Anales de la Real Academia de Doctores
Volumen 5, pp. 165-197, 2001

BÚSQUEDA DE UNA ESTRATEGIA METODOLÓGICA PARA RELACIONAR VARIABLES CUANTITATIVAS EN SITUACIONES LÍMITE

FCO JAVIER DÍAZ-LLANOS Y SAINZ-CALLEJA
CARMEN CERMENO CARRASCO

RESUMEN

El establecimiento de interrelaciones de una variable cuantitativa a explicar y con, p variables explicativas cuantitativas:

$$x_1, x_2, \ldots, x_p$$

(medidas en una muestra de tamaño n), mediante el **método de mínimos cuadrados** - «EN SITUACIONES LÍMITE» como: **presencia de multicolinealidad, número menor de individuos que variables o,datos ausentes -,** ofrece resultados —DE PREDICCIÓN— poco comprensibles para el investigador. Sin embargo, el **método de mínimos cuadrados parciales**, resuelve dichas situaciones con exito.En el presente trabajo, analizaremos ambos **métodos** para variables cuantitativas-**explicativas**-y,en especial, el de **mínimos cuadrados parciales** y,finalmente, aplicaremos 4 estrategias cuando estas variables están altamente correlacionadas.

INTRODUCCIÓN

El problema del ajuste de un conjunto de puntos representados en un sistema de ejes coordenados por una recta o —más generalmente— por una curva, «sensus lato»,era objeto de estudio desde mediados del siglo XVIII(Ledonhard EULER, 1749; Johan Tobías MAYER,1750).Sin embargo, la primera mención al **método de mínimos cuadrados,** fue atribuido a Adrien-Marie LEGENDRE (1805).En dicho estudio, se consideró este **método** como: «el más adecuado para relacionar variables de forma lineal» señalándose, —además— la conveniencia de la eliminación de individuos atípicos para optimizar el establecimiento de dichas interrelaciones. Dichos resultados, son concordantes con lo encontrado, actualmente (CERMEÑO y DÍAZ-LLANOS, 2000). Mas tarde, Benjamín PEIRCE (1852) (según apuntaría posteriormente, STIGLER,1873), establece un primer criterio para la detección de datos atípicos —el cual— fue criticado por AIRY 4 años más tarde (1856).

Por otro lado, LEGENDRE, utilizó en su realización un Teorema fundamental. Dicho Teorema, fue reivindicado - como suyo - por Carl Friedrich GAUSS, un año después(1806) quien, además, reivindicó su utilización desde 12 años antes de la publicación de LEGENDRE en 1805.

Por último, merece la pena destacar la introducción del **método de mínimos cuadrados,** realizada por Robert ADRAIN(1808), quien, aportó un punto de vista de gran interés, complementario al de los trabajos realizados por sus antecesores.

Sin embargo,dicho **método, no pudo ser justificado** hasta la llegada de la **ley de LAPLACE-GAUSS,** «bautizada» por Karl PEARSON, a finales del siglo XIX(1893), como «**ley normal**».

PASADO, PRESENTE Y FUTURO SOBRE LAS LIMITACIONES PARA RELACIONAR VARIABLES: se ha constatado, en numerosas ocasiones que, la presencia de la **multicolinealidad** conlleva a situaciones de **«inestabilidad» de los coeficientes de regresión** y,que,estos, pueden ser «no significativos» cuando las variables explicativas están muy correlacionadas con la variable a explicar, produciendo dificultades de interpretación de la ecuación de regresión lineal a causa, de signos erráticos en los coeficientes de regresión y, por tanto, la aplicación del **método de mínimos cuadrados,** conduce a resultados poco comprensibles para los investigadores que se dedican a las ciencias experimentales.

Aunque, es **interesante** no sólo **detectar la multicolinealidad** (BELSLEY,KUH y WELSH, 1980; FOUCART, 1992; TOMASSONE y colaboradores, 1993; ERKEL-ROUSSE,1994/1995,1995), sino también, **tomar medidas para atenuarla** (CAZES,1991;FOUCART,1992;IEMMA y PALM,1995) sin embargo, **la ecuación de predicción lineal** bajo estas medidas sigue siendo —en ocasiones— poco comprensible para el investigador.

Otras dos **situaciones límite** son:1)**número menor de individuos que variables** y **2)datos ausentes.**

En cuanto a la 1.ª: una situación que contempla menos individuos que variables, conlleva —sistemáticamente— a que el determinante de la matriz

$$X^T X$$

—que hay que resolver para la obtención de los coeficientes de regresión— «sea nulo» y, por tanto, no haya modo de encontrar tales coeficientes.

Así como, la situación anterior no ha sido estudiada por su propia trivialidad ésta, sí lo ha sido.

En cuanto a los **datos ausentes:**

El algoritmo NIPALS(Nonlinear iterative Partial Least Squares,H.WOLD y E.LYTTKENS,1969), permite la realización de un análisis de datos ausentes —**en componentes principales—** del triplete estadístico

$$(X, Q, D)$$

donde,

X: es la matriz de datos cuantitativos de dimensiones nxp.

Q: es la métrica que permite calcular la distancia entre individuos de dimensiones pxp.

D: es la métrica que permite calcular la distancia entre variables de dimensiones nxn.

En dicho análisis, no es necesario la supresión de los individuos, con datos ausentes ni, la estimación de dichos datos.

Debido a estos hechos, no comprendemos la «escasa difusión» de dicho algoritmo; el cual, sin embargo, si se encuentra contenido en el paquete de programas SIMCA (Svante WOLD, en sus tres versiones:1991,1996 y 1998,respectivamente).

De todos estos resultados concluimos que:el **método de mínimos cuadrados** —intensamente utilizado para relacionar variables— no funciona bien en las situaciones límite tales como:la **multicolinealidad, menor número de individuos que variables** y **datos ausentes, aconsejamos sustituirlo** por el:**método de mínimos cuadrados parciales.**La **regresión PLS** fue propuesta por Svante WOLD, Harald MARTENS y Herman WOLD en el año 1983.En este mismo año, Svante WOLD y colaboradores (1983) muestran la **regresión PLS** y su utilidad en las ciencias experimentales.

El hecho de que en octubre del año 1999 se celebrase un Simposium sobre los **métodos PLS** en Jouy-en-Josas (Francia), es un indicador positivo de que, los **métodos PLS** —son y serán— el punto de mira cuando se quiera **relacionar variables de cualquier naturaleza, bajo un contexto lineal o no lineal.**

Como aplicación al modelo no lineal, no contemplado en el libro de Michel TENENHAUS (1998) hemos de indicar que, en mayo del año 2000 en las XXXII Journées de Statistique en Fès-Maruecos, éste presentó una comunicación sobre la **regresión logística PLS.**Una línea de interés, es el estudio de la **regresión PLS2** cuando los dos bloques de variables sean cualitativas.Que nosotros sepamos, esta línea de investigación se está desarrollando en Francia bajo distintos enfoques (Michel TENENHAUS,1995,1998; Jérôme PÀGES y Michel TENENHAUS,1999;Pierre CAZES,1996,1997).

UNA REFLEXIÓN EN CUANTO A LA NORMALIZACIÓN DE LOS DATOS

Vamos a presentar a continuación dos tipos de **normalización de datos** que se encuentran con frecuencia en las referencias bibliográficas y en especial en (AUDRAIN,LESQUOY-de TURCKHEIM, MILLER y TOMASSONE,1992,pp.179-180).

— El primero, consiste en restar para cada una de las variables su media, y dividir por la raíz cuadrada de la suma de cuadrados de las desviaciones a su media.

$$y_i : y_i^{[1]} = \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^{i=n} (y_i - \bar{y})^2}} \quad (i = 1,2,...,n)$$

$$x_{ij} : x_{ij}^{[1]} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^{i=n} (x_{ij} - \bar{x}_j)^2}} \quad (i = 1,2,...,n;\ j = 1,2,..,p)$$

— El segundo, consiste en restar para cada una de las variables su media y, dividivir por la raíz cuadrada de la suma de cuadrados de las desviaciones a su media dividido por (n-1).

$$y_i : y_i^{[2]} = \frac{y_i - \bar{y}}{\sqrt{\sum_{i=1}^{i=n} \frac{(y_i - \bar{y})^2}{n-1}}} \quad (i = 1,2,...,n)$$

$$x_{ij} : x_{ij}^{[2]} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\sum_{i=1}^{i=n} \frac{(x_{ij} - \bar{x}_j)^2}{n-1}}} \quad (i = 1,2,...,n;\ j = 1,2,..,p)$$

Operaciones intermedias

Aunque, las operaciones intermedias que hay que realizar para llegar a los coeficientes de regresión difieran del tipo de normalización de los datos, los coeficientes de regresión asociados a las variables

$$x_1^{[1]}, x_2^{[1]}, ..., x_p^{[1]}$$

y a las variables

$$x_1^{[2]}, x_2^{[2]}, ..., x_p^{[2]}$$

son los mismos.

— En la primera normalización de los datos las operaciones son las siguientes:

$$\sum_{i=1}^{i=n} (y_i^{[1]})^2 = 1 \quad (i = 1,2,...,n)$$

$$\sum_{i=1}^{i=n} (x_{ij}^{[1]})^2 = 1 \quad (i=1,2,...,n;\ j=1,2,..,p)$$

$$\sum_{i=1}^{i=n}\left(y_i^{[1]}\, x_{ij}^{[1]} \right) = r_{y,x_j} \quad (i=1,2,...,n;\ j=1,2,..,p)$$

$$\sum_{i=1}^{i=n}\left(x_{ij}^{[1]}\, x_{ij'}^{[1]} \right) = r_{x_j,x_{j'}} \quad (i=1,2,..n;\ j \neq j')$$

— En la segunda **normalización de los datos** las operaciones son las siguientes:

$$\sum_{i=1}^{i=n}\left(y_i^{[2]} \right)^2 = (n-1) \qquad (i=1,2,...,n)$$

$$\sum_{i=1}^{i=n}\left(x_{ij}^{[2]} \right)^2 = (n-1) \qquad (i=1,2,...,n;\ j=1,2,..,p)$$

$$\sum_{i=1}^{i=n}\left(y_i^{[2]}\, x_{ij}^{[2]} \right) = (n-1)\, r_{y,x_j} \quad (i=1,2,..,n;\ j=1,2,..,p)$$

$$\sum_{i=1}^{i=n}\left(x_{ij}^{[2]}\, x_{ij'}^{[2]} \right) = (n-1)\, r_{x_j,x_{j'}} \quad (i=1,2,...,n;\ j \neq j')$$

Cálculo de los coeficientes de regresión

Tanto en la primera, como en la segunda **normalización de los datos**, los coeficientes de regresión afectados, tanto por unas como por otras variables, son los mismos.

Partimos de la expresión que nos permite calcular los coeficientes de regresión.

Para la primera normalización:

$$\hat{\beta}^{[1]} = \left(X^{[1]^T}\, X^{[1]} \right)^{-1} X^{[1]^T}\, y^{[1]}$$

Dado que

$$X^{[1]^T}\, X^{[1]} = R$$

$$X^{[1]^T}\, y^{[1]} = \begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \\ \cdot \\ r_{y,x_p} \end{pmatrix}$$

la fórmula para el cálculo de los coeficientes de regresión es la que a continuación mostramos,

$$\hat{\beta}^{[1]} = R^{-1} \begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \\ . \\ . \\ r_{y,x_p} \end{pmatrix}$$

Para la segunda normalización:

$$\hat{\beta}^{[2]} = \left(X^{[2]^T} X^{[2]} \right)^{-1} X^{[2]^T} y^{[2]}$$

Dado que

$$X^{[2]^T} X^{[2]} = (n-1) R$$

$$X^{[2]^T} y^{[2]} = (n-1) \begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \\ . \\ . \\ x_{y,x_p} \end{pmatrix}$$

la fórmula para el cálculo de los coeficientes de regresión es la que a continuación mostramos:

$$\hat{\beta}^{[2]} = R^{-1} \begin{pmatrix} r_{y,x_1} \\ r_{y,x_2} \\ . \\ . \\ r_{y,x_p} \end{pmatrix}$$

De lo que concluimos que los coeficientes de regresión son un invariante con respecto al tipo de normalización:

$$\hat{\beta}^{[1]} = \hat{\beta}^{[2]}$$

Ecuación de predicción lineal en función de las variables normalizadas por el primer caso.

$$y^{[1]^*} = \sum_{j=1}^{j=p} \hat{\beta}_j^{[1]^*} \, x_j^{[1]}$$

Deshaciendo el cambio, llegamos a la ecuación de predicción en función de las variables originales

$$y^* = \left(\bar{y} - \sum_{j=1}^{j=p} \sqrt{\frac{SCD_y}{SCD_{x_j}}} \, \hat{\beta}_j^{[1]^*} \, \bar{x}_j \right) + \sum_{j=1}^{j=p} \sqrt{\frac{SCD_y}{SCD_{x_j}}} \, \hat{\beta}_j^{[1]^*} \, x_j$$

Ecuación de predicción lineal en función de las variables normalizadas por el segundo caso.

$$y^{[2]^*} = \sum_{j=1}^{j=p} \hat{\beta}_j^{[2]^*} \, x_j^{[2]}$$

Deshaciendo el cambio, llegamos a la ecuación de predicción, en función de las variables originales

$$y^* = \left(\bar{y} - \sum_{j=1}^{j=p} \sqrt{\frac{SCD_y}{SCD_{x_j}}} \, \hat{\beta}_j^{[2]^*} \, \bar{x}_j \right) + \sum_{j=1}^{j=p} \sqrt{\frac{SCD_y}{SCD_{x_j}}} \, \hat{\beta}_j^{[2]^*} \, x_j$$

Dado que

$$\hat{\beta}_j^{[1]^*} = \hat{\beta}_j^{[2]^*} \quad (j = 1,..,p)$$

se concluye que, el tipo de normalización, no influye en la ecuación de predicción lineal.

VARIABLES EXPLICATIVAS ORTOGONALES

Cuando todas las variables explicativas

$$x_1 , x_2 , \ldots , x_p$$

sean ortogonales —condición que siempre se verifica en la **regresión en función de las componentes principales**— los coeficientes de regresión asociados a cada variable, no son más que los coeficientes de correlación lineal de BRAVAIS-PEARSON entre la variable a explicar y las variables explicativas respectivamente.

La ecuación de predicción lineal, en función de las variables normalizadas por el primer caso se expresa de la siguiente manera

$$y^{[1]^*} = \sum_{j=1}^{j=p} r^*_{y,x_j}\, x_j^{[1]}$$

Deshaciendo el cambio, llegamos a la ecuación de predicción, en función de las variables originales

$$y^* = \left(\bar{y} - \sum_{j=1}^{j=p} \sqrt{\frac{SCD_y}{SCD_{x_j}}}\, r^*_{y,x_j}\, \bar{x}_j \right) + \sum_{j=1}^{j=p} \sqrt{\frac{SCD_y}{SCD_{x_j}}}\, r^*_{y,x_j}\, x_j$$

La ecuación de predicción lineal en función de las variables normalizadas en el segundo caso, se expresa de la siguiente manera:

$$y^{[2]^*} = \sum_{j=1}^{j=p} r^*_{y,x_j}\, x_j^{[2]}$$

Deshaciendo el cambio, llegamos a la misma ecuación de predicción, en función de las variables originales que en el caso anterior.

PREFERENCIA POR UNA Y OTRA NORMALIZACIÓN DE LOS DATOS PARA DISTINTOS INVESTIGADORES

Dentro del conjunto de investigadores que escriben sobre el tema de Análisis Estadístico Multidimensional, unos utilizan el primer tipo y otros el segundo.

Así,citaremos, entre los que utilizan el primero a: DRAPER y SMITH (1981, p.263), IEMMA y PALM (1995,p.3) y, MONTGOMERY y RUNGER (1996, p.614) y, a: TENENHAUS, GAUCHI y MENARDO (1995, p.20), Louis LEGENDRE y Pierre LEGENDRE (1998, pp.132-139), JAMBU (1999, p.64) para el segundo tipo.

JAMBU justifica en su libro por qué divide por (n-1) y no por n para el cálculo de la varianza.

En el paquete de programas STATITCF 4.0(1991), la normalización de datos se hace mediante el segundo tipo.

Hacemos alusión a este programa porque, parte de los resultados que contemplamos en el ejercicio práctico contenido al final de esta nota, ha sido realizado con el programa de regresión lineal múltiple contenido en el mismo cuyo responsable científico es TOMASSONE(1987).

EL ALGORITMO DE REGRESIÓN PLS1

La exposición de la **regresión PLS** —que nosotros sepamos— se contempla tanto en libros como, en artículos en inglés y francés.En cuanto concierne a libros en inglés hemos de destacar, el de WOLD(1885), el de LOHMÖLLER(1989) y el de MARTENS

y NEAS(1989) y, en francés, el de BRY(1996,pp.73-82) y el de TENENHAUS (1998). En lo que se refiere a los artículos en ingles, hemos de destacar el de BOOKSTEIN (1982) y el de HÖSKULDSSON (1988) y, en francés, el de GAUCHI,MENARDO y TENENHAUS (1995).

Mientras que en la **regresión PLS1**, hay tan sólo una variable a explicar y p explicativas, en la **regresión PLS2** hay q variables a explicar (q>1) y p variables explicativas.Por el hecho de no estar contemplada en castellano, desarrollaremos en esta nota —de la manera lo más didáctica posible— la **regresión PLS1** lineal.

ETAPAS DE LA REGRESIÓN PLS1

1. CONSTRUIR LA PRIMERA COMPONENTE

$$t_1$$

DEFINIDA POR LA SIGUIENTE MANERA,

$$t_1 = w_{11}\, x_1^{[2]} + w_{12}\, x_2^{[2]} + \ldots + w_{1p}\, x_p^{[2]}$$

$$t_1 = \sum_{j=1}^{j=p} w_{1j}\; x_j^{[2]}$$

donde,

$$w_{1j} = \frac{cov(x_j^{[2]}, y^{[2]})}{\sqrt{\sum_{j=1}^{j=p} cov^2(x_j^{[2]}, y^{[2]})}} = \frac{< y^{[2]}, x_j^{[2]} >}{\sqrt{\sum_{j=1}^{j=p} \left[< y^{[2]}, x_j^{[2]} >\right]^2}} \quad (j=1,\ldots,p)$$

De esta fórmula se deduce que para obtener los coeficientes

$$w_{1j}\ (j = 1,\ldots, p)$$

tan sólo hay que calcular los siguientes productos escalares clásicos:

$$< y^{[2]}, x_j^{[2]} >\ \ (j = 1,\ldots, p)$$

DETECCIÓN DE INDIVIDUOS ATÍPICOS

La regla general de decisión para la detección de individuos atípicos esta basada en que la variable aleatoria

$$\frac{n(n-H)}{H(n^2-1)} T_i^2$$

sigue la ley de FISHER-SNEDECOR con H grados de libertad para el numerador y n-H grados de libertad para el denominador (MASSON, TRACY y YOUNG,1992), donde,

$$T_i^2 \text{ es la } T^2$$

de HOTELLING de la observación i, calculada utilizando H componentes siendo igual a

$$T_i^2 = \frac{n}{(n-1)} \sum_{h=1}^{h=H} \frac{t_{ih}^2}{s_h^2} \quad (i = 1,...,n)$$

donde

n : *es el número total de individuos*

$\|t_h\|^2$: *es la norma al cuadrado de la componente h*

s_h^2 : *es la varianza (con división n - 1) de la componente h*

DETECCIÓN DE INDIVIDUOS ATÍPICOS EN LA PRIMERA COMPONENTE

La regla general de decisión para la detección de individuos atípicos —para una sola componente— bajo nuestra nomenclatura, adopta la siguiente forma:

$$Si\ t_{i1}^A \geq {}^{-1}F_{F_{n-1}^1}(1-\alpha)$$ *se acepta la hipótesis que el individuo i es atípico*

$$Si\ t_{i1}^A < {}^{-1}F_{F_{n-1}^1}(1-\alpha)$$ *se rechaza la hipótesis que el individuo i es atípico*

Esta regla de decisión es equivalente a,

$$Si\ t_{i1}^A \geq \left[{}^{-1}F_{T_{n-1}}\left(1-\frac{\alpha}{2}\right)\right]^2$$ *se acepta la hipótesis que el individuo i es atípico*

$$Si\ t_{i1}^A < \left[{}^{-1}F_{T_{n-1}}\left(1-\frac{\alpha}{2}\right)\right]^2$$ *se rechaza la hipótesis que el individuo i es atípico*

donde

$$t_{il}^{A} = \frac{n^2}{(n+1)} \frac{t_{il}^2}{\|t_l\|^2}$$

$F_{F_{n-1}^{1}}^{-1}(1-\alpha)$: *es la función inversa de la función de distribución de la variable aleatoria F de FISHER - SNEDECOR con 1 grado de libertad para el numerador y n - 1 grados de libertad para el denominador para un área de* $1-\alpha$

$F_{T_{n-1}}^{-1}\left(1-\frac{\alpha}{2}\right)$: *es la función inversa de la función de distribución de la variable aleatoria T de STUDENT - FISHER con n - 1 grados de libertad para un área de* $\left(1-\frac{\alpha}{2}\right)$

$\|t_l\|^2$: *es la norma al cuadrado de la componente* t_l

En el caso de que la muestra haya sido homogénea, procederemos a realizar el siguiente apartado. En caso contrario, eliminaremos el individuo o individuos y recomenzaremos.

EFECTUAR LA REGRESIÓN LINEAL SIMPLE DE

$$y_{(1)}^{[2]} \text{ sobre } t_2$$

Y EL TEST DE SIGNIFICACIÓN GLOBAL DE LA REGRESIÓN.

En primer lugar, buscaremos la ecuación lineal de predicción estimada y, en segundo lugar comprobaremos si, la regresión lineal simple es globalmente significativa

Ecuación lineal de predicción estimada

La ecuación lineal de predicción estimada es de la siguiente forma:

$$y_{(1)}^{[2]^*} = \hat{\beta}_{1(1)}^{[2]^*} t_1$$

donde, la estimación del coeficiente de regresión ha sido calculada de la siguiente manera:

$$\hat{\beta}_{1(1)}^{[2]} = \frac{< y_{(1)}^{[2]}, t_1 >}{\|t_1\|^2} = \frac{\sqrt{n-1}}{\|t_1\|} r_{y_{(1)}^{[2]}, t_1}$$

De ésta fórmula se deduce que, el coeficiente de regresión es igual al coeficiente de correlación simple tan sólo cuando

$$\|t_1\| = \sqrt{n-1}$$

Situación que, **tan sólo se verifica** cuando, las variables originales están normalizadas mediante el **segundo tipo.** En estos momentos estamos en condiciones de calcular el residuo asociado a la recta de regresión mediante una simple sustracción:

$$e_1$$

Test de significación global de la regresión lineal

El test de FISHER permite determinar si, la regresión lineal simple es —globalmente— significativa.

La regla general de decisión del test de FISHER —para una componente explicativa— bajo nuestra nomenclatura, adopta la siguiente forma:

$$Si\ F_{n-2}^{1*} \geq F_{F_{n-2}^{1}}^{-1}(1-\alpha)\ la\ componente\ explicativa$$

$$es\ significativa$$

$$Si\ F_{n-2}^{1*} < F_{F_{n-2}^{1}}^{-1}(1-\alpha)\ la\ componente\ explicativa$$

$$no\ es\ significativa$$

$$donde,$$

$$F_{n-2}^{1} = (n-2)\frac{r^2_{y_{(1)}^{[2]}, t_1}}{1-r^2_{y_{(1)}^{[2]}, t_1}} = (n-2)\frac{\left[< y_{(1)}^{[2]}, t_1 >\right]^2}{(n-1)\|t_1\|^2 - \left[< y_{(1)}^{[2]}, t_1 >\right]^2}$$

$$F_{F_{n-2}^{1}}^{-1}(1-\alpha): es\ la\ función\ inversa\ de\ la\ función\ de$$

distribución de la variable aleatoria F de FISHER - SNEDECOR

con 1 grado de libertad para el numerador y n - 2 grados de

libertad para el denominador para un área de 1 - α

$< y_{(1)}^{[2]}, t_1 >$: *es el producto escalar clásico de* $y_{(1)}^{[2]}$ *con* t_1

$\|t_1\|^2$: *es la norma al cuadrado de la componente* t_1

En el caso hipotético de que la componente sea significativa procederemos a deshacer los cambios

$$1^o \text{ de } t_1 \text{ a } x_1^{[2]},\ldots,x_p^{[2]}$$

$$2^o \text{ de } x_1^{[2]},\ldots,x_p^{[2]} \text{ a } x_1,\ldots,x_p$$

y, llegamos a la ecuación de predicción estimada en función de las variables explicadas originales.

La ecuación de predicción estimada, en función de la variables explicadas —normalizadas— mediante el **segundo tipo** es,

$$y_{(1)}^{[2]^*} = \sum_{j=1}^{j=p} \hat{\beta}_{1(1)}^{[2]^*} w_{1j}\ x_j^{[2]}$$

Los coeficientes de regresión de esta ecuación de predicción estimada, **son más fáciles de interpretar para el investigador.**

Si el poder explicativo de esta regresión es demasiado débil, buscamos construir una segunda componente t_2, combinación lineal de las xj, no correlacionada con t_1 y, explicando bien el residuo. Esta componente t2 es combinación lineal de los residuos e1j de las regresiones de las variables xj sobre la componente t_1. Obtenemos t_2 con ayuda de la formula

$$t_2 = w_{21}\, e_{11} + \ldots. + w_{2p}\, e_{1p}$$

donde

$$w_{2j} = \frac{cov(e_{1j}, e_1)}{\sqrt{\sum_{j=1}^{j=p} cov^2(e_{1j}, e_1)}} = \frac{< e_1, e_{1j} >}{\sqrt{\sum_{j=1}^{j=p} [< e_1, e_{1j} >]^2}}$$

Para el cálculo de los residuos

$$e_{1j}\ (j = 1,\ldots, p)$$

efectuamos las regresions simples de

$$x_j^{[2]} \text{ sobre } t_1 \ (j = 1,...,p)$$

y obtenemos las rectas de predicción estimadas:

$$x_j^{[2]^*} = \hat{\alpha}_j^{[2]^*} t_1 \ (j = 1,...,p)$$

donde las estimaciones de los coeficientes de regresión han sido calculadas de la siguiente manera:

$$\hat{\alpha}_j^{[2]} = \frac{< x_j^{[2]}, t_1 >}{\|t_1\|^2} = \frac{\sqrt{n-1}}{\|t_1\|} r_{x_j^{[2]}, t_1} \ (j = 1,...,p)$$

En estos momentos, estamos en condiciones de calcular los residuos asociados a las rectas de regresión, mediante una simple sustracción,

$$e_{1j} \ (j = 1,...,p)$$

Dado que ya conocemos

$$e_1 \text{ y } e_{1j} \ (j = 1,...,p)$$

nos resta efectuar los productos escalares clásicos

$$< e_1, e_{1j} > \ (j = 1,...,p)$$

para calcular la componente

$$t_2$$

DETECCIÓN DE INDIVIDUOS ATÍPICOS PARA LA SEGUNDA COMPONENTE

Para no resultar reiterativos este apartado se hará de la misma manera que para la primera componente.

EFECTUAR LA REGRESIÓN LINEAL SIMPLE DE

$$y_{(1)}^{[2]} \text{ sobre } t_2$$

Y EL TEST DE SIGNIFICACIÓN GLOBAL DE LA REGRESIÓN.

En primer lugar, buscaremos la ecuación lineal de predicción estimada y, en segundo lugar comprobaremos si la regresión lineal simple es —globalmente— significativa.

Ecuación lineal de predicción estimada

La ecuación lineal de predicción estimada es de la siguiente forma,

$$y_{(1)}^{[2]^*} = \hat{\beta}_{2(1)}^{[2]^*} t_2$$

donde la estimación del coeficiente de regresión ha sido calculada de la siguiente manera,

$$\hat{\beta}_{2(1)}^{[2]} = \frac{< y_{(1)}^{[2]}, t_2 >}{\|t_2\|^2} = \frac{\sqrt{n-1}}{\|t_2\|} r_{y_{(1)}^{[2]}, t_2}$$

De ésta fórmula se deduce que el coeficiente de regresión es igual al coeficiente de correlación simple de BRAVAIS-PEARSON tan sólo cuando

$$\|t_2\| = \sqrt{n-1}$$

Situación que tan sólo se verifica cuando las variables originales están normalizadas mediante el **segundo tipo**.

En estos momentos, estamos en condiciones de calcular el residuo asociado a la recta de regresión mediante una simple sustracción:

$$e_2$$

Test de significación global de la regresión

Para no resultar reiterativos este apartado se realiza de la misma manera que el que mostramos con la primera componente.

3. DETECCIÓN DE INDIVIDUOS ATÍPICOS EN EL PLANO (t1-t2)

La regla general de decisión para la detección de individuos atípicos cuando se trata de dos componentes —adaptada a nuestra nomenclatura— adopta la siguiente forma:

$$Si\ t_{i(1-2)}^{A} \geq 1\ se\ acepta\ la\ hipótesis\ que$$

$$el\ individuo\ i\ es\ atípico$$

$$Si\ t_{i(1-2)}^{A} < 1\ se\ rechaza\ la\ hipótesis\ que$$

$$el\ individuo\ i\ es\ atípico$$

donde

$$t^{A}_{i(1-2)} = \frac{t^2_{i1}}{\frac{2(n^2-1)}{n^2(n-2)}\|t_1\|^2 \; F^{-1}_{F^2_{n-2}}(1-\alpha)} + \frac{t^2_{i2}}{\frac{2(n^2-1)}{n^2(n-2)}\|t_2\|^2 \; F^{-1}_{F^2_{n-2}}(1-\alpha)}$$

donde

$\|t_1\|^2$: *es la norma al cuadrado de la componente* t_1

$\|t_2\|^2$: *es la norma al cuadrado de la componente* t_2

$F^{-1}_{F^2_{n-2}}(1-\alpha)$: *es la función inversa de la función de distribución de la variable aleatoria F de FISHER-SNEDECOR con 2 grados de libertad para el numerador y n - 2 grados de libertad para el denominador para un área de* $1-\alpha$

En el caso de que la muestra haya sido homogénea, procederemos a realizar el siguiente apartado. En caso contrario, eliminaremos el individuo o individuos y recomenzaremos.

EFECTUAR LA REGRESIÓN LINEAL MÚLTIPLE

$$y^{[2]}_{(2)} \; sobre \; t_1 \; y \; t_2$$

Y EL TEST DE SIGNIFICATIVIDAD GLOBAL DE LA REGRESIÓN.

En primer lugar,buscaremos la ecuación lineal de predicción estimada y, en segundo lugar, comprobaremos si la regresión lineal múltiple es —globalmente— significativa.

Ecuación lineal de predición estimada

La ecuación lineal de predicción estimada es de la siguiente forma:

$$y^{[2]^*}_{(2)} = \hat{\beta}^{[2]^*}_{1(2)} t_1 + \hat{\beta}^{[2]^*}_{2(2)} t_2$$

donde las estimaciones de los coeficientes de regresión han sido calculadas a partir de las siguientes formulas:

$$\hat{\beta}^{[2]}_{1(2)} = \frac{\sqrt{n-1}}{\|t_1\|}\left[\frac{r_{y^{[2]}_{(2)},t1} - r_{y^{[2]}_{(2)},t2}\, r_{t1,t2}}{1 - r^2_{t1,t2}}\right]$$

$$\hat{\beta}_{2(2)}^{[2]} = \frac{\sqrt{n-1}}{\|t_2\|}\left[\frac{r_{y_{(2)}^{[2]},t_2} - r_{y_{(2)}^{[2]},t_1} r_{t_1,t_2}}{1 - r_{t_1,t_2}^2}\right]$$

y dado que las componentes

$$t_1 \text{ y } t_2 \text{ son ortogorales}$$

$$r_{t_1,t_2} = 0$$

y, por tanto, los estos dos estimadores, en nuestro caso concreto, toman la siguiente forma:

$$\hat{\beta}_{1(2)}^{[2]} = \frac{\sqrt{n-1}}{\|t_1\|} r_{y_{(2)}^{[2]},t_1}$$

$$\hat{\beta}_{2(2)}^{[2]} = \frac{\sqrt{n-1}}{\|t_2\|} r_{y_{(2)}^{[2]},t_2}$$

De lo que observamos de este hecho el siguiente resultado:

$$\hat{\beta}_{1(2)}^{[2]} = \hat{\beta}_{1(1)}^{[2]}$$

$$\hat{\beta}_{2(2)}^{[2]} = \hat{\beta}_{2(1)}^{[2]}$$

De estas fórmulas se deduce que los **coeficientes de regresión son iguales a los coeficientes de correlación lineal de BRAVAIS-PEARSON** cuando,

$$\|t_1\| = \sqrt{n-1} \quad \|t_2\| = \sqrt{n-1}$$

Situación que, **tan sólo se verifica** cuando las variables originales, están normalizadas mediante el **segundo tipo**. En estos momentos, estamos en condiciones de calcular el residuo asociado a la línea de regresión mediante una simple sustracción:

$$e_2$$

Test de significatividad global de la regresión

El test de FISHER permite determinar si, la regresión lineal múltiple es —globalmente— significativa.

La **regla general de decisión** del test de FISHER - **para dos componentes explicativas —ortogonales—** bajo nuestra nomenclatura, adopta la siguiente forma:

$$Si\ F_{n-3}^2 \geq F_{F_{n-3}^2}^{-1}(1-\alpha)\ las\ componentes\ explicativas$$

$$t_1 \text{ y } t_2$$

$$son\ significativas$$

$$Si\ F^{2}_{n-3} < F^{-1}_{F^{2}_{n-3}}(1-\alpha)\ las\ componentes\ explicativas$$

$$t_1\ y\ t_2$$

no son significativas

donde

$$F^{2}_{n-3} = \frac{n-3}{2}\left[\frac{\sum_{j=1}^{j=2} r^{2}_{y^{[2]}_{(2)},t_j}}{1-\sum_{j=1}^{j=2} r^{2}_{y^{[2]}_{(2)},t_j}}\right]$$

o bien

$$F^{2}_{n-3} = \frac{n-3}{2}\left[\frac{\|t_2\|^2\left[< y^{[2]}_{(2)}, t_1 >\right]^2 + \|t_1\|^2\left[< y^{[2]}_{(2)}, t_2 >\right]^2}{(n-1)\|t_1\|^2\|t_2\|^2 - \left[\|t_2\|^2\left[< y^{[2]}_{(2)}, t_1 >\right]^2 + \|t_1\|^2\left[< y^{[2]}_{(2)}, t_2 >\right]^2\right]}\right]$$

$F^{-1}_{F^{2}_{n-3}}(1-\alpha)$ *: es la función inversa de la función de distribución de la variable aleatoria F de FISHER - SNEDECOR con 2 grados de libertad para el numerador y n - 3 grados de libertad para el denominador para un área de* $1-\alpha$

$r^{2}_{y^{[2]}_{(2)},t_j}$ *: coeficientes de correlación lineal de BRAVAIS - PEARSON al cuadrado entre* $y_{(2)}\ y\ t_j\ (j = 1,...,p)$

$\|t_j\|^2$ *: normas al cuadrado de las componentes* $t_j\ (j = 1,2)$

$< y^{[2]}_{(2)}, t_j >$*: productos escalares clásicos de* $y^{[2]}_{(2)}$ *con* $t_j\ (j = 1,2$

En el caso hipotético de que las componentes

$$t_1\ y\ t_2$$

sean significativas procederemos a deshacer los siguientes cambios,

$$1^{o}: de\ t_1\ y\ t_2\ \ a\ \ x^{[2]}_1,...,x^{[2]}_p$$

$$2^{o}: de\ x^{[2]}_1,...,x^{[2]}_p\ \ a\ \ x_1,...,x_p$$

y, llegamos a la ecuación de predicción estimada en función de las variables explicativas originales.

Por el contrario, si el poder explicativo de esta regresión es todavía débil, buscamos construir una tercera componente.Esta componente, es combinación lineal de los residuos

$$e_{2j}$$

de las regresiones de los residuos

$$e_{1j}\ sobre\ t_2$$

Obtenemos

$$t_3$$

con ayuda de la fórmula

$$t_3 = w_{31}e_{21} + w_{32}e_{22} + ... + w_{3p}e_{2p}$$

donde

$$w_{3j} = \frac{cov(e_{2j}, e_2)}{\sqrt{\sum_{j=1}^{j=p} cov^2(e_{2j}, e_2)}} = \frac{< e_2, e_{2j} >}{\sqrt{\sum_{j=1}^{j=p} [< e_2, e_{2j} >]^2}} \quad (j = 1,..., p)$$

Para el cálculo de los residuos

$$e_{2j}\ (j = 1,..., p)$$

efectuamos las regresiones simples de

$$e_{1j}\ sobre\ t_2\ (j = 1,..., p)$$

y obtenemos las rectas de predicción estimadas

$$\overset{*}{e}_{1j} = \overset{*}{\hat{\alpha}}_{1j}\, t_2 \quad (j = 1,..., p)$$

donde las estimaciones de los coeficientes de regresión han sido calculadas de la siguiente manera:

$$\hat{\alpha}_{1j} = \frac{< e_{1j}, t_2 >}{\|t_2\|^2} \quad (j = 1,..., p)$$

En estos momentos, estamos en condiciones de calcular los residuos asociados a las rectas de regresión mediante una simple sustracción,

$$e_{2j} \quad (j = 1,...,p)$$

Dado que ya conocemos

$$e_2 \, y \, e_{2j} \quad (j = 1,...,p)$$

nos resta efectuar los productos escalares clásicos

$$< e_2 , e_{2j} > \quad (j = 1,...,p)$$

para calcular la componente

$$t_3$$

A continuación, seguiremos los mismos pasos que, los realizados para las dos componentes anteriores.

Este procedimiento iterativo continua hasta que el número de componentes a retener sea significativo.

UNA CONSIDERACIÓN PRÁCTICA REFERENTE A LA RETENCIÓN DE COMPONENTES

En el libro de TENENHAUS(1998,p.83), se contempla —de forma clara— el método de **validación cruzada,** contenido en el paquete de programas SIMCA(1991,1996,1998) de Svante WOLD. Este método, nos indica —de manera más precisa— de la que hemos expuesto con anterioridad, el número de componentes

$$t_1 , t_2 , ..., t_H$$

a retener.

EJERCICIO DIDÁCTICO ILUSTRATIVO DEL PROCESO METODOLÓGICO HECHO CON UNA CALCULADORA.

Para el desarrollo del ejercicio, hemos retenido una de las tablas de datos ya analizada en (CERMEÑO y DÍAZ-LLANOS,2000).

Se trata de la siguiente tabla:

Tabla de datos originales

y	*xi1*	*xi2*	*xi3*
15	1	2	3
31	2	5	6
37	3	6	7
49	4	7	10
57	5	9	11

En primer lugar, procederemos a la normalización de la tabla de datos originales, mediante el segundo procedimiento, ya aludido con anterioridad.

Tabla de datos normalizados

$y_i^{[2]}$	$x_{i1}^{[2]}$	$x_{i2}^{[2]}$	$x_{i3}^{[2]}$
–1,4000646400	–1,264911064	–1,468068078	–1,370989296
–0,4175631380	–0,632455532	–0,309066963	–0,436223867
–0,0491250750	0,000000000	0,077266740	–0,124635390
0,6877510510	0,632455532	0,463600445	0,810130038
1,1790018020	1,264911064	1,236267855	1,121718515

A partir de la tabla de datos normalizados, se deduce —fácilmente— aplicando las formulas ya contempladas en el apartado correspondiente:no sólo una reflexión, en cuanto a la normalización de los datos, para la matriz de correlaciones simples de BRAVAIS-PEARSON entre las variables explicativas, sino también, para las correlaciones simples entre la variable a explicar y cada una de las explicativas.

Matriz de correlaciones

	x1	*x2*	*x3*
x1	1.0000	0.9774	0.9853
x2	—	1.0000	0.9751
x3	—	—	1.0000

Correlaciones entre la variable a explicar y las explicativas

	x1	*x2*	*x3*
y	0,9903	0,9893	0,9969

En segundo lugar, procederemos a la comprobación de la posible presunción de multicolinealidad mediante un test no estadístico o estadístico.

ERKEL-ROUSSE(1995) señala que, no es aconsejable realizar la detección de la multicolinealidad en un modelo lineal ordinario mediante un test estadístico.Esta aseveración, esta avalada por el mismo autor (1994/1995) y por MADDALA(1977).

Nosotros aplicaremos, en esta ocasión, el indice de multicolinealidad contemplado en (BELSLEY,KUH,WELSH,1980; TOMASSONE y coll,1992,pp.149-150; FOUCART, 1992). El **indice de multicolinealidad** adopta la siguiente forma,

$$F = \frac{1}{p} \sum_{j=1}^{j=p} \frac{1}{\lambda_j}$$

$$1 \leq F \leq +\infty$$

donde

λ_j *son los valores propios de la matriz de*

correlaciones entre las p variables explicativas

Los valores propios de la matriz de correlaciones son los siguientes:

$$\lambda_1 = 2,9585 \quad \lambda_2 = 0,0270 \quad \lambda_3 = 0.0145$$

Aplicando la fórmula del indice de multicolinealidad obtenemos que

$$F = 35,4469$$

Recordemos que en (CERMEÑO y DIAZ-LLANOS,2000) se contrastó la multicolinealidad mediante el test de FARRAR y GLAUBER(1967) con un nivel de significación del 0,01, y la sospecha de presunción de multicolinealidad ya fue corroborada.

MONTGOMERY y RUGER (1996) proponen una regla rápida para detectar la multicolinealidad. Dicha regla consiste en:

$$Si \ \frac{\lambda_1}{\lambda_p} > 100 \ existe \ presunción \ de \ multicolinealidad$$

donde,

λ_1 *: es el primer valor propio de la matriz*

de correlaciones entre las variables explicativas

λ_p *: es el último valor propio de la matriz*

de correlaciones entre las variables explicativas

En nuestro caso concreto,

$$\frac{\lambda_1}{\lambda_3} = 204,0345$$

Por consiguiente, existe presunción de multicolinealidad.

Sin embargo, si se aplica el test de KLEIN(1962) que no esta sujeto a hipótesis distribucionales nos da que no hay sospecha de presunción de multicolinealidad.

Criterio del test de KLEIN

$$Si \ R^2_{y,x_1x_2....x_p} < r_{x_i,x_j} \quad hay \ presunción \ de \ multicolinearidad$$

$$Si \ R^2_{y,x_1x_2....x_p} \geq r_{x_i,x_j} \quad no \ hay \ presunción \ de \ multicolinealidad$$

$$(i \neq j)$$

Por tanto, tomamos la decisión de aplicar la regresión PLS1

PRIMERA ETAPA

CONSTRUCCIÓN DE LA COMPONENTE t1

Para la construcción de la componente t_1 vamos a seguir el siguiente proceso

1:**Cálculo de las estimaciones de los productos escalares clásicos entre**

$$y^{[2]} \text{ y } x_j^{[2]} \ (j = 1,2,3)$$

$$< y^{[2]}, x_1^{[2]} > = 3,961351751$$
$$< y^{[2]}, x_2^{[2]} > = 3,957053177$$
$$< y^{[2]}, x_3^{[2]} > = 3,987423300$$

2:**Cálculo de las estimaciones de los coeficientes**

$$w_{1j} \ (j = 1,2,3)$$

$$w_{11}^{*} = 0,5763 \quad w_{12}^{*} = 0,5757 \quad w_{13}^{*} = 0,5801$$

3: **Construcción de la estimación de la componente t1**

$$t_1^{*} = 0,5763\, x_1^{[2]} + 0,5757\, x_2^{[2]} + 0,5801\, x_3^{[2]}$$

De lo que se deduce que,

$$t_1^{*} = \begin{pmatrix} -2,369366884 \\ -0,795436602 \\ -0,027820325 \\ 1,101301703 \\ 2,091322107 \end{pmatrix}$$

SEGUNDA ETAPA

DETECCIÓN DE INDIVIDUOS ATÍPICOS EN LA PRIMERA COMPONENTE

Como resultado de la aplicación de la regla general de decisión, para la detección de individuos atípicos, cuando se ha retenido una sola componente, obtenemos la siguiente tabla:

Individuos	t_{i1}^{A}	*UCD*	*Diagnóstico de individuos atípicos*
1	1,976632770	7,7086	Negativo
2	0,222778104	7,7086	Negativo
3	0,000272512	7,7086	Negativo
4	0,427045337	7,7086	Negativo
5	1,539937941	7,7086	Negativo

donde

$$UCD = F^{-1}_{F_4^1}(0,95) = 7,7086$$

De los resultados de esta tabla se concluye que, la muestra estudiada puede ser considerada como homogénea.

TERCERA ETAPA

EFECTUAR LA REGRESIÓN LINEAL SIMPLE DE

$$y_{(1)}^{[2]} \text{ sobre } t_1$$

Y EL TEST DE SIGNIFICACIÓN GLOBAL DE LA REGRESIÓN

$$y_{(1)}^{[2]^*} = 0,5809\, t_1$$

$$r^{*2}_{y_{(1)}^{[2]}, t_1} = 0.9982$$

$$F_3^{1^*} = 1659,3650$$

$$\text{Como } F_3^{1^*} = 1659,3650 \text{ es mayor que } F^{-1}_{F_3^1}(0,95) = 10.128$$

la componente t_1 es significativa

Dado que la primera componente es significativa procedemos a deshacer el primer cambio, dando el siguiente resultado:

$$y^{[2]^*} = 0,3347\, x_1^{[2]} + 0,3344\, x_2^{[2]} + 0,3370\, x_3^{[2]}$$

El segundo cambio no lo haremos dado que, al final mostraremos las ecuaciones con las cuatro estrategias que hemos realizado, estando la ecuación bajo esta situación.

FE DE ERRATAS

En la página 189 dice:

A partir de estos momentos, estamos en condiciones de calcular los residuos asociados a las rectas de regresión, mediante una simple sustracción,

$$e_{11} = \begin{pmatrix} 0,105056868 \\ -0,172534089 \\ 0,016085711 \\ -0,004317112 \\ 0,055708622 \end{pmatrix} \quad e_{12} = \begin{pmatrix} -0,103075816 \\ 0,149184063 \\ 0,093294029 \\ -0,170859466 \\ 0,031457189 \end{pmatrix} \quad e_{13} = \begin{pmatrix} -0,002206047 \\ 0,023299858 \\ -0,108563589 \\ 0,173908045 \\ -0,086438266 \end{pmatrix}$$

Debe decir:

A partir de estos momentos, estamos en condiciones de calcular los residuos asociados a las rectas de regresión, mediante una simple sustracción,

$$e_{11} = \begin{pmatrix} 0,105056868 \\ -0,172534089 \\ 0,016085711 \\ -0,004317112 \\ 0,055708622 \end{pmatrix} \quad e_{12} = \begin{pmatrix} -0,103075816 \\ 0,149184063 \\ 0,093294029 \\ -0,170859466 \\ 0,031457189 \end{pmatrix} \quad e_{13} = \begin{pmatrix} -0,002206047 \\ 0,023299858 \\ -0,108563589 \\ 0,173908045 \\ -0,086438266 \end{pmatrix}$$

Aunque en nuestro caso, no es necesario mejorar ligeramente la regresión, buscando la segunda componente t_2, en este caso, la buscaremos —simplemente— para mostrar el proceso metodológico.

CUARTA ETAPA

CONSTRUCCIÓN DE LA COMPONENTE t_2

Para la construcción de la componente t_2 vamos a seguir el siguiente proceso:

1: **Cálculo de los residuos**

$$e_{1j} \ (j = 1,2,3)$$

Para el cálculo del residuos

$$e_{1j} \ (j = 1,2,3)$$

efectuamos las regresiones simples de,

$$x_j^{[2]} \ sobre \ t_1 \ (j = 1,2,3)$$

y obtenemos las rectas de predicción estimadas,

$$x_1^{[2]^*} = 0,5782\, t_1 \qquad x_2^{[2]^*} = 0,5761\, t_1 \qquad x_3^{[2]^*} = 0,5777\, t_1$$

A partir de estos momentos, estamos en condiciones de calcular los residuos asociados a las rectas de regresión, mediante una simple sustracción,

$$e_{11} = \begin{pmatrix} 0,105056868 \\ -0,172534089 \\ 0,016085711 \\ -0,004317112 \\ 0,055708622 \end{pmatrix} \quad e_{12} = \begin{pmatrix} -0,109075816 \\ 0,149184063 \\ 0,093294029 \\ -0,170859466 \\ 0,031457189 \end{pmatrix} \quad e_{13} = \begin{pmatrix} -0,002206047 \\ 0,023299858 \\ -0,108563589 \\ 0,173908045 \\ -0,086438266 \end{pmatrix}$$

2: **Cálculo de las estimaciones de los productos escalares clasicos entre**

$$e_1 \ y \ e_{1j} \ (j = 1,2,3)$$

$$\langle e_1, e_{11} \rangle = -0,012903080$$

$$\langle e_1, e_{12} \rangle = -0,003322689$$

$$\langle e_1, e_{13} \rangle = 0,016114989$$

3: **Cálculo de las estimaciones de los coeficientes**

$$w_{2j} \ (j = 1,2,3)$$

$$w^{*}_{21} = -0,6171 \quad w^{*}_{22} = -0,1593 \quad w^{*}_{23} = 0,7711$$

4: **Construcción de la estimación de la componente t2**

$$t^{*}_{2} = -0,6171 e_{11} - 0,1593 e_{12} + 0,7711 e_{13}$$

De lo que se deduce que,

$$t^{*}_{2} = \begin{pmatrix} -0,082993676 \\ 0,148271341 \\ -0,078784571 \\ 0,109548397 \\ -0,09604149 \end{pmatrix}$$

QUINTA ETAPA

DETECCIÓN DE INDIVIDUOS ATÍPICOS EN LA COMPONENTE t2

Como resultado de la aplicación de la regla general de decisión, para la detección de individuos atípicos, cuando se ha retenido una sola componente, obtenemos la siguiente tabla:

Individuos	t^{A}_{i2}	*UCD*	*Diagnóstico de individuos atípicos*
1	0,509727676	7,7086	Negativo
2	1,626906686	7,7086	Negativo
3	0,459336082	7,7086	Negativo
4	0,888096721	7,7086	Negativo
5	0,682599539	7,7086	Negativo

donde

$$UCD = F^{-1}_{F'_4}(0,95) = 7,7086$$

De los resultados de esta tabla se concluye que, la muestra estudiada puede ser considerada como homogénea.

SEXTA ETAPA

EFECTUAR LA REGRESIÓN LINEAL SIMPLE DE,

$$y_{(1)}^{[2]} \; sobre \; t_2$$

Y EL TEST DE SIGNIFICACIÓN GLOBAL DE LA REGRESIÓN

$$y_{(1)}^{[2]^*} = 0,3599\, t_2$$

$$r^{*}_{y_{(1)}^{[2]},t_2} = 0,04270$$

$$F_3^{l^*} = 0,005479$$

$$Como \; F_3^{l^*} = 0,005479 \; es \; menor \; que \; F^{-1}_{F_3^l}(0,95) = 10,128$$

$$la \; componente \; t_2 \; no \; es \; significativa$$

Como era de esperar, la segunda componente no es significativa y, el proceso se para.

OBSERVACIONES REFERENTES A LA LÍNEA DE PREDICCIÓN ESTIMADA POR CUATRO ESTRATEGIAS METODOLÓGICAS.

1. Si aplicamos el método de **regresión lineal múltiple** (TOMASSONE,1989;1991), considerando que, todas las variables están normalizadas, por el **segundo tipo** ya aludido con anterioridad, obtenemos la ecuación de predicción lineal que a continuación mostramos:

$$y^{[2]^*} = 0,0946\, x_1^{[2]} + 0,3183\, x_2^{[2]} + 0,5934\, x_3^{[2]}$$

2. Si aplicamos el método de **regresión progresiva** (TOMASSONE,1989), considerando que, todas las variables están normalizadas, por el **segundo tipo**, vamos obteniendo las ecuaciones de predicción lineal que a continuación mostramos:

$$y^{[2]^*} = 0,9969\, x_3^{[2]}$$

$$Desv\text{-}est.residual = 0,0915$$

$$r^{*}_{y,x_3} = 0,9969$$

$$r^{2*}_{y,x_3} = 0,9937$$

$$F_3^{l^*} = 474,829$$

$$y^{[2]^*} = 0,3506\, x_2^{[2]} + 0,6550\, x_3^{[2}$$

$$Desv\text{-}est.residual = 0,0211$$

$$R^{*}_{y,x_2x_3} = 0,9998$$

$$R^{2*}_{y,x_2x_3} = 0,9999$$

$$F_2^{*} = 4508,4907$$

$$y^{[2]^{*}} = 0,0946\, x_1^{[2]} + 0,3183\, x_2^{[2]} + 0,5934\, x_3^{[2]}$$

3. Si aplicamos el método de **regresión por etapas** (DRAPER y SMITH, 1981; BOURBONNAIS y USUNIER, 1992; BOURBONNAIS, 1998; CERMEÑO y DÍAZ-LLANOS, 2000), considerando que, todas las variables están normalizadas por el **segundo tipo**, obtenemos la ecuación de predicción lineal que a continuación mostramos,

$$y^{[2]^{*}} = 0,9969\, x_3^{[2]}$$

4. Si aplicamos el método de **mínimos cuadrados parciales** (WOLD, 1985; MARTENS y NAES, 1989; BRY, 1996; TENENHAUS, 1998), obtenemos la ecuación de predicción lineal que a continuación mostramos,

$$y^{[2]^{*}} = 0,3347\, x_1^{[2]} + 0,3344\, x_2^{[2]} + 0,3370\, x_3^{[2]}$$

Nota:los resultados de los apartados 2 y 3 han sido obtenidos mediante el programa de regresión lineal múltiple contenido en el paquete de programas STATITCF 4.0 (1991).

La última ecuación del apartado 2, es la misma que la del apartado 1.Ha sido obtenida, mediante una calculadora ya que, el programa de regresión lineal múltiple no la contempla dado que, el determinante de la matriz que hay que invertir, es muy próximo a cero.

CONCLUSIONES

— Aunque el ajuste de una nube de puntos a una línea y la detección de datos atípicos datan del siglo XVIII, aún a principios del siglo XXI, se sigue —y se seguirá— investigando, sobre estos temas.

— En nuestro caso concreto, hemos de indicar que, la regresión obtenida con la cuarta estrategia (**regresión PLS1**), es tan buena como la de la primera estrategia (**regresión lineal múltiple**) pero, la de la cuarta (**regresión PLS1**) tiene la ventaja de, ser perfectamente comprensible para el investigador.

— El hecho de que la ecuación de predicción lineal obtenida mediante la **regresión PLS1,** es más comprensible para el investigador que, la obtenida mediante la **regresión lineal múltiple,** se pone de manifiesto en las aplicaciones del método a datos concretos (TENENHAUS,GAUCHI y MENARDO, 1995, pp.29,43; TENENHAUS, 1998, p.81).

Aclaraciones más importantes de la nomenclatura utilizada

$y^{[1]}$: *matriz de la variable a explicar normalizada de la siguiente manera*

$$y_i^{[1]} = \frac{y_i - \bar{y}}{\sqrt{SCD_y}}$$

$y^{[2]}$: *matriz de la variable a explicar normalizada de la siguiente manera*

$$y_i^{[2]} = \frac{y_i - \bar{y}}{\sqrt{\frac{SCD_y}{n-1}}}$$

$X^{[1]}$: *matriz de las variables explicativas normalizadas de la siguiente manera*

$$x_{ij}^{[1]} = \frac{x_{ij} - \bar{x}_j}{\sqrt{SCD_{x_j}}}$$

$X^{[2]}$: *matriz de las variables explicativas normalizadas de la siguiente manera*

$$x_{ij}^{[2]} = \frac{x_{ij} - \bar{x}_j}{\sqrt{\frac{SCD_{x_j}}{n-1}}}$$

R : *matriz de coerrelaciones entre las variables explicativas*

r_{y,x_j} : *coeficiente de correlacción lineal simple de BRAVAIS - PEARSON entre y* y x_j $(j = 1,..., p)$

r^*_{y,x_j} : *estimación del coeficiente de correlacción lineal simple de BRAVAIS - PEARSON entre y* y x_j $(j = 1,..., p)$

$R_{y,x_1x_2....x_p}$: *coeficiente de correlacción lineal múltiple entre y* y $x_1x_2.....x_p$

$R^*_{y,x_1x_2....x_p}$: *estimación del coeficiente de correlacción lineal múltiple entre y* y $x_1x_2....x_p$

$R^2_{y,x_1x_2....x_p}$: *coeficiente de determinación*

$R^{2*}_{y,x_1x_2....x_p}$: *estimación del coeficiente de determinación*

β_j : *coeficientes de regresión lineal de*

y sobre x_j $(j = 1,..., p)$

$\beta_j^{[1]}$: *coeficientes de regresión lineal de*

$y^{[1]}$ *sobre* $x_j^{[1]}$ $(j = 1,..., p)$

$\beta_j^{[2]}$: *coeficientes de regresión lineal de*

$y^{[2]}$ *sobre* $x_j^{[2]}$ $(j = 1,..., p)$

$\hat{\beta}_j^{[1]}$: *estimadores de los coeficientes de regresión lineal*

de $y^{[1]}$ *sobre* $x_j^{[1]}$ $(j = 1,..., p)$

$\hat{\beta}_j^{[2]}$: *estimadores de los coeficientes de regresión lineal*

de $y^{[2]}$ *sobre* $x_j^{[2]}$ $(j = 1,..., p)$

$\hat{\beta}_j^{[1]^*}$: *estimaciones de los coeficientes de regresión lineal*

de $y^{[1]}$ *sobre* $x_j^{[1]}$ $(j = 1,..., p)$

$\hat{\beta}_j^{[2]^*}$: *estimaciones de los coeficientes de regresión lineal*

de $y^{[2]}$ *sobre* $x_j^{[2]}$ $(j = 1,..., p)$

$\hat{\beta}_{1(1)}^{[2]^*}$: *estimación del coeficiente de regresión de*

$y_{(1)}^{[2]}$ *sobre* t_1

$\hat{\beta}_{2(1)}^{[2]^*}$: *estimación del coeficiente de regresión de*

$y_{(1)}^{[2]}$ *sobre* t_2

$\hat{\beta}_{1(2)}^{[2]^*}$: *estimación delprimer coeficiente de regresión de*

$y_{(2)}^{[2]}$ *sobre* t_1 *y* t_2

$\hat{\beta}_{2(2)}^{[2]^*}$: *estimación del segundo coeficiente de regresión de*

$y_{(2)}^{[2]}$ *sobre* t_1 *y* t_2

$\hat{\beta}^{[2]^*}_{1(3)}$: *estimación del primer coeficiente de regresión de*

$y^{[2]}_{(3)}$ *sobre* t_1, t_2 *y* t_3

$\hat{\beta}^{[2]^*}_{2(3)}$: *estimación del segundo coeficiente de regresión de*

$y^{[2]}_{(3)}$ *sobre* t_1, t_2 *y* t_3

$\hat{\beta}^{[2]^*}_{3(3)}$: *estimación del tercer coeficiente de regresión de*

$y^{[2]}_{(3)}$ *sobre* t_1, t_2 *y* t_3

$\hat{\alpha}^{[2]^*}_j$: *estimaciones de los coeficientes de regresión*

de $x^{[2]}_j$ *sobre* $t_1 (j = 1,..., p)$

$\hat{\alpha}^*_{1j}$: *estimaciones de los coeficientes de regresión*

de e_{1j} *sobre* $t_2 (j = 1,..., p)$

$\hat{\alpha}^*_{2j}$: *estimaciones de los coeficientes de regresión*

de e_{2j} *sobre* $t_3 (j = 1,..., p$

e_1 : *residuo de la regresión de* $y^{[2]}_{(1)}$ *sobre* t_1

e_{1j} : *residuos de las regresiones de* $x^{[2]}_j$ *con* $t_1 (j = 1,..., p)$

e_2 : *residuo de* $y^{[2]}_{(2)}$ *con* t_1 *y* t_2

e_{2j} : *residuos de las regresiones de* e_{1j} *con* t_2 $(j = 1,..., p)$

BIBLIOGRAFÍA

ADRAIN R.(1808):Research concerning the probabilities of the errors which happen in making observations.Analyst,vol 1,p.93-109.

AIRY G.B.(1856):Letter from Professor Airy, Astronomer Royal, to the editor. Astronomical Journal,4, pp.137-138.

ALBANO C.,DUNN III W.J.,ESBENSEN K., HELLBERG S.,JOHANSSON E., SJOSTROM H., WOLD S.(1983):Pattern Recognition: Finding and using regularities in Multivariate Data in Proc.IFOST Cont.» Food Research and Data Analysis», MARTENS J(Ed), Applied Science Publications, London.

AUDRAIN S.,LESQUOY-de TURCKHEIM., MILLEIR C.,TOMASSONE R.(1992):La régression, nouveaux regards sur une ancienne méthode statistique. INRA et MASSON. Paris.

BELSLEY D.A.,KUH E.,WELSH R.E.(1980): Regression diagnostics: identifying influential data and sources of collinearity. Willey,New York.

BERTRAND J.(1855):Méthode des moindres carrés.Mémoire sur la combinaison des

observations.Traduction française de l'oeuvre de C.F.GAUSS par J BERTRAND (autorisé par C.F GAUSS lui-même).Mallet-Bachelin.Paris

BOOKSTEIN F.L.(1982):The geometric meaning of soft modeling, with some generalizations, in system under indirect observation, vol.2, K.G. JÖRESKOG & H. WOLD (Eds), North-Holland, Amsterdam, pp.55-74

BOURBONNAIS R. (1998): Econometrie. 2ème édition. Dunod.

BOURBONNAIS R., USUNIER J-CL. (1992): Pratique de la prévision des ventes. Conception de systèmes. Economica.

BRY X. (1996): Analyses factorielles multiples. Ed Economica.

CAZES P. (1996): Méthodes de Régression, polycopié de 3ème cycle. Université Paris IX Dauphine,Paris.

CAZES P. (1997): Adaptation de la régression PLS au cas de la régression après Analyse des Correspondances Multiples. Revue de Statistique Appliquée, vol.45, n12, pp.89-99.

CERMEÑO CARRASCO C.,DÍAZ-LLANOS y SAINZ-CALLEJA Fco (2000): Efecto de la eliminación progresiva de individuos atípicos en la regresión por etapas. Anales de la Real Academia de Doctores. Volumen 4, pp.267-297.

DRAPER N., SMITH H. (1981): Applied Regression Analysis. Second Edition. Jonh Wiley & Sons, Inc.

ERKEL-ROUSSE H. (1994/1995): Multicolinéarité dans le modèle linéaire ordinaire: définition, détection, propositions de solutions, in «Introduction à l'économétrie du modèle linéaire», polycopié ENSAE, pp.177-252.

ERKEL-ROUSSE H. (1995): Détection de la multicolinéarité dans un modèle linéaire ordinaire:quelques éléments pour un usage averti des indicateurs de BELSLEY, KUH et WELSCH. Revue de Statistique Appliquée,XLIII(4), 19-42.

EULER L.(1749): Recherches sur la question des inégalités du mouvement de Saturne et de Jupiter, pièce ayant remporté le prix de l'année 1748, par l'Académie royale des sciences de Paris.Republié en 1960, dans Leonhardi Euleri, Opera Omnia, 2 ème série,25, pp.47-157. Turici,Bâle.

FARRAR D.E.,GLAUBER R.R. (1967): Multicolinearity in regression analysis. Review of Economics and Statistics,vol.49.

FOUCART T. (1992): Colinéarité dans une matrice de produit scalaire. Revue de Statistique Appliquée, XXXX (3), 5-17.

FOUCART T. (1996): Analyse de la colinéarité. Classification de variables. Revue de Statistique Appliquée, XLIV(4), 41-57.

GAUSS C.F.(1806): «II Comet vom Jahr 1805» Monatliche Correspondenz zur Beförderung der Erd-un Himmelskunde, vol.14, p.181-186.

GAUCHI J-P.,MENARDO C.,TENENHAUS M.(1995): Régression PLS et applications. Revue de Statistique Appliquée, XLIII (1), 7-63.

HÖSKULDSSON A.(1988):PLS regression methods. Journal of Chemometrics, vol. 2, 211-228.

IEMMA A.F.,PALM R. (1995): Quelques alternatives à la régression classique dans le cas de la colinéarité. Revue de Statistique Appliquée, 43,(2), p 5-33.

JAMBU M.(1999): Méthodes de base de l'analyse des données. Eyrolles.

JOHNSTON J (1975): Métodos de econometría. Tercera edición. Editorial Vicens-Vives, 464 p.

KLEIN L.R. (1962): An introduction to econometrics. Prentice Hall.

LEGENDRE A-M.(1805):Nouvelles méthodes pour la détermination des orbites des Comètes.Courcier.Paris.

LEGENDRE L., LEGENDRE P. (1998): Numerical Ecology. Second english edition. Elsevier, pp. 132-139.

LOHMÖLLER J.B.(1989): Latent variables Path Modeling with Partial Least Squares, Physica-Verlag, Heildelberg.

LYTTKENS, E., WOLD. H. (1969). Nonlineal iterative partial least squares (NIPALS) estimation procedures. Bull.Intern. Statist. Inst: Proc, 37 th session, London,1-15.

MADDALA G.S. (1977): Econometrics, McGraw-Hill Ed.

MARTENS H.,NAES T. (1989): Multivariate calibration. New York, Wiley, 419 p.

MASSON R.L., TRACY N.D., YOUNG J.C.(1992): Multivariate Control Charts for individual observations. Journal of Quality Technology, vol. 24, pp. 88-95.

MAYER J.T. (1750): Abhandlung über die Umwalzung des Monds um seine Axe and die scheinbare Bewegung der Mondsflecken. Kosmographische Nachrichten und Sammlungen auf das Jahr 1748, pp. 52-183.

MONTGOMERY D.C., RUNGER G.C. (1996): Probabilidad y Estadística aplicadas a la Ingeniería. McGRAW-HILL INTERAMERICANA EDITORES.

NORDBERGERG L. (1982): A procedure of determination of a good ridge parameter in linear regression. Commun Statist-Simula.Comouta., 11 (3), 285-289.

PAGÈS J., TENENHAUS M. (1999): Analyse Factorielle Multiple et approche PLS, Actes du Symposium PLS'99, Groupe HEC/CISIA- CERESTA, 5 et 6 octobre 1999, Jouy-en-Josas

PEIRCE B. (1852): Criterion for the rejection of doubtful observations. Astronomical Journal, 2, pp. 161-163.

SIMCA.(1991): «Soft Independant Modeling of Class Analogy» Version 4.3R, Umetri AB Box 1456, S-90124 Umea.

SIMCA-P for Windows. (1996): Graphical Software for Multivariate Process Modeling. Umetri AB, Box 7960, S-90719 Umea, Sweden.

SIMCA 7.0.(1998): Graphical Software for Multivariate Modeling. Umetri AB,Box 7960,S-90719 Umea, Sweden.

STATITCF 4.0.(1991): Institut Technique de Céréales et des Fourrages. 8 avenue du Président Wilson 75116, Paris.

STIGLER S.M.(1973):Simon Newcomb,Percy Daniell, and the history of robust estimation 1885-1920.Journal of the American Statistical Association,vol.68,number 344,pp.872- 879.

TENENHAUS M. (1995): A partial least squares approach to multiple regression, redundancy analysis, and canonical analysis. Les Cahiers de Recherches de HEC, CR 550.

TENENHAUS M. (1998): La régression PLS. Théorie et pratique. Editions Technip, 254 p.

TOMASSONE R. (1989): Comment interpréter les résultats d'une régression lineal. Institut Technique des Céréales et des Fourrages.

TROTTER H.F. (1957): GAUSS'work (1803-1826). On the theory of Least Squares. Traduction anglaise par H.F Trotter. Statistical Techniques Research Group, Technical Report, n1 5. Princeton, N.J Princenton University.

TZE-SAN LEE. (1998): Optimum ridge parameter selection. Applied Statistics, 36 (1), 112-118.

WOLD H. (1985): Partial Least Squares, in Encyclopedia of Statistical Sciences, vol. 6, KOTZ. S. & JOHNSON N.L. (Eds), John Wiley & Sons, New York, pp.581-591