependiente y el remanente de la variable en cuestión que no tiene correlación con los modelos anteriores.

Desafortunadamente, los coeficiente secuenciales dependen de la "secuencia del modelo", el coeficiente secuencial de una variable dos dada la uno, es diferente que el coeficiente secuencial de la variable uno dada la dos. El orden en que se acomoden las variables en la matriz simétrica será la que determine cada coeficiente secuencial.

Interpretación de los Coeficiente Parciales.

Se han obtenido las dos regresiones simples de dos variables independientes respecto a una dependiente, las ecuaciones de las lineas son las siguientes.

$$y = 41.35 - 0.10(x_1 - \bar{x}_1)$$
$$y = 41.35 - 2.12(x_2 - \bar{x}_2)$$

En ambos casos se han utilizado variables corregidas para que no exista diferencia en las constantes. Como puede verse, ambas variables independientes tienen una pendiente negativa respecto a la dependiente. Al formar la regresión múltiple con las mismas variables, se tiene la siguiente ecuación.

$$y = 41.35 + 0.00069(x_1 - \bar{x}_1) - 2.13(x_2 - \bar{x}_2)$$

La variable dos no ha sufrido cambios significativos, la que si ha cambiado radicalmente es la variable uno, ya que cambia de un coeficiente negativo a un positivo.

Una característica importante de los coeficiente parciales, es que no cambian aún cuando se cambie el orden de las variables en la matriz X.

Un coeficiente secuencial es la regresión de la variable correspondiente corregida por las restantes variables. Una variable que se coloca al final en el orden de variables, en cada uno de los pasos se irá corrigiendo, si es un coeficiente que se coloca en las primeras posiciones, la corrección se llevará a cabo al despejar los valores de los coeficientes.

CAPITULO VII

VECTORES ORTOGONALES

Obtención de los Vectores Ortogonales

En algunos libros, la obtención de vectores ortogonales es considerado como una reducción de datos, en este caso, los vectores ortogonales resultantes tienen la misma cantidad de elementos que los vectores originales, por eso se debe cuidar el concepto de reducción, se considera que el proceso es meramente una transformación.

Cuando las variables independientes es la aplicación consecutiva de un tratamiento, el manejo de los vectores ortogonales pueden ayudar para ajustar una curva que explique el comportamiento de las variables dependientes, de otra forma, los vectores ortogonales tienen pocas aplicaciones. En el presente trabajo, los vectores servirán para explicar el origen de la media y la varianza de cada una de las variables de la regresión.

Definitivamente, la metodología propuesta para la obtención de los vectores ortogonales dista mucho de ser

rápida, ya que implica muchas operaciones, aunque sencillas, por su volumen pueden complicar el proceso.

Para explicar el procedimiento, se presenta el siguiente ejemplo, donde se tienen dos variables independientes y una dependiente, las observaciones sólo son cinco para facilitar las operaciones. Nótese que la variable dependiente se ha quedado como tal, sin tomar valores específicos.

$$X = \begin{matrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \\ 1 & 4 & 16 \\ 1 & 5 & 25 \end{matrix} \qquad Y = \begin{matrix} y_1 \\ y_2 \\ y_3 \\ y_4 \\ y_5 \end{matrix}$$

Al aplicar el método ABDO a estos datos, conservando la estructura de no sustituir la variable dependiente, se tiene el siguiente resultado.

5	15	55	$y_1 + y_2 + y_3 + y_4 + y_5$
	55	225	$y_1 + 2y_2 + 3y_3 + 4y_4 + 5y_5$
		979	$y_1 + 4y_2 + 9y_3 + 16y_4 + 25y_5$
5	15	55	$y_1 + y_2 + y_3 + y_4 + y_5$
1	3	11	$(y_1 + y_2 + y_3 + y_4 + y_5)/5$
	10	60	$-2y_1 - y_2 + y_4 + 2y_5$
	1	6	$(-2y_1 - y_2 + y_4 + 2y_5)/10$
		14	$2y_1 - y_2 - 2y_3 + y_4 + 2y_5$
		1	$(2y_1 - y_2 - 2y_3 + y_4 + 2y_5)/14$

Como puede apreciarse, los coeficiente de la regresión, no son más que combinaciones lineales de los

valores de la variable dependiente.

Los coeficientes obtenidos en los pasos nones (donde todavía no se realiza la división) son los siguientes.

1	1	1	1	1
-2	-1	0	1	2
2	-1	-2	-1	2

Los vectores que forman cada uno de los renglones son ortogonales respecto a los otros, cualquier multiplicación entre ellos tendrá como resultado cero.

Era de esperarse que el primer vector representara una línea horizontal, el segundo una línea recta con cierta pendiente y la tercera una línea curva, ya que los datos fueron expresamente diseñados para mostrar los resultados obtenidos; se puede verificar que la primera variable es el vector unidad, la segunda es una recta creciente y la tercera es el cuadrado de la anterior, así se generan vectores que describen los datos originales.

Como se mencionó al principio, las cinco observaciones se han tranformado en tres vectores que describen el comportamiento de las variables originales, definitivamente ha habido una reducción en conceptos, más no en elementos, ya que se transportó de quince datos a quince datos.

Los vectores resultantes no describen íntegramente a las variables originales, el primero sí describe el comportamiento de la primera variable, el segundo vector representa a la segunda variable corregida respecto a la primera, y el tercer vector representa a la tercera variable corregida respecto a las anteriores.

Si se extraen los pasos nones del procedimiento y se agrupan uno enseguida del otro, se pueden reconocer datos importantes.

5	15	55	$y_1 + y_2 + y_3 + y_4 + y_5$
	10	60	$-2y_1 - y_2 + y_4 + 2y_5$
		14	$2y_1 - y_2 - 2y_3 + y_4 + 2y_5$

En el lado derecho se encuentran los coeficientes de los vectores ortogonales, que precisamente reciben el nombre de "coeficientes polinomiales ortogonales", los elementos de la variable original X recibirán el nombre de "coeficientes polinomiales" solamente.

En el lado izquierdo se tienen los coeficiente de las medias de cada uno de los vectores, como se muestra a continuación:

Media vector ortogonal 1	=	$5b_o + 15_1 + 55_2$
Media vector ortogonal 2	=	$10_1 + 60_2$
Media vector ortogonal 3	=	14_2

Estos resultados se obtuvieron por las reglas de las combinaciones lineales, ya que la media de una suma de variables es la suma de las medias. Y con el mismo criterio de las combinaciones lineales, la diagonal de la matriz simétrica que se forma del lado izquierdo del procedimiento, representa la desviación estándar de cada uno de los vectores, ya que es la suma de cuadrados.

Varianza vector 1 : 5 s^2
Varianza vector 2 : 10 s^2
Varianza vector 3 : 14 s^2

Donde s^2 es la varianza generada en el modelo, hay que recordar que esta varianza está absorvida por el error experimental, que, dicho sea de paso, tiene media cero.

Intervalo de Confianza para las estimaciones

Al momento de realizar una prueba de regresión se tienen dos objetivos primordiales, el entender el comportamiento de las variables y el de poder inferir nuevos resultados a partir de la ecuación obtenida. Hasta aquí ya se ha cumplido, en parte, el primer objetivo, ahora, para cumplir con el segundo objetivo hay que encontrar la seguridad que representa el modelo encontrado. Una solución al cuestionamiento es definir el procedimiento para delimitar, mediante intervalos de confianza, las estimaciones resultantes de la ecuación utilizada. El

intervalo de confianza, obviamente, depende de la variabilidad que se tenga en el modelo, un modelo más variable es más inseguro.

En una ecuación de regresión de dos variables independientes (con tres coeficientes), al probar dos valores las respectivas variables, se genera un valor en la variable de respuesta, cuyo valor esperado es sí mismo, ya que es producto de una línea que en sí es un valor esperado. Para encontrar la varianza de ese número es necesario hacer algunos cálculos.

Para facilitar los cálculos de la obtención de la varianza del error, se recomienda incluir la suma de cuadrados de la variable dependiente al final de la matriz ABDO, así, en el útimo paso se tendrá la suma de cuadrados del error, como se muestra a continuación.

	17	1,244	400.9	703
		109,328	30,228.0	49,557
			9,555.6	16,363
				30,210
Etapa 0				
Paso 1	17	1,244	400.9	703
Paso 2	1	73.17	23.58	41.35
Etapa 1				
Paso 3		18,297.05	892.45	-1,885.72
Paso 4		1	0.04	-0.10
Etapa 2				
Paso 5			57.95	-123.29
Paso 6			1	-2.12
Etapa 3				
Paso 7				683.14
Paso 8				1

La suma de cuadrados del error es el valor 683.14 que se ha obtenido en el paso siete. Los grados de libertad son el número de observaciones menos el número de modelos generados, así, la varianza del error es:

$$s^2 = \frac{683.14}{(17 - 3)} = 48.79$$

Para encontrar la varianza de un valor esperado, digamos, la media que se genera al sustituir los siguientes datos: $x_o = 1$, $x_1 = 87$, $x_2 = 24.4$, es necesario encontrar la varianza "homogénea" de cada una de las variables (ya que las variables no presentan la misma varianza). Se seleccionan los pasos nones de la tabla final de la ABDO.

Paso 1	17	1,244	400.9	703
Paso 3		18,297.05	892.45	-1,885.72
Paso 5			57.95	-123.29

De estos valores se obtienen la medias de cada uno de los vectores ortogonales.

	Media	Var
vector 1:	$17\ b_o + 1{,}244\ b_1 + 400.9\ b_2$	$17\ s^2$
vector 2:	$18{,}297.05 b_1 + 892.45\ b_2$	$18{,}297.05\ s^2$
vector 3:	$57.95\ b_2$	$57.95\ s^2$

Para localizar el vector correspondiente a los puntos $x_o = 1$, $x_1 = 87$, $x_2 = 24.4$, generado a partir de los vectores ortogonales, es necesario encontrar los factores necesario, por ejemplo, si se divide el vector uno entre 17, el coeficiente resultante es uno, lo que se estaba buscando, entonces el factor es 1/17. La varianza de la primera variable es:

$$(1/17)^2(17)\ s^2 = 2.87$$

Todo el vector uno se debe dividir entre 17, el resultado es:

	Media			Var
vector 1:	$1\ b_o$ +	$73.17\ b_1$ +	$23.58\ b_2$	$18,297.05\ s^2$
vector 2:		$18,297.05 b_1$ +	$892.45\ b_2$	$57.95\ s^2$
vector 3:			$57.95\ b_2$	

La suma del coeficiente de la segunda variable del primer vector, sumado al coeficiente de la segunda variable del segundo vector, multiplicado por algún factor, debe de resultar el coeficiente deseado: 87, expresado matemáticamente:

$$(73.17) + f\ (18,296.5) = 87$$

$$f = 13.82/18,296.5 = 0.00075$$

La varianza de la segunda variable es el factor, elevado al cuadrado, más la suma de cuadrados de la variable multiplicada por la varianza del error experimental.

$$(0.00075)^2 \ (18,296) \ s^2 = 0.5$$

Al multiplicar el segundo vector el resultado es el siguiente:

	Media						Var
vector 1:	$1\ b_o$	+	$73.17\ b_1$	+	$23.58\ b_2$		
vector 2:			$13.72\ b_1$	+	$0.67\ b_2$		
vector 3:					$57.95\ b_2$		$57.95\ s^2$

El último factor se localiza al despejar la siguiente ecuación.

$$23.58 + 0.67 + f\ (57.95) = 24.4$$

$$f = 0.00247$$

Entonces la varianza de la variable tres es:

$$(0.00247)^2 \ (57.95) \ s^2 = 0.01$$

La varianza del punto referido es la suma de las tres varianzas.

$$2.87 + 0.5 + 0.01 = 3.38$$

Para encontrar el intervalo de confianza para cada uno de los puntos que se deseen estimar, se debe buscar el grado de significancia, y buscar el valor en la distribución t-student. El intervalo de confianza es:

$$\bar{x} + \sqrt{(\text{varianza})\ t_{\alpha/2}}$$

Si se desea realizar una prueba de hipótesis acerca de algún valor estimado, el estadístico de prueba es:

$$t = \frac{\hat{y}}{\sqrt{\text{Varianza}}}$$

CONCLUSIONES

El uso del método ABDO para el cálculo de los primeros resultados, es decir, en la obtención de la ecuación de la regresión y la varianza, es muy sencillo y útil, sin embargo, para poder describir el comportamiento de las variables, se requieren de conocimientos profundos sobre el álgebra de matrices así como de los espacios vectoriales, lo que vuelve al método innaccesible para muchas personas.

Para un análisis completo de un conjunto de variables, no existe mejor opción que el uso de la computadora, pero para poder interpretar los resultados que pudiera generar un programa computacional, es necesario conocer de la prueba, y para conocer la prueba lo mejor es usar un método como el de ABDO, que no cae en los tecnicismos y en la frialdad de la teoría estadística, además de que promueve el entendimiento mediante la práctica.

En conclusión, el método ABDO es un excelente método para entender la regresión lineal múltiple, pero para los trabajos cotidianos, se recomienda un equipo computacional.

Dada la naturaleza de la Universidad Autónoma Agraria Antonio Narro, donde se requieren conocimientos prácticos y la preparación estadística de sus alumnos no es profunda, es importante definir un método como el ABDO para la capacitación, después, se podrá revisar software para la aplicación de la prueba e interpretación de resultados.

RESUMEN

El método abreviado de Doolittle (ABDO) es un procedimiento de métodos numéricos, que resuelve regresiones lineales múltiples al encontrar la ecuación de predicción y la varianza adjunta de cada uno de los modelos generados.

Su forma de operar es la elaboración de modelos lineales consecutivos que van aumentando de complejidad hasta llegar al modelo final. La forma de encontrar los coeficientes es despejando las ecuaciones secuenciales generadas.

El método es fácil de manejar y de entender, cada uno de sus pasos es una explicación del comportamiento de los datos, permite llegar a un análisis extenso, aunque para encontrar detalles muy específicos, el método se complica y requiere de conocimientos especiales para poderlo interpretar.

LITERATURA CITADA

Allen M.D. and Foster B.D., 1982, Analyzing Experimental Data by Regresion, Lifetime Learning Publications.

Anderson R.L and T.A. Bancroft, 1952, Statistical Tehory in Research, New York, McGraw-Hill.

Dudley J., 1958. A Procedure for Computing Regression Coefficients. JASA, vol 53, p 144-150

Graybill F.A.,1976, Theory and Application of the Linear Model, Duxbury Press.

Goodnight, James. 1979. A tutorial om the sweep operator. American Statician, v3, p149-158

Rohde, C.A., and Harves, J.R., 1965. Unified Least Squares Analysis, JASA, vol 60, p 523-527

Searle S.R.,1980, Linear Models, Ed. Bradley-Hunty-Kendall

Snedecor, G.W., and Cochran W.G.,1981, Statistical Methods, Ed. Iowa State University Press.