CUBO, Vol. 3. Octubre 1987
Págs. 194-237

ANALISIS MULTIDIMENSIONAL DE DATOS

Raul Gouet *

COMPONENTES PRINCIPALES REGRESION LINEAL.-

El propósito de este cursillo es presentar dos métodos clásicos del análisis estadístico de datos multivariados, intentando un equilibrio razonable entre teoría y aplicaciones. La teoría, que espero despierte interés en aquellos con vocación matemática, usa herramientas clásicas de Algebra Lineal y Probabilidades, contenidas normalmente en los programas de carreras científico-tecnológicas.

Completando la teoría, veremos también elementos de interpretación de resultados que luego aplicaremos a algunos conjuntos de datos en el microcomputador. Todos están invitados a cocinar las recetas teóricas, incluso los matemáticos.

* Facultad de Ciencias Físicas y Matemáticas, Universidad de Chile.

1.- EL ANALISIS DE DATOS.-

En estudios de diversa naturaleza y origen, se presenta generalmente una etapa de mediciones empíricas que arrojan volúmenes importantes de datos. De ellos esperamos una mejor comprensión del fenómeno en estudio.

El AD se ocupa del análisis estadístico de datos que surgen en muestras multivariadas. Su propósito primario es reducir grandes masas de números, produciendo una síntesis significativa y razonablemente completa de toda información que contenga la muestra. En este proceso de resumen de los datos, aceptamos perder información para ganar significación.

En general, la práctica del AD es descriptiva, en el sentido de rechazar la imposición de un modelo probabilístico para los datos, por ejemplo normalidad. Esta posición, de aparente objetividad, impide estudios de estabilidad de resultados, a través de procedimientos clásicos de inferencia estadística.

Hay una gran variedad de técnicas y recetas que suelen llamarse de AD. Por una parte están los métodos lineales, que tienen su origen en la escuela anglo sajona durante la primera mitad de este siglo, y que adquieren verdadera popularidad con el uso masivo de computadores.

La escuela francesa, en los años 60 y 70, remoza y formaliza elegantemente estas técnicas, llamadas también factoriales, revelando por ejemplo la potencia del Análisis de Correspondencias.

La otra cara de la medalla reúne una extensa fauna de métodos exóticos, ni lineales ni convencionales, sobre los cuales

hay pocos resultados teóricos pero abundante controversia.

En este curso abordaremos, con la perspectiva de la escuela francesa, los métodos de Análisis en Componentes principales y Regresión Lineal.

1.1.- POBLACION Y MUESTRA.-

Antes de detallar el formalismo matemático, vamos a precisar el sentido de algunos términos de uso corriente en estadística: Se entenderá que una población es una colección de objetos que pertenecen a una clase bien definida e identificable. Una muestra será cualquier subconjunto de la población.

La distinción entre muestra y población cobra interés, cuando buscamos hacer inferencias que generalicen las conclusiones válidas en la muestra hacia la población completa. Ello obliga en principio a diseñar cuidadosos esquemas de muestreo, pero la realidad suele apartarse de los buenos deseos del estadístico, y de la muestra a menudo no se sabe de donde viene ni como llegó. En ciertas disciplinas, como la Arqueología, el investigador debe conformarse con lo escasamente disponible.

Aunque habitualmente el objetivo final de un estudio es el conocimiento de la población, nuestro análisis estará centrado en la muestra, la que exploraremos desde diversos ángulos .

1.2.- VARIABLES E INDIVIDUOS.-

A los términos variable e individuo asignaremos un significado preciso. Un individuo es cualquier objeto o entidad que forme parte de la muestra. Una variable (también llamada carácter) es un atributo cuantitativo que poseen los individuos. Podemos imaginarla como una regla o aplicación que asocia a cada individuo un número real.

En un estudio se medirán típicamente muchas variables, sobre uno o más conjuntos de individuos. Estos números, dispuestos en arreglos rectangulares (matrices), llamadas tablas de datos, constituyen el punto de partida para el AD.

1.3.- COMPONENTES PRINCIPALES (ACP).-

Desarrollado por Hotelling en los años 30, el ACP es un típico representante de los métodos factoriales, que tiene actualmente gran popularidad, aunque se reconoce que no es ni el único ni el mejor método de análisis factorial disponible.

Desde el punto de vista de la representación de los individuos en un espacio vectorial, el objetivo del ACP es simplificar, reduciendo la dimensión del espacio, con la menor pérdida de información posible. Desde el punto de vista de las variables, el ACP pretende concentrar la información en una base de variables no correlacionadas, que tengan máxima importancia.

1.4.- REGRESION LINEAL (RL).-

Los métodos de regresión son una panacea estadística. Todo el mundo tiene a su alcance, ya sea en calculadora o computador un programa que hace regresiones. Se usa la regresión con múltiples propósitos y frecuentemente se abusa.

Nadie quiere dejar de ajustar un modelo si tiene la oportunidad: para predecir, extrapolar, aproximar, explicar, descubrir una ley, verificar una hipótesis, etc.

Es la opinión de muchos que la RL es probablemente una de las técnicas más potentes para analizar datos, y se abusa de ella porque parece decirnos mucho más de lo que los datos son capaces de dar.

Los métodos de regresión se ocupan del problema general de describir relaciones entre variables (estadísticas), "explicando" una de ellas en términos de las restantes.

La RL se practica con variados objetivos: remover efectos de variables; descubrir una ley empírica; predecir comportamiento; controlar un sistema etc.

El tratamiento sistemático de los problemas relacionados con la regresión necesita dos o más cursos de un semestre. Aquí estaremos satisfechos mirando algunas "componentes principales" que sirvan como estímulo en la investigación personal posterior.

2.- DEFINICIONES BASICAS.-

Supongamos que la muestra consiste en n individuos $\iota = 1,2,\ldots,n$ sobre los cuales se han medido p variables cuantitativas $\upsilon = 1,2,\ldots,p$.

Cada variable v en υ puede verse como una aplicación

$$X^v : \mathfrak{l} \longrightarrow \mathbb{R}$$

$$i \longrightarrow X^v(i) = X^v_i$$

Siendo X^v_i el valor que toma la variable v sobre el individuo i.

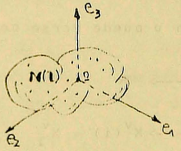
El arreglo rectangular de p filas y n columnas conteniendo la información relativa a las p variables sobre todos los individuos, lo llamaremos tabla de datos

$$X = \begin{array}{|c|c|} \hline \multicolumn{2}{|c|}{\longleftarrow i \longrightarrow} \\ \hline & \\ X^v_i \ldots & \\ & \\ \hline \end{array} \quad \Big\} \; v$$

Un individuo i en ι queda (para efectos del análisis) completamente caracterizado por los valores $X^1_i, X^2_i, \ldots, X^p_i$. Esto permite representarlo en el espacio $E = R^p$ como el vector X^i cuyas coordenadas en la base canónica son los números X^v_i.

Es natural que el espacio E sea llamado Espacio de individuos.

La colección de vectores de E que representa a los individuos de la muestra se llama Nube de individuos: $N(\iota)=\{ X_i ; i \in \iota \}$



Tenemos una primera representación de los datos en un espacio vectorial, desde la perspectiva de los individuos.

En ciertos problemas es útil (y necesario) asignar a cada individuo un número no negativo que refleje su importancia relativa. Esto se logra con pesos P_i, normalizados de manera que $0 \leq P_i \leq 1$ y $\Sigma_i P_i = 1$. Cuando no hay forma de justificar un trato discriminatorio se escogen valores $P_i = 1/n$.

Consideremos ahora una variable $v \in \iota$. Todo lo que sabemos de ella está en la colección de valores que toma en la muestra ι. Es entonces razonable asociarle el vector X^v en $F = R^n$, cuyas coordenadas en la base canónica $\{f_j ; i \in \iota\}$ son los datos $x^v{}_1, x^v{}_2, \ldots, x^v{}_n$.

El espacio F se conoce como Espacio de variables y la colección $N(\upsilon) = \{ X^v ; v \in \upsilon \}$ es la nube de variables.

2.1.- REPRESENTACION EN ESPACIOS DUALES.-

Sean E^*, F^* los espacios duales de E y F respectivamente. Para designar la imagen u(x) de un vector X en E (Resp.F) por la forma lineal u en E^* (resp. F^*) escribiremos $\langle u,x\rangle$ O $\langle x,u\rangle$, identificando E con el doble dual E^{**}.

Supongamos que E^* y F^* están dotados de las bases duales $\{e^*_v ; v \in u\}$ y $\{f^*_i ; i \in \iota\}$. Estas bases se caracterizan porque asignan a un vector sus coordenadas en las bases canónicas. Por ejemplo $\langle e^*_v, X_i \rangle = X^v_i$ y $\langle F^*_j, X^v \rangle = X^v_i$.

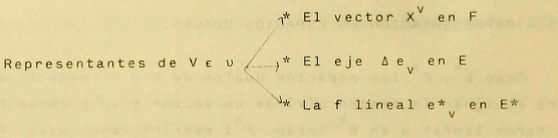
Lo anterior revela que e^*_v representa efectivamente a la variable v puesto que asocia al individuo i, a través de su vector X_i, el valor $X^v_i = X^v(i)$. Con idéntico argumento podemos convencernos que f^*_i representa al individuo i.

Finalmente, con un poco de imaginación aceptaremos que los ejes de E representan variables y los de F individuos.

Para designar al eje (subespacio vectorial (sev) de dimensión 1) engendrado por el vector u, escribimos Δu.

A manera de resumen tenemos:

Representantes de $i \in \iota$ ⟵
- * El vector X_i en E
- * El eje Δf_i en F
- * La f lineal f^*_i en F^*



2.2.- LAS APLICACIONES X y X'.-

Veremos como se relacionan las distintas representaciones de individuos o variables y el papel que tiene la tabla de datos en dichas relaciones.

Pongamos en correspondencia cada f^*_i en F* con X_i en E. Esto define la aplicación lineal L : F*-----> E caracterizada por $L(f^*_i) = X_i$.

Esta transformación tiene asociada la matriz M con respecto a las bases respectivas de F* y E, cuyas columnas son las coordenadas en la base de E, de las transformadas de la base de F*: $L(f^*_i) = X_i = \Sigma_v X^v{}_i e_v$.

Luego, la i-ésima columna de M está formada por los números $X^1{}_i, X^2{}_i, \ldots, X^p{}_i$, de la i-ésima columna de la tabla de datos X. Por lo tanto, X es la matriz asociada a L.

Sea L' la aplicación (dual) transpuesta de L. Esta transformación lineal lleva E* en F** = F, de manera que $\langle L'(u), v \rangle = \langle u, L(v) \rangle$ para u en E* y v en F*. Se comprueba fácilmente que la matriz M asociada a L' es la tabla de datos "transpuesta" X'.

2.3.- METRICAS.-

Las acciones de clasificar y comparar, requieren una no - ción de proximidad entre los objetos estudiados. Por esta razón debemos metrizar los espacios E y F.

Para disponer del concepto de ángulo trabajaremos con mé - tricas euclidianas, es decir, que se originen en un producto in terno.

A una forma bilineal simétrica definida positiva $M : E \times E \longrightarrow \mathbb{R}$, $(x,y) \longrightarrow M(x,y)$ podemos asociarle:

* Una forma cuadrática $M : E \longrightarrow \mathbb{R}$, $x \longrightarrow M(x) = M(x,x)$

* Un isomorfismo $M : E \rightarrow E^*$, $x \longrightarrow M(x)$ en E^*
$\langle M(x), y \rangle = M(x,y)$

* Una norma euclidiana $\|x\|_M = \sqrt{M(x)}$

* Una distancia euclidiana $d(x,y) = \|x-y\|_M$

* Una noción de M-ortogonalidad $x' y$ sii $M(x,y) = 0$

* Una matriz M simétrica definida positiva (dp)

2.4.- METRICAS EUCLIDIANAS EN E,E*,F,F*.-

Sea M una métrica sobre E. Para cada par de puntos de la nube N(ı) calculamos su distancia como $||x_i - x_k||_M$. Dado que los individuos también se representan en F*, es natural pensar que la transformación X induce una métrica "heredada de M" que llamaremos W. Para el par de individuos i,k pedimos que se verifique $||X_i - X_k||_M = ||f^*_i - f^*_k||_W$ o más generalmente $||a-b||_W = ||X(a)-X(b)||_M$ para a,b en F* cualesquiera.

Puesto de manera equivalente tendremos:

$||a||_W = ||X(a)-X(b)||_M$ para cualquier a en F*, o bien

$W(a,b) = M(X(a),X(b))$ a,b en F*

Pero $M(X(a),X(b)) = \langle M \cdot X(a), X(B) \rangle$

$= \langle X' \cdot M \cdot X(a), b \rangle$

luego $\langle W(a), b \rangle = \langle X' \cdot M \cdot X(a), b \rangle$ para todo a,b en F*

por lo tanto $W = X' \cdot M \cdot X$

En el espacio de variables procedentes de manera análoga. Dada una métrica euclidiana N sobre F construimos la métrica V sobre E* a través de la aplicación X'. Para ello imponemos la condición $||a-b||_V = ||X'(a)-X'(b)||_N$ para todos a,b en E*, es decir,

$V(a,b) = N(X'(a),X'(b))$

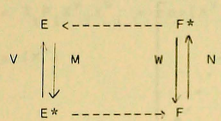
$= \langle N \cdot X'(a), X'(b) \rangle$

$= \langle X^{*}N^{*}X'(a), b \rangle$ para todos a,b en E*, luego

$$V = X^{*}N^{*}X'$$

2.5.- ESQUEMA DE DUALIDAD.-

Los espacios y transformaciones con los que hemos complicado progresivamente el panorama, son presentados en el esquema siguiente:



Donde podemos apreciar como las métricas sobre E y F inducen de manera natural las estructuras euclidianas sobre los duales.

NOTA:

Las formas bilineales V y W serán dp o semi dp según si X es inyectiva o no. En efecto, si $a \in \ker(X)$ entonces $W(a,a) = M(X(a),X(a)) = 0$.

El rango de W es el mismo que el de X. Como generalmente tendremos más individuos que variables ($n > p$), y $n = \dim(\ker(X)) + \text{rango}(x)$, resulta $\dim(\ker(X)) \geq n-p > 0$ luego W no es dp.

Razonando de manera idéntica se verifica que V es dp si $\text{rango}(x') = p$ lo que equivale a pedir que no haya variables

colineales.

2.6.- METRICA DE LOS PESOS EN F.-

Hay una métrica para las variables que permite interpre - tar geométricamente medias, varianzas y covarianzas. Para calcular la distancia entre dos variables se consideran las diferencias en cada individuo ponderado por su respectivo peso.

Esta métrica tiene asociada la matriz diagonal Dp dada por:

$$Dp = \begin{bmatrix} P_1 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ & & & & P_n \end{bmatrix}$$

La media aritmética de la variable v e υ se calcula como

$$me(x^v) = \Sigma_i P_i x^v{}_i \qquad v \text{ en } \upsilon$$

Sea 1 el vector de F de componentes 1,1,1...1 en la base canónica. Entonces $Dp(X^v,1) = \Sigma_i P_i X^v{}_i$, v en υ .

Por otra parte, $\|1\|_{Dp} = \Sigma P_i = 1$, luego 1 es vector unitario y $me(X^v)$ puede interpretarse como la norma de la proyección Dp-ortogonal de X^v sobre el eje $\Delta 1$. Puesto que la proyección ortogonal es el vector a mínima distancia, $me(x^v)1$ en F es la variable constante que mejor aproxima a X^v en el sentido de la métrica Dp.

La variable $X^v - me(X^v)1$, que es la proyección sobre el ortogonal $\Delta'1$, se llama centrada porque su media es nula.

Para mayor comodidad vamos a suponer en adelante que todas las variables de la tabla de datos son centradas, es decir $Dp(X^v,1) = 0$ para todo v en u. Veremos luego que esto equivale a fijar el origen del sistema de coordenadas de E en el centro de gravedad de la nube N(ι).

Sean X^v y X^s dos variables (centradas). Entonces

$$Dp(X^v,X^s) = \Sigma_i P_i X^v{}_i X^s{}_i = cov(X^v,X^s) \text{ y } \|X^v\|^2_{Dp} = Dp(X^v,X^v) = \Sigma_i P_i (X^v{}_i)^2 = var(X^v)$$

Si miramos el esquema de dualidad con N = Dp encontramos que la métrica V sobre el dual E* tiene asociada la matriz de covarianzas $X'Dp'X'$.

2.7.- CENTRO DE GRAVEDAD E INERCIA.-

Como si se tratara de un sistema de partículas, el centro de gravedad de la nube de individuos N(ι) se calcula como

$$g = \Sigma_i P_i X_i$$

Y veremos que las coordenadas de g son las medias: $g^v = \Sigma_i P_i X^v{}_i = me(x^j)$. Luego, las variables que describen la muestra son centradas sii g = 0, es decir, el origen de E esta en g.

Continuando con la analogía física definamos el momento de inercia de N(ι) con respecto a un punto a en E como:

$$Ia = \Sigma_i p_i \|x_i - a\|^2_M$$

Este valor lo interpretamos como una medida de la deformación que debemos introducir para proyectar la nube N(ι) sobre el punto a.

N(ι) se concentra en a sii Ia = 0.

Un resultado importante sobre la inercia se obtiene del siguiente cálculo.

$$\|x_i - a\|^2_M = \|x_i - g - (g-a)\|^2_M$$

$$= \|x_i - g\|^2_M + \|g-a\|^2_M - 2M(x_i - g, g-a)$$

Sumando sobre ι tenemos $\Sigma_i p_i \|x_i - a\|^2_M = \Sigma_i p_i \|x_i - g\|^2_M + \|g-a\|^2_M - 2M(\Sigma_i p_i (x_i - g), g-a)$, para llegar a :

$$Ia = Ig + \|g-a\|^2_M$$

Lo que significa que g es el punto con respecto al cual la inercia de N(ι) es mínima. Podemos imaginar que g es el subespacio (afín si las variables no fueran centradas) de dimensión cero que mejor aproxima a la nube.

3.0.- COMPONENTES PRINCIPALES.-

El ACP es un método que busca reducir la complejidad de la representación de los datos. Desde el punto de vista de los individuos, nos interesamos en proyectar la nube $N(\iota)$ sobre un SEV (afín) de baja dimensión, de tal manera que se preserve toda o gran parte de la estructura de ángulos y proximidades de la representación original.

¿Por qué reducir la dimensión?. Simplemente porque somos incapaces de ver en dimensión mayor que 3.

Desde la perspectiva de las variables, el ACP intenta sustituirlas por una colección reducida de descriptores no correlacionados, que capturen la mayor parte de la variabilidad de la muestra (máxima varianza).

Nos ubicamos en el espacio E de los individuos para plantear una sucesión de problemas que corresponden a reducciones óptimas de los datos, en orden creciente de complejidad.

P0. Determinar el punto más cercano a $N(\iota)$. Es decir, el sev (afín) de dimensión cero, óptimo para proyectar la nube. Averiguar si $N(\iota)$ se concentra en dicho punto.

P1. Determinar la recta más cercana a $N(\iota)$. Es decir, el sev (afín) de dimensión uno que provoque la menor distorsión al proyectar la nube. Averiguar si $N(\iota)$ se concentra en dicha recta.

.
.
.

Pk. Determinar el mejor sev (afín) de dimensión k para proyectar la nube, etc.

Usando el criterio de la inercia, la solución de PO ya la encontramos en el centro de gravedad g. Necesitamos extender este criterio a los sev para resolver P1, P2,...

Sea $W \subset E$ un sev(no afín). $E = W \oplus W'$. Un vector x_i de la nube $N(\iota)$ se descompone de manera única como $X_i = a_i + b_i$ donde $a_i \in W$ y $b_i \in W'$.

Si proyectamos x_i M-ortogonalmente sobre W hay una pérdida equivalente a la norma de la parte de x_i ortogonal a W, es decir, $\|b_i\|^2_M$. Sumando las pérdidas individuales adecuadamente ponderadas, llegamos a la inercia de $N(\iota)$ con respecto a W.

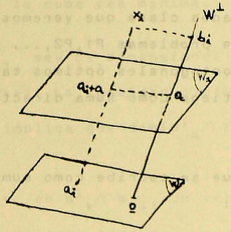
$$Iw = \Sigma_i p_i \|b_i\|^2_M$$

Es evidente que Iw = 0 sii $N(\iota)$ está contenida en W.

Del teorema de Pitágoras ($\|x_i\|^2 = \|a_i\|^2 + \|b_i\|^2$) recuperamos una identidad bastante útil:

$$Ig = Iw + Iw' \qquad (g=0)$$

La inercia con respecto a un subespacio afín W_1 (no contiene al cero) se define de manera muy natural. Sea W un sev de E, a en W' y $W_1 = W + a$. Entonces $W_1 \cap W' = \{a\}$ y



$x_i = a_i + b_i = (a_i+a) + (b_i-a)$ $a_i \in W$, $b_i \in W'$, luego $Iw_1 = \Sigma_i p_i \|b_i - a\|^2_M$. Usando la identidad $\|b_i-a\|^2 = \|b_i\|^2 + \|a\|^2 - 2M(b_i,a)$ llegamos a $Iw_1 = Iw + \|a\|^2 - 2M(\Sigma_i p_i b_i, a)$ pero $g = 0 = \Sigma_i p_i x_i = \Sigma_i p_i a_i + \Sigma_i p_i b_i$. Finalmente

$$Iw_1 = Iw + \|a\|^2 \geq Iw$$

Esto significa que W es mejor (en términos de menor distorsión) que cualquier ser afín paralelo. Por lo tanto, los problemas P1,P2,...,Pk tendrán por solución sólo sev (conteniendo al cero = g).

3.1.- DESCOMPOSICION DE LA INERCIA.-

Un par de resultados clave que veremos a continuación, permitirán reemplazar los problemas P1,P2,... por la búsqueda de una sucesión de ejes ortogonales óptimos tales que el sev W_k, solución de Pk, se obtiene como suma directa de los primeros k ejes.

Sea W sev de E que se escribe como suma directa de dos sev M-ortogonales w_1 y W_2:

$$W = W_1 \oplus W_2 \qquad W_1 \perp W_2 \qquad y$$

$E = W \oplus W' = W_1 \oplus W_2 \oplus W'$, luego $x_i = a_i + b_i = c_i + d_i + b_i$ con $a_i \in W$, $b_i \in W'$, $c_i \in W_1$, $d_i \in W_2$.

Entonces

$$Iw' = \Sigma_i p_i \|a_i\|^2 = \Sigma_i p_i \|c_i\|^2 + \Sigma_i p_i \|d_i\|^2$$

$$Iw' = Iw_1' + Iw_2'$$

Si preferimos una expresión en términos de Iw, podemos escribir $Iw = Ig - Iw'_1 - Iw'_2$.

El subespacio óptimo de dimensión k + 1 contiene un subespacio óptimo de dimensión k.

Esto significa que el sev W_{k+1}, solución de Pk+1, puede sustituirse a partir de W_k y un eje adicional, con respecto al cual la inercia de la nube sea mínima.

Sean W_k y W_{k+1} sev de dimensiones respectivas k y k+1, soluciones de Pk y Pk+1. La igualdad $\dim(w'_k) + \dim(W_{k+1}) = (p-k)+(k+1) = p+1$ implica que $\dim(W'_k \cap W_{k+1}) \geq 1$.

Sea entonces v en $W'_k \cap W_{k+1}$ un vector no nulo y descompongamos W_{k+1} en suma directa como $W_{k+1} = \Delta v \oplus R$, donde R(de dim k) es el suplementario M-Ortogonal de Δv en W_{k+1}.

Definamos también el sev u (de dim k+1 porque $v \in W'_k$) como $u = \Delta v \oplus W_k$ y comparemoslo con W_{k+1} en términos de la inercia.

Aplicando las fórmulas recientemente deducidas para la descomposición de la inercia llegamos a:

$$Iw_{k+1} = Ig - I\Delta'v - Ir' \qquad y$$

$$Iu = Ig - I\Delta'v - Iw'_k$$

Puesto que W_k es el sev óptimo de dimensión k, tenemos $Iw'_k \geq Ir'$ luego, de las ecuaciones anteriores deducimos que

$$Iw_{k+1} \geq Iu$$

Pero W_{k+1} es óptimo y por lo tanto debe verificarse la igualdad $Iw_{k+1} = Iu$, que implica finalmente $I_R = Iw_k$.

Hemos comprobado que W_{k+1} contiene un sev óptimo de dimensión k y que además, en virtud de las relaciones

$$u = \Delta v \oplus W_{k+1} \qquad y \qquad Iu = Ig - I\Delta' v - Iw'$$

es posible construir W_{k+1} "pegando" a W_k un eje Δv ortogonal a W_k, de mínima inercia. Resulta claro ahora que los problemas P1, P2,... son equivalentes a la siguiente secuencia de tareas:

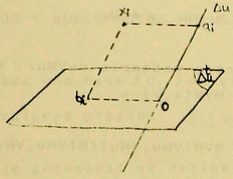
T1 : Determinar Δu_1 más cercano a N(ι) (mínima inercia)

T_2 : Determinar Δu_2 de mínima inercia, M-ortogonal a Δu_1.

TK. Determinar Δu_k de mínima inercia, M-ortogonal a los ejes Δu_1, Δu_2,..., Δu_{k-1} (es decir, al sev engendrado por $u_1, u_2, \ldots, u_k$)

3.2.- CALCULOS EXPLICITOS.-

Ahora comenzamos a resolver concretamente las tareas $T_1, T_2, \ldots$



$x_i = a_i + bi \qquad I\Delta^t u = \Sigma_i p_i \|a_i\|^2$

Asociemos a u una forma lineal v = Mu en E* y una variable c = X'v = X'Mu en F (ver esquema de dualidad). Entonces $a_i = M(x_i,u)u$ y $M(x_i,u) = \langle x_i, Mu \rangle = \langle Xf^*_i, Mu \rangle = \langle f^*_i, X'Mu \rangle = \langle f^*_i, c \rangle$.

Luego $c = \Sigma_i M(x_i,u) f_i$ y además $I\Delta\,'u = \Sigma_i p_i M(x_i,u)^2 = \|c\|^2_{Dp} = \langle DpC, C \rangle = \langle DpX'Mu, X'Mu \rangle = \langle xDpX'Mu, Mu \rangle$

$= \langle VMu, Mu \rangle = M(u,VMu)$

$= \langle MVMu, u \rangle = MVM(u,u)$.

Llegamos así a la fórmula

$$I\Delta\,'u = M(u,VMu) = M(u,VMu) = MVM(u,u)$$

que nos indica que $N(\iota) \subset \Delta'u$ sii VMU = 0, es decir, si u es vector propio de VM asociado al valor propio cero.

Supongamos que $N(\iota)$ no está contenida en $\Delta' u$ (hay parte de la nube que se estira a lo largo de Δu). Aplicando la desigualdad de Schwarz veremos que el eje engendrado por VMu es mejor que Δu, en el sentido que $I_\Delta VMu \leq I_\Delta u$. Luego, un eje inmejorable tendrá que ser invariante frente a VM. En otras palabras $I_\Delta VMu = I_\Delta u$ sii $VMu = \mu u$ $\mu \neq 0$.

Consideremos el vector unitario $v = VMu / \sqrt{M(VMu, VMu)}$. Es claro que

$$I_\Delta 'VMu = I\Delta' v = MVM(v,v) =$$

$$= MVM(VMu, VMu)/M(VMu, VMu)$$

$$= MVM(VMu, VMu)/MVM(u, VMu)$$

Pero

$$I\Delta' u = M(u, VMu) \leq M(VMu, VMu)/M(u, VMu) =$$

$$= MVM(u, VMu)/MVM(u, u) \leq$$

$$\leq MVM(VMu, VMu)MVM(u,u)/MVM(u,u)MVM(u, VMu) = I\Delta' v$$

lo que equivale a $I_\Delta VMu \leq I_\Delta u$.

Si Δu es un eje inmejorable, es decir, si $I_\Delta VMu = I_\Delta u$ entonces $M(u, VMu)^2 = M(VMu, VMu)$, igualdad que se cumple sii $VMu = \mu u$ $\mu \neq 0$. Además $I\Delta' u = M(u, VMu) = \mu M(u,u) = \mu$

La inercia $I_\Delta u$ es mínima si u pertenece al sev propio asociado al mayor valor propio de VM.

Hemos ligado el problema de los ejes de mínima inercia con el estudio de valores y vectores propios de VM. Es útil mencio-

nar ahora que VM tiene valores propios reales no negativos $\mu_1 \geq \mu_2 \geq \ldots \mu_p \geq 0$. Además podemos construir una base M-ortonormada $\{u_j ; j = 1,2,..,p\}$ de vectores propios de VM.

3.3.- EJES, PLANOS Y SUBESPACIOS PRINCIPALES.-

Suponemos dada una base $\{U_v\}$ de vectores propios de VM asociados a los valores propios $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_p \geq 0$.

El primer eje principal se define como el sev de dimensión 1 Δu tal que $I\Delta u$ es mínima.

Escribiendo u en la base de vectores propios, tendremos $u = \Sigma_v a_v U_v$ y $\Sigma_v (a_v)^2 = 1$. Además $I\Delta' u = M(u,VMu) = \Sigma_v \mu_v (av)^2$.

El problema de cálculo del primer eje principal se reduce a

$$\text{Minimizar} \quad \Sigma_v \mu_v (a_v)^2$$

$$\text{s.a} \quad \Sigma_v (a_v)^2 = 1$$

Cuya solución es $a_1 = 1$, $a_v = 0$ $v \neq 1$. Esto muestra que el primer eje principal es Δu_1 y $I\Delta' u_1 = \mu_1$.

El segundo eje principal corresponde al eje de mínima inercia, orotogonal a Δu_1. Se puede probar, con un cálculo muy similar al anterior, que dicho eje es Δu_2 y que $I\Delta' u_2 = \mu_2$.

Los primeros dos ejes principales determinan el plano principal $W_2 = \Delta u_1 \oplus \Delta u_2$, sev de mínima inercia: $Iw'_2 = \mu_1 + \mu_2$.

Continuando con el proceso descrito llegamos los k primeros ejes principales $\Delta u_1, \Delta u_2, \ldots \Delta u_k$, los que engendran el sev principal $W_k = \Delta u_1 \oplus \ldots \oplus \Delta u_k$ y $Iw'_k = \mu_1 + \ldots + \mu_k$.

3.4.- CALIDAD DE LA REPRESENTACION.-

Lo primero es saber que entendemos por representación. Recordemos que estamos tras una presentación simple de los datos y para ello vamos a proyectar la nube N(ı) en su sev de dimensión baja, digamos k, perdiendo un mínimo de información.

Todo el desarrollo teórico anterior nos lleva a escoger W_k para proyectar. La pregunta es ¿qué tan buena es la representación? y la respuesta dependerá del valor de un índice que definiremos más abajo.

La información que contienen los datos podemos imaginarla asociada a la variabilidad de los mismos.

Si proyectamos N(ı) al centro de gravedad perdemos toda variabilidad, luego es natural interpretar el valor Ig como una medida de la información total de la muestra. Por otra parte, puesto que Iw_k mide la distorsión debida a la proyección sobre W_k, debemos interpretarla como pérdida de información al proyectar. Alternativamente, $Iw'_k = Ig - Iw_k$ puede verse como la parte de información que explica el sev W_k.

El índice porcentaje de inercia explicada por el sev principal W_k se define como

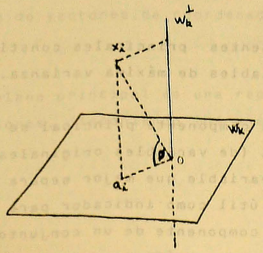
$$i_k = Iw'_k/ig$$

pero $Iw'_k = \mu_1 + \ldots + \mu_k$, $Ig = I_{E'} = I\Delta' u_1 + \ldots + I\Delta' u_p = \mu_1 + \ldots + \mu_p$ = traza(VM). Luego

$$i_k = \Sigma_{i=1}^{k} \mu_i / \text{trazar}(VM)$$

Si $i_k = 1$ la representación en W_k es perfecta y $\mu_v = 0$ para $v = k+1, K+2, \ldots p$. En tal caso rango(VM) = k = rango(X) lo que significa (suponiendo $n > p$) que hay multicolinealidad (redundancia lineal) en las variables.

El porcentaje de inercia es un índice global de calidad. Es posible y a veces recomendable recurrir a índices que muestran si uno o varios individuos están bien representados por su proyección en W_k.



Por ejemplo $\cos^2(\theta) = (M(x_i, U_1)^2 + \ldots + M(X_i, U_k)^2) / \|x_i\|^2_M$

3.5.- LAS COMPONENTES PRINCIPALES.-

Asociemos a los ejes principales ΔU_v en E, variables C^j en F, llamadas componentes principales, a través de las aplicaciones M y X' del esquema de dualidad:

$$C^v = X'MU_v \qquad v = 1,2,\ldots,p$$

Notemos sus propiedades más importantes:

1) Son centradas, es decir de media nula $me(C^v)=Dp(C^v,1) =$

$$<Dp1,X'MU_v> = <XDp1,MU_v> = <0,MU_v> = 0.$$

2) Son no correlacionadas: $COV(C^v,c^s) = Dp(C^v,C^s) =$

$$<DpX'MU_v,X'MU_s> = <XDpX'MU_v,MU_s> =$$

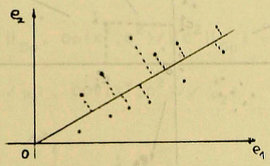
$$<VMu_v,MU_s> = \mu_v M(U_s,U_v) = 0$$

3) $var(C^v) = Dp(C^v) = <VMU_v,MU_v> = \mu_v M(u_v,U_v) = {}_v.$

Las componentes principales constituyen un sistema Dp-ortogonal de variables de máxima varianza. Una base para F.

La primera componente principal se interpreta como la combinación lineal (de variables originales) que tiene varianza máxima. Es la variable que mejor separa a los individuos y en ese sentido es útil como indicador para ordenarlos. Por ejemplo la primera componente de un conjunto de variables económicas, culturales, demográficas, etc., serviría para definir un

indicador de desarrollo de países o regiones.



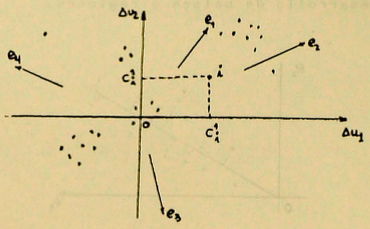
3.6.- SALIDAS GRAFICAS.-

Gran parte del interés del ACP está en los gráficos de los resultados, que permiten visualizar relaciones entre variables así como la organización de los individuos.

En aplicaciones prácticas del método, se seleccionan los primeros ejes tales que el porcentaje de inercia explicado por ellos sea al menos 90%, pero habitualmente nunca se analizan más de 4 ó 5.

La proyección de la nube $N(^1)$ sobre el plano principal produce una colección de vectores de coordenadas $M(x_i,U_1)$, $M(x_i,U_2)i = 1,2,..,n$.

El gráfico del plano principal es una representación cartesiana de las proyecciones, disponiendo los ejes ΔU_1 y ΔU_2 perpendicularmente.



El gráfico puede enriquecerse incorporando por ejemplo identificadores de individuos y los porcetajes de inercia de los ejes.

Al clasificar visualmente a los individuos es interesante dar sentido a los grupos en términos de las variables antiguas. Si proyectamos los vectores de la base canónica de $E\{e_1, e_2, .. e_p\}$ (que representan a las viejas variables), orientaremos el plano y podremos explicar las clasificaciones y proponer interpreta - ción de las componentes principales.

Para completar el análisis puede ser interesante mirar otros planos como ΔU_1 @ ΔU_3, ΔU_2 @ ΔU_3 etc.

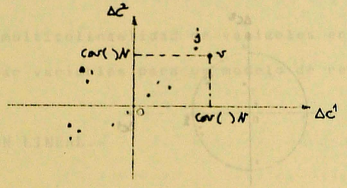
3.7.- REPRESENTACION DE VARIABLES.-

Usando la base de F, de las componentes principales normalizadas $\{C^1/\sqrt{\mu_1}, \ldots, C^p/\sqrt{\mu_p}\}$ representamos las variables en un gráfico cartesiano, cuya calidad depende del índice de inercia. Si éste es alto, los ángulos y distancias entre las proyeccio - nes reflejarán los ángulos y distancias entre las variables ori ginales.

Escogiendo las dos primeras componentes, la variable x^v estará representada por el punto de coordenadas

$$(Dp(x^v,c^1)/\|c^1\|_{Dp},\ Dp(x^v,c^2)/\|c^2\|_{Dp}) =$$

$$(COV(x^v,c^1)/\sqrt{\mu_1}\ ,\ COV(x^v,c^2)/\sqrt{\mu_2})$$



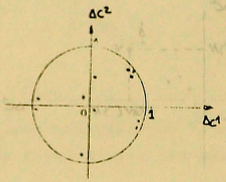
3.8.- LA ELECCION DE LAS METRICAS.-

Dos son las métricas M que se usan comúnmente en el espacio E. La primera de ellas es la identidad o métrica euclidiana usual. Según esta métricas todas las direcciones tienen igual importancia en el cálculo de la distancia y esto hace que los resultados del ACP dependan de las unidades de medición de las variables.

La otra alternativa es una distancia que otorga pesos a las direcciones, en razón inversa a la dispersión de las variables que representa. Esta métrica se apoya en el principio según el cual se otorga menor peso a las medidas hechas con instrumentos de poca precisión. Su matriz con respecto a la base canónica es diagonal con los recíprocos de las varianzas de $x^1, x^2, \ldots, x^p$.

Se puede demostrar que la elección de esta métrica es equivalente a usar la identidad sobre los datos estandarizados (variables centradas y de varianza 1).

En estas condiciones la matriz V = XDpX' de covarianzas se convierte en matriz de correlaciones y el gráfico de las variables se transforma en el círculo de correlaciones,



justamente porque ahora las coordenadas de X^v son $cor(X^v,C^1)$ y $cor(X^v,C^2)$. Debido a que $||X^v|| = 1 = var(X^v)$ resulta

$$1 = || \Sigma_s (Dp(X^v,C^s)/\sqrt{\mu_s}) C^s/\sqrt{\mu_s} ||^2_{Dp} = \Sigma_s (Dp(X^v,C^s)/\sqrt{\mu_s})^2,$$

luego los puntos del gráfico de variables caen en el interior (o el borde) del círculo unitario. La cercanía al borde indica mejor calidad de la representación.

3.9.- APLICACIONES DEL ACP.-

El ACP puede ser empleado provechosamente para responder a problemas como los siguientes:

1) Clasificar individuos descritos por variables cuantitativas.

2) Detectar observaciones aberrantes o anómalas.

3) Detectar factores o variables "escondidas"

4) Economizar espacio de almacenamiento de datos.

5) Construir indicadores (desarrollo, inteligencia, etc.)

6) Representar gráficamente una estructura de correlaciones.

7) Hacer una representación euclidiana de una tabla de distancias.

8) Detectar multicolinealidad de variables en un modelo lineal.

9) Seleccionar variables para un modelo de regresión lineal.

4.- REGRESION LINEAL.-

Entra en escena una nueva variable que vamos a representar como el vector y en el espacio F.

El problema que abordaremos en esta sección del curso consiste en explicar el comportamiento de y en términos de las variables $x^1, x^2, \ldots, x^p$.

El enunciado del problema es particularmente vago porque un estudio de regresión puede tener muy variados objetivos:

1) Describir la dependencia entre $x^1, x^2, \ldots, x^p \in y$.

2) Estudiar el comportamiento de y retirando los efectos de confusión de otras variables.

3) Determinar una ley empírica que relacione y con las otras variables.

4) Ajustar un modelo para suavisar datos.

5) Contribuir con evidencia empírica a un análisis causal.

6) Predecir el comportamiento de y cuando se conocen los valores de $X^1,\ldots,X^p$ para un nuevo individuo.

7) Controlar un sistema complejo.

La perspectiva de la regresión para responder a esta serie de tareas distintas es la construcción de un modelo que ligue y con $X^1,\ldots,X^p$, a través de una función f sencilla (interpretable si es posible) tal que para cada i en ı el valor y_i se determine aproximadamente como $f(X^1{}_i,\ldots,X^p{}_i)$. La diferencia se atribuye a un error aditivo e_i desconocido.

Tenemos entonces:

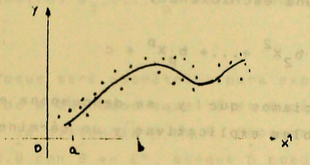
$$y_i = f(X^1{}_i, X^2{}_i, \ldots, X^p{}_i) + e_i \qquad i = 1,2,..,n$$

Para escoger adecuadamente la función f es necesario definir un criterio, una familia de candidatos (paramétrica o no) y dejar que los datos decidan.

En el marco de la regresión lineal la función f se toma lineal en los argumentos, es decir, los candidatos para modelo tienen la forma $f(X_1,\ldots,X_p) = \Sigma_1 b_1 x_1$ donde $b_1, b_2, \ldots, b_p$ son números reales llamados parámetros del modelo

$$y_i = b_1 X^1{}_i + b_2 X^2{}_i + \ldots + b_p X^p{}_i \qquad i = 1,2\ldots n \qquad (1)$$

La razón para trabajar con modelos aparentemente tan sim - ples como los lineales se debe a que por una parte la aproxima- ción lineal puede ser suficiente en cierto rango de las varia - bles $X^1,\ldots,X^p$, lo que no justifica una función más compleja. Por otra parte, el cálculo numérico y el análisis teórico de los modelos lineales es harto más simple que las alternativas no lineales.



En la figura anterior se presenta una forma de dependencia entre y y x^1 que sería perfectamente aproximada por una recta en el rango [a,b].

La variable y y las restantes $X^1,\ldots,X^p$, tienen roles asi- métricos y suelen llamarse respectivamente variable a explicar y variables explicativas.

Tradicionalmente el estudio del modelo de RL va precedido de una serie hipótesis probabilísticas sobre las variables y los errores. En una primera etapa vamos a prescindir de ellas para enfrentar el problema con espíritu descriptivo.

4.1.- NOTACION MATRICIAL-VECTORIAL.-

Presentaremos el modelo de RL usando los mismos espacios que intervienen en el ACP (ver esquema de dualidad).

Definamos el vector de errores e en F cuyas coordenadas en la base canónica son los e_i. Entonces las ecuaciones en (1) la fundimos en una escribiendo:

$$y = b_1x^1 + b_2x^2 + \ldots + b_px^p + c \qquad (2)$$

Donde apreciamos que y se descompone en una combinación lineal de variables explicativas y un término de error e.

Sea w el sev de F engendrado por las variables explicativas $x^1,\ldots,x^p$. La ecuación (2) nos dice que y se explica mediante una componente a lo largo de w más un error.

En la ecuación (1) podemos interpretar al vector de coeficientes B como una forma lineal en E* porque asocia al individuo X_i en E el valor $\Sigma_v b_v x^v{}_i = B(X_i) = \langle B, X_i \rangle$.

En este sentido B se comporta como el representante en E* de una variable. Para recuperar la auténtica variable de F que B representa recurrimos a la aplicación X' y obtenemos X'B. Pero si $B = \Sigma_v b_v e^*_v$ tenemos $X'B = \Sigma_v b_v X'e^*_v = \Sigma_v b_v x^v$.

De manera que la ecuación (2) es equivalente a la expresión matricial

$$y = X'B + e$$

Es bueno notar que W = X'(E*) y que dim(W) = rango(X') = número de variables linealmente independientes.

PROBLEMA CENTRAL: ENCONTRAR B EN E*, OPTIMO.

El problema planteado se conoce como problema de estima - ción de los parámetros $b_1,\ldots,b_p$. Para resolverlo necesitamos un criterio.

Nuestro enfoque será geométrico para explicar o aproximar y con un vector de W buscamos aquel $\hat{y}$ en W más cercano de a y en el sentido de la distancia N de F. Necesariamente $\hat{y}$ se es - cribe como $\hat{y} = X.\hat{B}$ con $\hat{B}$ en E*, aunque $\hat{B}$ puede no ser único. Consideramos a $\hat{B}$ como solución del problema central.

La unicidad de $\hat{B}$ se consigue sii X' es inyectiva en efec- to si $X'B_1 = X'B_2$ tendríamos $\ker(X') \neq \{0\}$

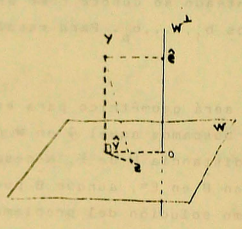
Si X' no es inyectiva se dice que el parámetro no es identificable, problema que está ligado a la estimabilidad (in- sesgada) de combinaciones lineales de los b_v.

Desde el punto de vista de las variables, cuando X'no es inyectiva se habla de multicolinealidad.

El enfoque geométrico que hemos adoptado corresponde al famoso criterio de mínimos cuadrados. Si N = 1 se habla de mí - nimos cuadrados ordinarios. Si N es cualquier métrica euclidia- na se le llama mínimos cuadrados generalizados.

4.2.- SOLUCION DEL PROBLEMA CENTRAL.-

Nos basamos en el siguiente resultado muy conocido, en el punto $\hat{y}$ de W más cercano de de y en el sentido de N es la proyección N-ortogonal de y sobre W. Para convencernos miremos la figura y los cálculos a continuación



Designemos por A el proyector N-ortogonal sobre W. Entonces $y = \hat{y} + \hat{e} = A(y) + (1-A)(y)$.

Si Z en W es cualquiera tenemos: $d(y,x)^2 = \|y-Z\|^2 = \|y-\hat{y}+y-Z\|^2 = \|y-\hat{y}\|^2 + \|\hat{y}-Z\|^2 \geq \|y-\hat{y}\|^2 = d(y,\hat{y})^2$.

Busquemos ahora la forma explícita para $\hat{y}$ (o para el proyector A), partiendo del simple hecho que $\hat{e}$ (vector de residuos) es ortogonal a $W = X'(E^*)$.

Para cualquier e^*_v, $\hat{e}) = \langle NX'e^*_v, \hat{e}\rangle = \langle e^*_v, XN\hat{e}\rangle$

$= \langle e^*_v, XN(y-\hat{y})\rangle = 0$

de donde salen las famosas ecuaciones normales:

$$XNy = XN\hat{y} = XNX'\hat{B} = V\hat{B}$$

Si X' es inyectiva V es invertible y podemos despejar la solución $\hat{B}$ única:

$$\hat{B} = V^{-1}XNy = (XNX')^{-1}XNy.$$

Cuando X' no es inyectiva se puede recurrir a inversas generalizadas para seleccionar alguna solución de las ecuaciones normales.

La expresión para el proyector en el modelo de rango completo (X'inyectiva) es

$$A = X'(XNX')^{-1}XN.$$

4.3.- CALIDAD DE LA REGRESION.-

La regresión será buena en la medida que la parte de y que no se logra explicar sea pequeña. En otras palabras, mejor es la calidad mientras más chico sea ê.

El índice r^2 de correlación lineal múltiple mide la calidad global de la regresión en los términos expresados más arriba:

$$r^2 = \|\hat{y}\|^2 / \|y\|^2$$

y no es otra cosa que el $\cos^2$ del ángulo entre y y el sev W.

En general la calidad de la regresión depende del comportamiento global de los residuos $\hat{e}$, cuyo análisis detallado sirve para comprender y corregir el modelo.

4.4.- HIPOTESIS PROBABILISTICAS.-

El modelo lineal $y = X'B + e$ suele adornarse con muchas hipótesis probabilísticas que permiten, entre otras cosas, demostrar la optimalidad del criterio de mínimos cuadrados y efectuar test de hipótesis estadísticas sobre los parámetros.

Estas hipótesis facilitan el tratamiento teórico elegante del modelo de RL pero la validez de las mismas es difícilmente verificable a partir de los datos.

Consideremos algunas de ellas:

H1) y es un vector aleatorio.
H2) Las variables explicativas no son aleatorias
H3) El vector de errores tiene esperanza nula. $E(e_i)=0$.
H4) Los errores son no correlacionados. $E(e_i e_k) = 0$.
H5) Los errores son independientes.
H6) Las varianzas de los e_i son iguales.
H7) e tiene distribución normal multivariada.

4.5.- TEOREMA DE GAUSS MARKOV.-

Con H1, H2 y H3 se puede demostrar un interesante teorema sobre la calidad del estimador de mínimos cuadrados $\hat{B}$ y además indicar la métrica N que se debe escoger.

En la fórmula $\hat{B} = (XNX')^{-1}XNy$ observamos que $\hat{B}$ es función lineal de y.

El teorema de Gauss Markov afirma que entre todos los estimadores $\tilde{B}$ lineales en y, $\hat{B}$ es mejor, en el sentido que la diferencia de las matrices de varianza-covarianza $V(\tilde{B}) - V(\hat{B})$ es semi dp, siempre que la métrica N se tome como la inversa de la matriz de varianzas covarianzas del vector de errores $E(ee') = \Sigma$

Si agregamos la hipótesis H7 se puede reforzar la conclusión del teorema de Gauss Markov, diciendo que $\hat{B}$ es el mejor entre todos los estimadores insesgados.

Finalmente corresponde notar que la hipótesis de normalidad permite, al menor en principio conocer la distribución de $\hat{y}$, $\hat{B}$, $\hat{e}$ y de otros estadísticos de interés, brindando al usuario la teoría de la inferencia estadística clásica; estimación y test de hipótesis.

4.6.- INTERPRETACION Y ESTABILIDAD DE B.

El modelo lineal provoca en el analista la legítima tentación de interpretar los coeficientes. Imaginarlos por ejemplo como elasticidad precio-demanda, índice de reactividad, coeficiente de aislación, etc. Esta práctica es interesante pero peligrosa puesto que el significado de un coeficiente depende no sólo de la variable que multiplica sino también de las restantes (recordar que en general las variables explicativas están correlacionadas).

Por otra parte el coeficiente que multiplica a una variable dada cambia si el conjunto de variables "acompañantes" es alterado.

Otro problema delicado es la estabilidad de $\hat{B}$, que podemos formular como una pregunta: ¿qué confianza se puede tener en el valor de $\hat{B}$ calculado. La respuesta depende, por un lado de la precisión de los algoritmos de cálculo, (factor no despreciable) y por otro de las relaciones de colinealidad entre las variables explicativas.

Para enfrentar situaciones concretas se procede a estimar la varianza del estimador del coeficiente de interés y luego se usa ese valor como indicador de calidad.

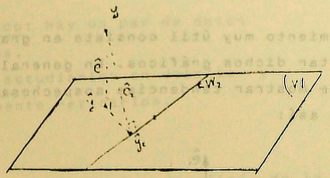
Hay un último aspecto importante que mencionar la significación de $\hat{B}$. Esto se refiere a decidir si uno o varios betas son significativamente no nulos, o si por el contrario, son distintos de cero solo por azar.

En términos del modelo, la significación de los betas habla de la relevancia que las variables correspondientes tienen en el modelo. Si demostramos (o decidimos) que b_v debería ser cero, estamos diciendo que X^v debe salir dle modelo, porque en nada contribuye.

Este tipo de problema cae en el ámbito del análisis de varianza y los test F que no se tratan en este apunte. Sin embargo hay una interesante aproximación geométrica al problema de la significación que vale la pena considerar:

Supongamos que B se particiona en dos pedazos B_1 y B_2. Sospechamos que las variables asociadas a B_1 no son relevantes, es decir, creemos que B_1 debería ser cero.

Para comprobar esto procedemos como sigue: estimar el modelo completo, luego estimar el modelo botando las variables sospechosas y finalmente comparar. ¿Cómo?, decidiendo en base a los tamaños de los respectivos vectores de residuos.



W = Espacio engendrado por todas las variables explicativas.

W_2 = Espacio de las variables no sospechosas

$\hat{y}$ = Proyección N-ortogonal en W.

$\hat{y}_2$ = Proyección N-ortogonal en W_2

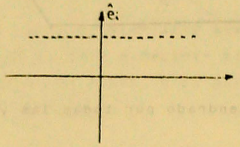
$\hat{e} = y-\hat{y}$, $\hat{e}_2 = y-\hat{y}_2$

La regla de decisión es botar las variables sospechosas si $\|\hat{e}_2\|$ no es significativamente más grande que $\|\hat{e}\|$

4.7.- LOS RESIDUOS.-

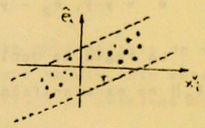
El vector de los residuos $\hat{e}$ contiene valiosa información sobre datos aberrantes, errores en la especificación del modelo, validez de los supuestos probabilísticos, etc.

Un procedimiento muy útil consiste en graficar los residuos e interpretar dichos gráficos. En general un gráfico de residuos no debe mostrar tendencias sospechosas. Debe verse aproximadamente así:

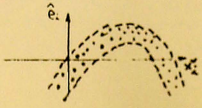


Situaciones que se apartan de este caso ideal, admiten algunas explicaciones como veremos en las figuras a continuación:

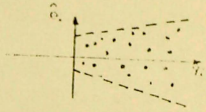
Diagnóstico: falta término lineal en la variable X^v.
Remedio: agregarlo



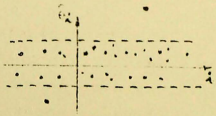
Diagnóstico: falta término cuadrático en X^v.
Remedio: agregarlo



Diagnóstico: la igualdad de varianzas de los y_i es falsa.
Remedio: Cambiar métrica N o transformar la variable y.



Diagnóstico: hay un par de datos aberrantes.
Remedio: estudiarlos en detalle y eventualmente retirarlos.



Apenas hemos tocado la superficie de la regresión lineal, en teoría o práctica. A quienes deseen profundizar en el tema recomiendo el excelente clásico "Applied Regression Analysis" de N. Draper y H. Smith (2° edición).

5.- REFERENCIAS.-

[1] Draper N. Smith H., Applied Regression Analysis, Sec. ed. Wiley. 1979.

[2] Mosteller F. Tuckey J., Data Analysis and Regression, Ad. Wesley, 1977.

[3] Pages J.P. Cailliez F., Introducción a L'analyse des Donnees, Smash, 1976.

6.- ABREVIATURAS.-

ACP = Análisis en Componentes Principales
dp = definida positiva
etc = etcétera
RL = Regresión lineal
sev = subespacio vectorial
sii = si y sólo si