

EL ENFOQUE GEOMETRICO EN LA ESTADISTICA *

Roberto Behar Gutiérrez.
Profesor Asociado.
Universidad del Valle.
Cali - Colombia.

RESUMEN

El objetivo del presente artículo es deducir algunos resultados elementales de la estadística, al enmarcar el problema estadístico en el contexto geométrico que surge de un espacio vectorial con un producto interior bien definido. Con esto se pretende motivar sobre la ganancia conceptual y la capacidad de generalización que se logra al involucrar el enfoque geométrico en la estadística.

Se trata inicialmente la situación muestral (n -uplas) y posteriormente la situación poblacional (variables aleatorias), deduciendo en ambos casos las expresiones para la media, la desviación estándar, el coeficiente de correlación, el modelo de regresión y los indicadores de bondad, desde el punto de vista geométrico.

INTRODUCCION

En la investigación en estadística ha imperado el enfoque algebraico sobre el enfoque geométrico, aunque en múltiples ocasiones se recurre a conceptos geométricos para justificar ó argumentar algunos pasos de un desarrollo algebraico. Se conocen también algunos casos, aunque poco frecuentes, en que aportaciones estadísticas han sido basadas casi exclusivamente en un enfoque geométrico, como en los trabajos de Fisher (1915) sobre la distribución del coeficiente de correlación de una muestra en n pares de una distribución normal

bivariada; también los trabajos de Durbin y Kendall (1951) en el estudio de la geometría de la estimación.

Como estos trabajos hay otros como los de Kruskal (1961, 1968, 1975), Zyskind (1967), Watson (1967); para una breve revisión de estos trabajos y una bibliografía más amplia se remite al lector a la referencia (7).

Aunque en los trabajos citados se visualiza la gran solvencia del enfoque geométrico y su fuerza conceptual, éste no ha sido el predominante, por varias razones entre las cuales está la tradición (Inercia) del enfoque algebraico en la enseñanza de la estadística.

Vale la pena sembrar algunas inquietudes sobre el enfoque geométrico en la estadística, por sus múltiples ventajas no solo en la investigación y generación de conocimientos sino también en el ámbito de la formación de potenciales investigadores, sobre todo si aceptamos que conceptos como puntos, líneas, planos, perpendicularidad, etc, están dentro de nuestra cotidianidad, lo cual hace más "natural" el enfoque geométrico.

Con el propósito de unificar el lenguaje de las ideas geométricas que se tratarán más adelante se presentará a continuación algunos elementos de geometría.

* Ponencia presentada en el XX Congreso Nacional de la Sociedad Matemática Mexicana. Xalapa-Veracruz-México. Noviembre de 1987.

GEOMETRIA

Cuando se habla de geometría casi inmediatamente se asocia con esta palabra ideas como punto, línea, plano, ángulo, etc, y además se hace corresponder con ellas una representación gráfica, ligando así el concepto a lo que se conoce comunmente como **geometría analítica**.

En forma un poco más general podría asociarse a geometría, ideas como vectores entendidos como segmentos de recta que parten del origen y están caracterizados por su magnitud, dirección y sentido; se pensaría posiblemente en **ángulos** entre dichos vectores, en rectas que los continene, en planos que pueden ser generados por estos y en cierta idea de ortogonalidad asociada casi siempre con una posición espacial y con un ángulo entre vectores (factible de medir con un transportador); esta concepción podría llamarse **Geometría Vectorial**.

No obstante la aparente universalidad de las concepciones mencionadas sobre geometría, ellas constituyen solo casos particulares de conceptos más generales de estructuras que permiten definir todas las ideas propuestas con niveles más altos de abstracción, como se revela en la siguiente definición de espacio vectorial.

DEFINICION 1. (Espacio Vectorial).

Un conjunto V con una operación suma $(+)$ y una operación producto por un escalar (λ) :

$+: V \times V \rightarrow V$ y $\lambda: R \times V \rightarrow V$;
donde R es el campo de los números reales, tales que para todo $v_1, v_2, v \in V$; $\lambda, c, d \in R$ se cumple que:
a) operación suma es cerrada, conmutativa, asociativa, además en V existe un elemento neutro con respecto a la suma; cada elemento de V tiene su inverso aditivo;
 $\exists 1 \in R / 1v = v, \forall v \in V$.

La operación producto por un escalar cumple que:

- i) $\lambda v \in V$
- ii) $c(dv) = (cd)v$
- iii) $c(v_1 + v_2) = cv_1 + cv_2$
- iv) $(c + d)v = cv + dv$

si esto sucede se dice que $(V, +, \lambda)$ es un espacio vectorial.

DEFINICION 2. Vector es todo elemento de un espacio vectorial.

NOTACION: Los vectores que son arreglos ordenados,

se denotan con letra en negrilla.

$$Y = (Y_1, Y_2, \dots, Y_n)'$$

Con base en estas definiciones se presentan algunos ejemplos de espacios vectoriales.

ALGUNOS ESPACIOS VECTORIALES DE INTERES ESTADISTICO

- a. $(R^n, +, \lambda)$, $\lambda \in R$. Con la suma y el producto por un escalar definidos componente a componente. Podría ser el espacio vectorial de las realizaciones de una muestra aleatoria de tamaño n , de una variable aleatoria Y .

En este caso un vector es $(y_1, y_2, \dots, y_n)' = Y$.

- b. $(M_{m \times p}, +, \lambda)$, $\lambda \in R$. donde $M_{m \times p}$, representa el conjunto de las matrices de orden $m \times p$ con entradas reales. De nuevo la suma y el producto por un escalar definido componente a componente respectivamente.

Podría ser el **espacio vectorial** de las realizaciones de una muestra aleatoria de tamaño n de una variable multivariada $(Y_1, Y_2, \dots, Y_p)'$.

En este caso un vector es una matriz de $m \times p$.

- c. $(F, +, \lambda)$, $\lambda \in R$ donde F representa el conjunto de las funciones continuas de variables real y valor real. La suma definida como

$f, g \in F$; $(f + g)(x) = f(x) + g(x)$ y el producto por un escalar:

$$(\lambda f)(x) = \lambda f(x).$$

En este caso un vector es una función.

- d. $(A, +, \lambda)$, $\lambda \in R$. Donde A representa el conjunto de las variables aleatorias continuas con dominio común. Donde la suma y el producto por un escalar pueden mirarse como el definido en funciones continuas. En este caso un vector es una variable aleatoria.

Nótese que en la estructura de espacio vectorial solo están definidas las operaciones suma y producto por un escalar, las cuales no posibilitan por sí solas la definición de características importantes de los espacios vectoriales como son los conceptos de norma, ángulo, ortogonalidad etc. Se hace necesario dotar los espacios vectoriales con nuevas operaciones que permitan la introducción de estas ideas; surgen así los denominados espacios vectoriales con producto interior.

DEFINICION 3. Espacio Vectorial con producto interior es un espacio vectorial $(V, +, \lambda)$, $\lambda \in R$ en el cual se define una operación llamada **producto interior entre vectores**, $\langle, \rangle$, que satisface los siguientes axiomas.

$\langle, \rangle : V \times V \rightarrow R$, tal que si $v_1, v_2, v_3 \in V$ y $\lambda \in R$ entonces

- i) $\langle v_1 + v_2, v_3 \rangle = \langle v_1, v_3 \rangle + \langle v_2, v_3 \rangle$
- ii) $\langle \lambda v_1, v_2 \rangle = \lambda \langle v_1, v_2 \rangle$
- iii) $\langle v_1, v_1 \rangle \geq 0$
- iv) $\langle v_1, v_1 \rangle = 0$ si y solo si $v_1 = 0$

DEFINICION 4. espacio vectorial normado es un espacio vectorial $(V, +, \lambda)$ sobre el cual se ha definido una función de valor real la cual aplica cada elemento v de V en un número real $\|v\|$, llamado norma de v . La norma satisface los siguientes axiomas:

- a. $\|v\| > 0$ para todo $v \in V$, $\|v\| = 0$ si y solo si $v = 0$
- b. $\|v_1 + v_2\| \leq \|v_1\| + \|v_2\|$ para cada, $v_1, v_2, \in V$
- c. $\|\lambda v\| = |\lambda| \|v\|$, para cada escalar λ y cada $v \in V$. Puede probarse con base en el axioma b) que $\|v_1\| - \|v_2\| \leq \|v_1 - v_2\|$ para cualquiera $v_1, v_2 \in V$.

LEMA. La función $\|v\| = \sqrt{\langle v, v \rangle}$ define una norma en el espacio vectorial $(V, +, \lambda)$; en el sentido de la definición 4.

DEFINICION 5. (Distancia)

Se llama distancia entre elementos de un conjunto V , a la función con dominio $V \times V$ que aplica en los números reales, y que satisface los siguientes axiomas, para cualesquiera $v_1, v_2, v_3, v \in V$

- i) $d(v_1, v_2) \geq 0$
- ii) $d(v, v) = 0$
- iii) $d(v_1, v_2) = d(v_2, v_1)$
- iv) $d(v_1, v_2) + d(v_2, v_3) \geq d(v_1, v_3)$

LEMA. La norma de la definición 4, induce una función de distancia en un espacio vectorial con producto interior:

$$d(v_1, v_2) = \|v_1 - v_2\|$$

DEFINICION 6. (Ortogonalidad)

Se dice que los vectores v_1, v_2 de un espacio vectorial con producto interno, son ortogonales si $\langle v_1, v_2 \rangle = 0$.

Por otro lado, de la desigualdad de Cauchy -Schwarz, se sabe que

$$\frac{|\langle v_1, v_2 \rangle|}{\langle v_1, v_1 \rangle^{1/2} \langle v_2, v_2 \rangle^{1/2}} \leq 1$$

nótese que cuando v_1 y v_2 son ortogonales la expresión de la izquierda toma el valor cero y cuando $v_1 = cv_2$, con $c \in R$, entonces dicha expresión es la unidad. Este hecho da soporte intuitivo a la siguiente definición.

DEFINICION 7. El ángulo θ entre dos vectores v_1, v_2 de un espacio vectorial con producto interior se define a través de

$$\cos \theta = \frac{\langle v_1, v_2 \rangle}{\langle v_1, v_1 \rangle^{1/2} \langle v_2, v_2 \rangle^{1/2}}$$

nótese que dos vectores v_1, v_2 son ortogonales si y solo si $\cos \theta = 0$.

PRODUCTOS INTERIORES PLAUSIBLES EN ESPACIOS DE INTERES

a) **Espacio Vectorial:** $(R^n, +, \lambda)$, $\lambda \in R$

$$\text{producto interior: } \langle x, y \rangle = \sum_{j=1}^n x_j y_j = x' y$$

donde las x_j, y_j son las componentes respectivas de los vectores x y y .

Criterio de ortogonalidad: $x' y = 0$,

$$\text{es decir } \sum_{j=1}^n x_j y_j = 0$$

$$\cos \theta = \frac{x' y}{(x' x)^{1/2} (y' y)^{1/2}}$$

$$= \frac{\sum x_j y_j}{[\sum x_j^2]^{1/2} [\sum y_j^2]^{1/2}}$$

b) **Espacio Vectorial:**

$(M_{n \times p}, +, \lambda)$, $\lambda \in \mathbb{R}$. Espacio de las matrices de $n \times p$.

Producto interior: $\langle M_1, M_2 \rangle = \sum_{j=1}^n m'_{1j} m_{2j}$

donde m'_{kj} es el vector de $\mathbb{R}^p$ constituido por la j -ésima fila de la matriz M_k , ($k = 1, 2$).

Criterio de ortogonalidad:

$$\sum_{j=1}^n m'_{1j} m_{2j} = 0$$

Angulo entre matrices:

$$\cos \theta = \frac{\sum m'_{1j} m_{2j}}{[\sum m'_{1j} m_{1j}]^{1/2} [\sum m_{2j} m_{2j}]^{1/2}}$$

c) **Espacio Vectorial:** $(F, +, \lambda)$, $\lambda \in \mathbb{R}$ donde F es el conjunto de funciones de $\mathbb{R}$ en $\mathbb{R}$ tales que

$$\int_{-\infty}^{\infty} [f(x)]^2 dx < \infty \text{ para cualquier } f(x) \in F.$$

$$\text{Producto Interior : } \langle f, g \rangle = \int_{-\infty}^{\infty} f(x) g(x) dx$$

Criterio de ortogonalidad:

$$\int_{-\infty}^{\infty} f(x) g(x) dx = 0$$

Angulo entre funciones

$$\cos \theta = \frac{\int_{-\infty}^{\infty} f(x) g(x) dx}{[\int_{-\infty}^{\infty} [f(x)]^2 dx]^{1/2} [\int_{-\infty}^{\infty} [g(x)]^2 dx]^{1/2}}$$

$$\text{Nótese que } \|f\| = \langle f, f \rangle^{1/2} = \left[\int_{-\infty}^{\infty} [f(x)]^2 dx \right]^{1/2}$$

$$= \left[\int_{-\infty}^{\infty} [f(x)]^2 dx \right]^{1/2}$$

d) **Espacio Vectorial:** $(A, +, \lambda)$, $\lambda \in \mathbb{R}$ donde A es el conjunto de variables aleatorias continuas con espacio muestral común tales que $E(X^2) < \infty$ para cualquier $X \in A$. Aquí $E(\cdot)$ es el operador valor esperado.

Producto interior : $\langle X, Y \rangle = E[XY]$

DEFINICION 8 (Proyección Ortogonal)

Sea L un subespacio de V y sea $z \in V$. La proyección ortogonal de z sobre L es un vector $P_L(z)$ tal que:

- i) $P_L(z) \in L$
- ii) $\langle z - P_L(z), x \rangle = 0, \forall x \in L$

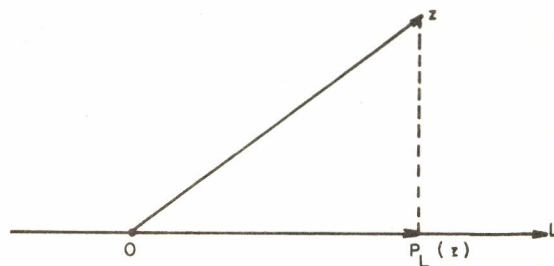


Figura No. 1

Nota

A la función $P_L(\cdot)$ de V en L , de la definición anterior se le llama **proyector ortogonal** sobre L .

Propiedades

1. $z \in L \Leftrightarrow P_L(z) = z$
2. z es ortogonal a $L \Leftrightarrow P_L(z) = 0$
3. $\|z\|^2 = \|P_L(z)\|^2 + \|z - P_L(z)\|^2$
4. Para cada $z \in V$, su proyección ortogonal sobre L es única.

5. De todos los vectores en L , el más cercano (mínima distancia) a un vector $z \in V$, es su proyección ortogonal sobre L , $P_L(z)$.

6. La función proyector ortogonal es un operador lineal, es decir $P(ay + z) = a P_L(y) + P_L(z)$

Nota

La función norma y distancia mencionadas, corresponden a las inducidas por el producto interior definido en el vectorial V .

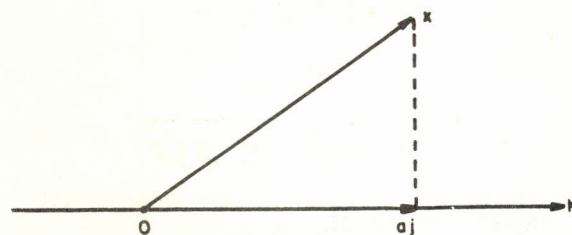


Figura No. 2

ALGUNOS PROBLEMAS ESTADÍSTICOS ENFOCADOS GEOMETRICAMENTE

I. Situación Muestral

Se dispone de la realización de una muestra aleatoria de tamaño n , sobre una variable X :

$$\mathbf{x} = (x_1, x_2, \dots, x_n).$$

espacio vectorial : $(\mathbb{R}^n, +, \lambda), \lambda \in \mathbb{R}$

$$\text{producto interior: } \langle \mathbf{x}, \mathbf{y} \rangle = \mathbf{x}' \mathbf{y} = \sum_{i=1}^n x_i y_i$$

DEFINICION 9. Subespacio de las constantes es:

$$K = \{ \mathbf{c} \in \mathbb{R}^n / \mathbf{c} = (c, c, \dots, c)', c \in \mathbb{R} \}$$

Nota

Los elementos del subespacio K son de la forma

$$\mathbf{c} = c (1, 1, \dots, 1)' = c \mathbf{J}$$

donde $\mathbf{J} = (1, 1, \dots, 1)'$ por lo tanto $\{ \mathbf{J} \}$ es una base para K

I.1. Problema estadístico: Encontrar un valor central (una constante) que sintetice la información sobre la muestra : $\mathbf{x} = (x_1, x_2, \dots, x_n)'$

Problema geométrico: ¿Cuál es el vector en el subespacio K de las constantes, que se encuentra más "cerca" de $\mathbf{x} = (x_1, x_2, \dots, x_n)'$?

Solución: La proyección ortogonal de $\mathbf{x}$ sobre el subespacio K (propiedad 5 de la proyección ortogonal).

así que el valor $P_K(\mathbf{x})$ debe ser tal que:

i) $P_K(\mathbf{x}) \in K$, es decir, $P_K(\mathbf{x}) = a \mathbf{J}$ con $a \in \mathbb{R}$.

ii) $\langle \mathbf{x} - a \mathbf{J}, \mathbf{J} \rangle = 0$, es decir,

$$(\mathbf{x} - a \mathbf{J})' \mathbf{J} = 0,$$

$$\rightarrow \mathbf{x}' \mathbf{J} - a \mathbf{J}' \mathbf{J} = 0$$

$$\rightarrow \sum_{i=1}^n x_i - a n = 0$$

$$\rightarrow a = \bar{x}, \text{ La media aritmética muestral}$$

así $a \mathbf{J} = (\bar{x}, \bar{x}, \dots, \bar{x})$ es el vector de constantes más próximo a $\mathbf{x} = (x_1, x_2, \dots, x_n)$.

Lo cual significa que, bajo la función de distancia definida, la media aritmética es la mejor medida para representar el conjunto de datos.

I.2. Problema estadístico: ¿Cuál es la "bondad" de

$\bar{x} (1, 1, \dots, 1)$ para representar la muestra $\mathbf{x} = (x_1, x_2, \dots, x_n)'$?

Problema geométrico. ¿Cómo medir la "cercanía" entre los vectores

$$\bar{x} (1, 1, \dots, 1) \text{ y } \mathbf{x} = (x_1, x_2, \dots, x_n) ?$$

Solución 1: A través de la distancia inducida por el producto interior definido, así:

$$d(\bar{x} \mathbf{J}, \mathbf{x}) = \|\bar{x} \mathbf{J} - \mathbf{x}\| =$$

$$= \langle \mathbf{x} - \bar{x} \mathbf{J}, \mathbf{x} - \bar{x} \mathbf{J} \rangle^{1/2}$$

$$= \left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^{1/2} = \sqrt{n-1} S$$

donde S es la desviación estándar muestral.

Solución 2: La "cercanía" puede medirse a través del ángulo que forma los vectores

$\bar{x}$ J y x o lo que es un espacio igual al ángulo entre los vectores J y x , así:

$$\begin{aligned} \cos^2 \theta &= \frac{\langle x, J \rangle^2}{\langle x, x \rangle \langle J, J \rangle} \\ &= \frac{\left[\sum_{i=1}^n x_i \right]^2}{\left[\sum_{i=1}^n x_i^2 \right] n} \\ &= \frac{(\bar{x})^2}{\bar{x}^2} \\ &= \frac{\text{promedio al cuadrado}}{\text{promedio de los cuadrados}} \end{aligned}$$

Nótese que si $x_i = \bar{x}$, para todo i , resulta que $\cos^2 \theta = 1$, es decir $\theta = 0$,

Lo cual significa que los vectores están superpuestos; caso en el cual la media identifica totalmente la muestra. En general, cuanto más cerca estén los

x_i de $\bar{x}$, mas cerca estará $(\bar{x})^2$ de $\bar{x}^2$, es decir, más cerca de 1 estará $\cos^2 \theta$.

- 1.3. Problema estadístico:** ¿Cuál es la correlación lineal entre x y y en el sentido de cuán "cerca" están las variables X, Y de poderse expresar como $Y = \beta X + K$?

Problema geométrico. Nótese que:

$y = \beta x + c$ J $\Leftrightarrow x, y, J$ son linealmente dependientes $\Leftrightarrow x, y, J$ son vectores que pertenecen a un mismo plano $\Leftrightarrow$ el subespacio generado por $\{x, J\}$ coinciden con el subespacio generado por $\{y, J\}$.

Así el problema geométrico puede plantearse como: ¿Qué tan "cerca" está el subespacio generado por $\{x, J\}$ del generado por $\{y, J\}$

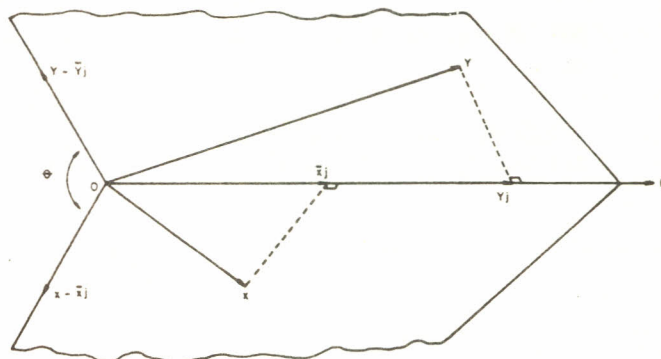


Figura No. 3

Solucion: Una medida de la "cercanía" de los espacios en cuestión se logra a través del ángulo entre los vectores

$x - \bar{x}J$ y $y - \bar{y}J$.

así:

$$\begin{aligned} \cos^2 \theta &= \frac{\langle x - \bar{x}J, y - \bar{y}J \rangle}{[\langle x - \bar{x}J, x - \bar{x}J \rangle \langle y - \bar{y}J, y - \bar{y}J \rangle]^{1/2}} \\ &= \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{[\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2]^{1/2}} \\ &= \text{coeficiente de correlación lineal} \end{aligned}$$

Lo cual significa que un indicador del grado de relación lineal entre dos variables en una muestra, es el coeficiente de correlación lineal.

- 1.4. Problema estadístico:** ¿Cuál es el "mejor" modelo lineal que permite predecir la variable Y en términos de las variables $x_1, x_2, \dots, x_p$, de acuerdo con la información que suministra la muestra? suponemos que los vectores $x_1, x_2, \dots, x_p$ corresponden a las n observaciones de cada variable, respectivamente.

Problema geométrico: Se desea encontrar las constantes $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ que determinan la combinación lineal de los vectores $x_1, x_2, \dots, x_p$ digamos:

$\hat{y} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$, tal que $\hat{y}$ es el vector más cerca de y .

así el problema geométrico puede plantearse como:

¿Cuál es el vector $\hat{y}$ en el espacio $L = L(j, x_1, \dots, x_p)$ que está a menor distancia del vector y ?

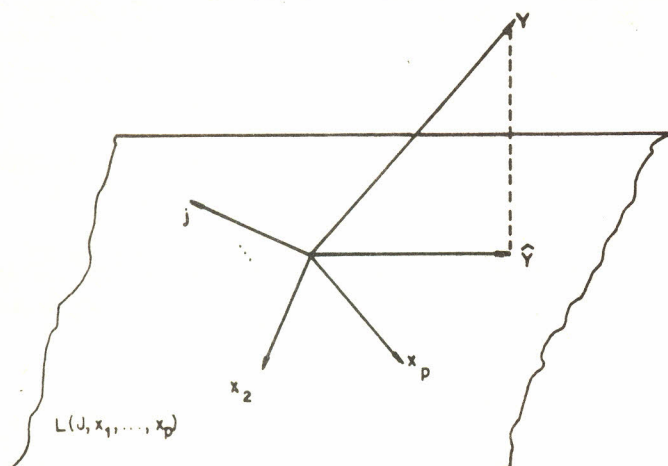


Figura No. 4

Solución: el vector $\hat{y} \in L$, cuya distancia al vector y es mínima es el vector ortogonal de y sobre el subespacio $L = L(j, x_1, \dots, x_p)$

nótese que en forma general $\hat{y}$ puede expresarse

como $\hat{y} = X\hat{\beta}$, donde $X = [j \ x_1 \ \dots \ x_p]$ y $\hat{\beta} = (\beta_0 \ \beta_1 \ \dots \ \beta_p)'$.

La intención ahora es tratar de caracterizar el vector buscado $\hat{y}$. Usando la segunda condición de proyección ortogonal debe suceder que:

$\langle y - \hat{y}, z \rangle = 0$ " Para todo $z \in L$, es decir
 $\langle y - X\hat{\beta}, z \rangle = 0$ " Para todo $z \in L$, lo cual es equivalente con:

$\langle y - X\hat{\beta}, j \rangle = 0; \langle y - X\hat{\beta}, x_1 \rangle = 0; \dots;$
 $\langle y - X\hat{\beta}, x_p \rangle = 0$

Estas ecuaciones pueden reescribirse como:

$J'(y - X\hat{\beta}) = 0; \quad x_1'(y - X\hat{\beta}) = 0$
 $\vdots; \quad x_p'(y - X\hat{\beta}) = 0$

resultando en forma matricial:

$$\begin{pmatrix} J' \\ X_1' \\ \vdots \\ X_p' \end{pmatrix} (y - X\hat{\beta}) = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix} \Leftrightarrow X'(y - X\hat{\beta}) = 0$$

es decir:

$$X'X\hat{\beta} = X'y,$$

que corresponden a las llamadas **ecuaciones normales**.

Es muy importante resaltar que las ecuaciones normales obtenidas de esta manera contienen más información que las obtenidas en forma algebraica, puesto que:

- i) La proyección sobre L siempre existe entonces $X\hat{\beta}$ existe y por lo tanto se sabe que las ecuaciones normales son consistentes.

(Aquí $\hat{\beta}$, es una solución).

- ii) La proyección de Y sobre L es única, por lo tanto

el vector $X\hat{\beta}$ que da solución a las ecuaciones normales es único. (Invariante con respecto a los posibles distintos valores de $\hat{\beta}$ que son solución de las ecuaciones normales).

- iii) La proyección de Y sobre L es el vector en L de mínima distancia a Y , entonces, el vector

$$\hat{y} = X\hat{\beta}$$

que satisface las ecuaciones normales es óptima desde el punto de vista de la distancia inducida por el producto interior definido sobre $\mathbb{R}^n$.

Cuando se obtienen las ecuaciones normales es forma algebraica, solo se sabe que si existe solución a dicho sistema, estas corresponde solo a puntos críticos; es decir que para verificar la consistencia del sistema y las condiciones de optimalidad de sus soluciones es necesario realizar procedimientos adicionales.

I.5. Problema Estadístico : Con respecto al problema anterior, ¿ Qué tan "bueno" es el modelo ajustado?

Problema geométrico : ¿ Qué tan próximo está

el vector $\hat{y}$ al vector y ?

Solución 1: La primera idea que surge es a tra-

vés de la distancia entre $\hat{y}$ y y , por ejemplo

$$\begin{aligned}\|\hat{y} - y\|^2 &= (\hat{y} - y)'(\hat{y} - y) = \sum_{j=1}^n (\hat{y}_j - y_j)^2 \\ &= (n-p) \text{ CME},\end{aligned}$$

donde CME es el cuadrado medio del error, el cual es un indicador de la calidad (bondad) del modelo ajustado.

Solución 2: La proximidad entre $\hat{y}$ y y puede medirse a través del ángulo entre los subespacios

$L(y, j)$, y $L(\hat{y}, j)$ con el mismo criterio que en el problema I.3, como se muestra en la figura No. 5.

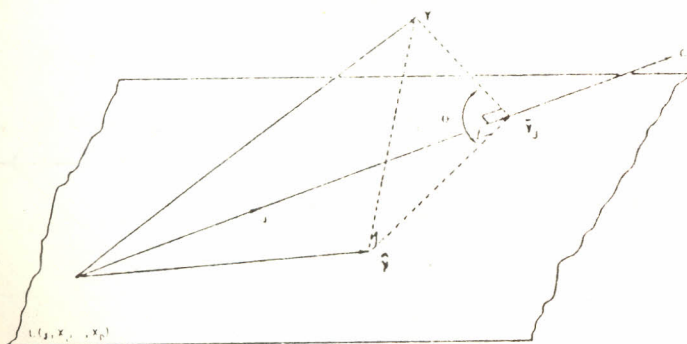


Figura No. 5

De esta manera :

$$\cos \theta = \frac{\langle y - \bar{y}j, \hat{y} - \bar{\hat{y}}j \rangle}{\|y - \bar{y}j\| \|\hat{y} - \bar{\hat{y}}j\|}$$

$$= \frac{\sum (y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\|y - \bar{y}j\| \|\hat{y} - \bar{\hat{y}}j\|}$$

nótese que: $(y - \bar{y}j)'j = 0$, por lo tanto:

$$y'j = \hat{y}'j \Leftrightarrow \bar{y} = \bar{\hat{y}}, \text{ así que}$$

$$\begin{aligned}\cos \theta &= \frac{\sum (y_j - \bar{y})(\hat{y}_j - \bar{\hat{y}})}{[\sum (y_j - \bar{y})^2]^{1/2} [\sum (\hat{y}_j - \bar{\hat{y}})^2]^{1/2}} = r_{y, \hat{y}} \\ &= r_{y, x_1 x_2 \dots x_p}\end{aligned}$$

coeficiente de correlación múltiple muestral.

II. SITUACION POBLACIONAL (VARIABLES ALEATORIAS)

ESPACIO VECTORIAL $\langle A, +, \lambda \rangle$ donde A es un conjunto de variables aleatorias continuas con dominio común Ω , tales que $E[X^2] < \infty$ y producto interior: $\langle X, Y \rangle = E[XY]$.

DEFINICION 10.

El subespacio de las **variables aleatorias constantes**, se define como:

$$C = \{X \in A / P[\{w \in \Omega / X(w) = c, c \in R\}] = 1\}$$

es decir el conjunto de variables aleatorias tales que $X = c$ con probabilidad uno, para alguna constante. Note que una base para el subespacio C es $\{X=1\}$.

II.1. Problema estadístico : ¿Cuál es la constante que "mejor" sintetiza la información sobre una variable aleatoria continua X ?

Problema geométrico : ¿Cuál es el vector del espacio C de las constantes "mas" cerca del vector X ?

Solución: el vector proyección de X sobre el subespacio C , $P_C(x)$. Así que

$$i) P_C(x) \in C \rightarrow P_C(x) = \lambda \text{ (constante)}$$

$$ii) \langle X - \lambda, K \rangle = 0 \rightarrow K \in C, \text{ en particular:}$$

$$\langle X - \lambda, 1 \rangle = 0 \rightarrow E[(X - \lambda) 1] = 0 \rightarrow$$

$$E[X - \lambda] = 0 \rightarrow \lambda = E[X] \text{ media poblacional}$$

Con base en la distancia inducida por el producto interior definida en el espacio de variables aleatorias, puede decirse que la media poblacional es el valor que mejor sintetiza la información sobre una variable aleatoria.

II.2. Problema estadístico : ¿Cuál es la "bondad" de $\lambda = E(x)$ para representar a X ?

Problema geométrico : ¿Cómo medir la "cercanía" del vector λ al vector X ?

Solución 1: a través de la distancia entre $\lambda = E(X)$ y X .

$$d(X, \lambda) = |X - \lambda| = (\langle X - \lambda, X - \lambda \rangle)^{1/2}$$

$$= [E|(X - \lambda)^2|]^{1/2}$$

$$= [E(X - E(x))^2]^{1/2}$$

$$= \sigma_X \text{ desviación estándar poblacional}$$

Es decir que una medida de la bondad de la media para sintetizar la información sobre una variable aleatoria, es su desviación estándar, la cual representa la distancia entre X y $E(X)$.

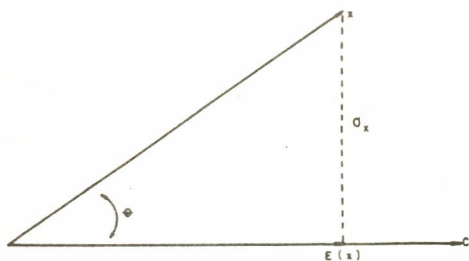


Figura No. 6

Solución 2: Otra forma de indicar la "cercanía" del vector X al vector $E(X)$, es a través del ángulo entre ellos, que es igual al ángulo entre los vectores X y 1 .

$$\cos \theta = \frac{\langle X, 1 \rangle}{[\langle X, X \rangle \langle 1, 1 \rangle]^{1/2}}$$

$$\cos \theta = \frac{E(X)}{[E(X^2)]^{1/2}}$$

$$\text{así } \cos \theta = \frac{E(X)}{[E(X^2)]^{1/2}}$$

nótese que si $X = C$, entonces: $E(X) = C$ y $E(X^2) = C^2$, por lo tanto $|\cos \theta| = 1$.

Caso en el cual la media $E(X)$ da toda la información sobre X .

Note que

$$\tan \theta = \frac{|X - E(x)|}{|E(x)|}$$

$$= \frac{\sigma_X}{[E(x)]}$$

el cual corresponde al comunmente usado **coeficiente de variación**.

II.3. Problema estadístico : ¿Cuál es la medida de la correlación lineal entre las variables aleatorias X e Y ?, es decir en qué medida las variables aleatorias X e Y se acercan a la relación $Y = aX + C$?

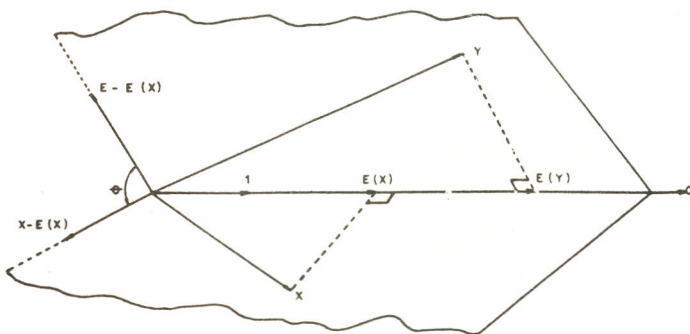


Figura No. 7

Problema geométrico: Note que como $\{1\}$ es una base del espacio de las variables aleatorias constantes, entonces $C = C1$. Así el problema geométrico podría plantearse como ¿Qué tan próximos están X , Y , y 1 de estar contenidos por un subespacio de dimensión dos (2), ó ¿Qué tan "próximos" están los subespacios $L(X, 1)$ y $L(Y, 1)$?

Solución : El ángulo θ entre dos vectores ortogo-

nales a cada subespacio respectivamente, es un indicador del grado de dependencia lineal de los vectores $X, Y, 1$.

El mencionado ángulo θ es equivalente al formado por los vectores $X - E(X)$ y $Y - E(Y)$, así que:

$$\begin{aligned} \cos \theta &= \frac{\langle Y - E(X), Y - E(Y) \rangle}{|X - E(X)| |Y - E(Y)|} \\ &= \frac{E\{[Y - E(X)][Y - E(Y)]\}}{\sigma_X \sigma_Y} \\ &= \frac{\text{COV}(X, Y)}{\sigma_X \sigma_Y} \end{aligned}$$

Donde $\text{COV}(X, Y)$ es la covarianza entre las variables aleatorias X y Y .

$$= \rho_{xy} \quad \text{coeficiente de correlación lineal poblacional.}$$

El cual es una medida de la dependencia lineal de los vectores $X, Y, 1$.

II.4. Problema estadístico: ¿Cuál es la mejor función lineal que relaciona la variable aleatoria Y con un conjunto de variables aleatorias $1, X_1, X_2, \dots, X_p$?

Problema geométrico: ¿Cuál es el vector

$\hat{Y}$ en $L(1, X_1, \dots, X_p)$ que está más "cerca" del vector Y ?

Solución: La proyección ortogonal de Y sobre el subespacio $L(1, X_1, \dots, X_p)$.

Puede notarse que dicha proyección es

$\hat{Y} = E[Y / 1, X_1, \dots, X_p]$, esperanza condicional, siempre y cuando resulte ser una función lineal de $1, X_1, \dots, X_p$, es decir:

$$\hat{Y} \in L(1, \dots, X_p).$$

Esto equivale a suponer que $y = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \varepsilon$, para alguna variable aleatoria ε tal que $E[\varepsilon] = 0$ y además las X_j están no correlacionadas con ε .

Veamos que se cumple la condición ii) de la proyección ortogonal:

$\langle \hat{Y} - Y, Z \rangle = 0$ " Para todo $Z \in L(1, X_1, \dots, X_p)$, veamos:

$$\begin{aligned} \langle Y - \hat{Y}, Z \rangle &= \langle Y, Z \rangle - \langle \hat{Y}, Z \rangle = \\ &= E[YZ] - E[\hat{Y}Z] \\ &= E[YZ] - E[E(Y / 1, X_1, \dots, X_p) Z] \\ &= E[YZ] - E[E(YZ / 1, X_1, \dots, X_p)] \\ &\quad \text{puesto que } Z \in L(1, X_1, \dots, X_p) \\ &= E[YZ] - E[YZ] = 0 \quad \text{" Para todo } Z \in L. \end{aligned}$$

Así, con respecto a la distancia inducida por el producto interior definido en el espacio de variables aleatorias, la media condicional de Y , constituye el mejor modelo lineal para el propósito buscado. (Ver Figura No. 8)

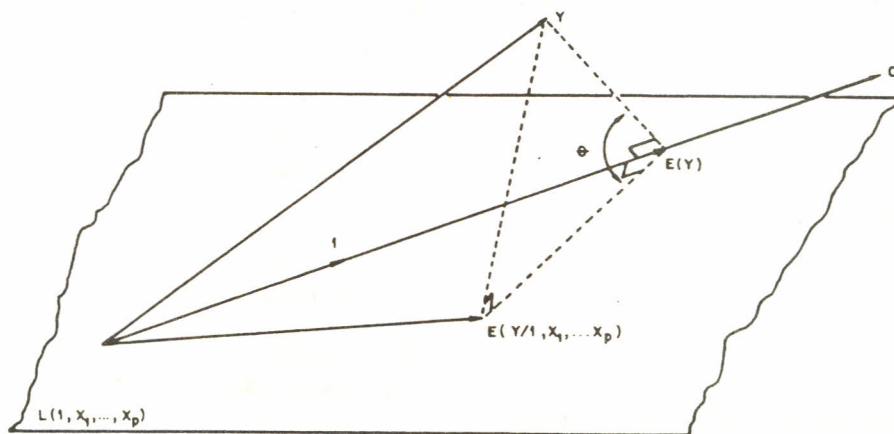


Figura No. 8

II.5 Problema estadístico :

¿Qué tan buena es $\hat{Y} = E[Y/1, X_1, \dots, X_p]$ para predecir a Y?

Problema geométrico:

¿Qué tan "cerca" está el vector $\hat{Y}$ del vector Y?

Solución 1 : Puede medirse a través de la distancia

$d(\hat{Y}, Y) :$

$$d(\hat{Y}, Y) = |Y - \hat{Y}| = \sqrt{\langle Y - \hat{Y}, Y - \hat{Y} \rangle}^{1/2}$$

$$= \{E(Y - E(Y/1, X_1, \dots, X_p))^2\}^{1/2}$$

$$= \sigma_{Y/1, X_1, \dots, X_p} \text{ desviación estándar condicional poblacional}$$

Por lo tanto, la desviación estándar condicional, es un buen indicador de la bondad de la media condicional para "predecir" Y.

Solución 2: La "cercanía" entre Y y $\hat{Y}$ puede medirse a través del ángulo θ que forma el subespacio

$L(Y, 1)$ con $L(\hat{Y}, 1)$, el cual es equivalente con el ángulo entre los vectores $Y - E(Y)$, y

$\hat{Y} - E(\hat{Y})$, es decir

$$\cos \theta = \frac{\langle Y - E(Y), \hat{Y} - E(\hat{Y}) \rangle}{|Y - E(Y)| |\hat{Y} - E(\hat{Y})|} = \rho_{Y\hat{Y}}$$

$$= \frac{\langle Y - E(Y), E(Y/1, X_1, \dots, X_p) - E(Y) \rangle}{|Y - E(Y)| |E(Y/1, X_1, \dots, X_p) - E(Y)|}$$

puesto que $E[\hat{Y}] = E[Y]$

$$= \frac{\text{COV}[Y, E(Y/1, X_1, \dots, X_p)]}{\sigma_Y \sigma_{Y/1, X_1, \dots, X_p}}$$

$$= \rho_{Y, 1, X_1, \dots, X_p}$$

así que $\cos \theta$ representa el coeficiente de correlación entre Y y $\hat{Y}$, o lo que es igual es el coeficiente de correlación múltiple entre Y y 1, $X_1, \dots, X_p$.

CONCLUSIONES.

Aunque no se pretende haber agotado el tema de del enfoque geométrico de la estadística, si se puede notar en los elementos presentados, cómo a partir de la definición de algunos pocos conceptos de geometría, pueden homologarse problemas estadísticos con problemas geométricos, los cuales tienen soluciones naturales desde el punto de vista geométrico y heredan una riqueza interpretativa, como por ejemplo que al medir una característica a través de un ángulo entre vectores no influye la norma de los vectores, lo cual es equivalente a un proceso de estandarización estadística.

Se hace evidente, en el tratamiento de estos elementales problemas estadísticos, las ventajas del enfoque geométrico.

BIBLIOGRAFIA.

1. BESLEY, WELSH, KUH, (1980). *Regression Diagnostics*. John Wiley. New York.
2. BRAMBILA F; BEHAR R. "Geometría y Estadística". Centro de Estadística y Cálculo, Colegio de Postgraduados, Chapingo, México, 1986. (Inédito).
3. BRYANT, Peter (1984). "Geometary Statistics, Probability: Variations a common theme". The American statistician, February 1984, Vol. 38 No. 1.
4. DRAPER, SMITH. (1981). *Applied Regression Analysis*. New York. John Wiley.
5. GUNST R.; MASON R. (1980). *Regression Analysis and its Application*. New York: Marcel Dekker INC.
6. HABERMAN, SHELBY J. (1978). *Analysis of Qualitative Data*. New York: Academic Press.
7. HERR, David G. (1980). "On the History of the use of Geometry in the General Linear Model". The American Statistician, February 1980, Vol. 34 No. 1.
8. MARGOLIS, Marvin S. (1979). "Perpendicular Projections and Elementary Statistics". The American Statistician, August 1979, Vol. 33 No. 3.
9. SEBER, G.A.F. (1977). *Linear Regression Analysis*. New York: John Wiley.
10. TAKEUCHI, K; YANAI H.; MUKHERJEE B.N. (1982). *The Foundations of Multivariate Analysis*. New York: John Wiley.