REGRESIÓN ORTOGONAL Y COMPONENTES PRINCIPALES

J. ALBERTO MARTÍNEZ ARNAIZ*

Escuela de Empresariales. Bilbao

In this work the Principal Components Analysis is presented, starting from the orthogonal regression plane. On this basis, the data reduction technique is exposed in the three-dimensional case. Finally, the correlation matrix analysis is considerated, as well as its extension to p dimensions.

INTRODUCCIÓN

Parece que el objetivo final al que debe apuntar la enseñanza de la Estadística en una Escuela de Empresariales se podría formular como sigue: capacitación al futuro diplomado para realizar, mediante un Ordenador y el software estadístico adecuado, el análisis de los datos reales existentes en su empresa y en el ámbito económico en el que ésta se encuentra inmersa.

Si se entra en el detalle de objetivos concretos, se daría un consenso amplio en torno a la utilidad del Análisis de Componentes Principales mediante el programa SPAD. En relación con el dilema en torno a si se opta por una presentación rigurosa de tal técnica, o bien se prefiere ofrecer al alumno los conocimientos precisos para saber interpretar las "salidas" que suministra el Ordenador, cabría preguntarse, ¿existe alguna opción intermedia?

Este trabajo pretende defender precisamente una posible solución de compromiso, consistente en:

- exponer al alumno, ya familiarizado con la regresión convencional, el plano de regresión ortogonal.

*J. Alberto Martínez Arnaiz. Doctor en Ciencias Económicas. Profesor de Estadística. Escuela de Empresariales. Bilbao.

- a continuación, presentar la técnica de reducción de rango en el caso tridimensional.

- por último, tratar el caso del análisis de la matriz de correlación, así como su extensión a p dimensiones.

2. PLANO DE REGRESIÓN ORTOGONAL Y VARIANZA RESIDUAL

La regresión convencional exige la separación de las variables en dos clases: la variable a explicar por un lado, y los regresores por otro. La regresión ortogonal, en cambio, estudia el conjunto de variables en un solo bloque: todas las variables son a la vez explicativas y explicadas.

El objetivo de la regresión ortogonal (en $\mathbb{R}^3$) consiste en ajustar un plano a la nube de puntos. El criterio utilizado es el mínimo cuadrático; la diferencia con la regresión convencional radica en que el error de regresión ya no se define como diferencia entre valor predicho y valor observado.

Sea (x, y, z) una variable tridimensional, y $M_0(x_0, y_0, z_0)$ una de las observaciones de la misma. Sea, además,

$$\alpha x + \beta y + \gamma z + \delta = 0$$

la ecuación del plano de regresión ortogonal que deseamos obtener. Impondremos al plano la restricción:

$$\alpha^2 + \beta^2 + \gamma^2 = 1$$

Pues bien, el error de regresión se define aquí como la distancia del punto observado al plano:

$$e_0 = |\alpha x_0 + \beta y_0 + \gamma z_0 + \delta|$$

El criterio de obtención del plano se establecerá del siguiente modo : "el plano de regresión ortogonal es aquel cuya media de cuadrados de distancias a los puntos observados sea mínima".

Si suponemos que hay m observaciones, la función a minimizar será:

$$\frac{1}{m}\sum_{i=1}^{m}(e_i)^2 = \frac{1}{m}\sum(\alpha x_i + \beta y_i + \gamma z_i + \delta)^2, \quad i = (1, 2, \ldots, m)$$

función que al incluir la restricción impuesta se transforma en:

$$L = \frac{1}{m}\sum(\alpha x_i + \beta y_i + \gamma z_i + \delta)^2 - \mu(\alpha^2 + \beta^2 + \gamma^2 - 1)$$

que, a su vez, derivando con respecto a δ da lugar a la condicion necesaria:

$$\frac{1}{m}\sum 2(\alpha x_i + \beta y_i + \gamma z_i + \delta) = 0 \Rightarrow$$
$$\frac{1}{m}\sum(\alpha x_i + \beta y_i + \gamma z_i + \delta) = 0 \Rightarrow$$
$$\alpha\overline{x} + \beta\overline{y} + \gamma\overline{z} + \delta = 0$$

Esta última expresión prueba que el plano buscado pasa por el centro de gravedad de la nube:

$$(\overline{x}, \overline{y}, \overline{z})$$

De esta primera conclusión extraemos una recomendación operativa:

- en primer lugar, debemos centrar las variables (restar de cada una de ellas su media)
- en segundo lugar, obtendremos el plano de regresión ortogonal que pasa por el origen de coordenadas.

El resultado que acabamos de obtener nos invita a replantear ligeramente nuestro problema teórico. Supondremos que operamos sobre variables centradas, y nuestro objetivo consistirá en hallar el plano de regresión ortogonal que pase por el origen de coordenadas.

Sean, (x, y, z) una variable centrada tridimensional, y $M_0(x_0, y_0, z_0)$ una de las observaciones de la misma. Sea, además,

$$\alpha x + \beta y + \gamma z = 0$$

la ecuación del plano de regresión ortogonal que deseamos obtener. Impondremos al plano la restricción:

$$\alpha^2 + \beta^2 + \gamma^2 = 1$$

El error de regresión se define como la distancia del punto observado al plano:

$$e_0 = |\alpha x_0 + \beta y_0 + \gamma z_0|$$

La función a minimizar será:

$$\frac{1}{m}\sum_{i=1}^{m}(e_i)^2 = \frac{1}{m}\sum_{i=1}^{m}(\alpha x_i + \beta y_i + \gamma z_i)^2, \quad i = 1, 2, \ldots, m)$$

función que al incluir la restricción impuesta se transforma en:

$$L = \frac{1}{m}\sum_{i=1}^{m}(\alpha x_i + \beta y_i + \gamma z_i)^2 - \mu(\alpha^2 + \beta^2 + \gamma^2 - 1)$$

que, a su vez, da lugar a las condiciones necesarias de mínimo:

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}2(\alpha x_i + \beta y_i + \gamma z_i)x_i - 2\mu\alpha &= 0 \\
\frac{1}{m}\sum_{i=1}^{m}2(\alpha x_i + \beta y_i + \gamma z_i)y_i - 2\mu\beta &= 0 \\
\frac{1}{m}\sum_{i=1}^{m}2(\alpha x_i + \beta y_i + \gamma z_i)z_i - 2\mu\gamma &= 0 \\
-(\alpha^2 + \beta^2 + \gamma^2 - 1) &= 0
\end{aligned}$$

Las tres primeras ecuaciones son equivalentes a:

$$\begin{aligned}
\alpha(s_x)^2 + \beta s_{xy} + \gamma s_{xz} &= \mu\alpha \\
\alpha(s_{xy})^2 + \beta(s_y)^2 + \gamma s_{yz} &= \mu\beta \\
\alpha(s_{xz})^2 + \beta s_{yz}^2 + \gamma(s_z)^2 &= \mu\gamma
\end{aligned}$$

que, en expresión matricial:

$$\begin{bmatrix} s_x^2 & s_{xy} & s_{xz} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{xz} & s_{yz} & s_z^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} \mu$$

nos indican que $(\alpha\ \beta\ \gamma)$ forman un vector propio de la matriz de covarianzas. ¿Cuál de los tres ? Para resolver esta duda es preciso volver a la función media de cuadrados de los errores:

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}(e_i)^2 &= \frac{1}{m}\sum_{i=1}^{m}(\alpha x_i + \beta y_i + \gamma z_i)^2 = \\
&= \alpha^2(s_x)^2 + \beta^2(s_y)^2 + \gamma^2(s_z)^2 + 2\alpha\beta s_{xy} + 2\alpha\gamma s_{xz} + 2\beta\gamma s_{yz} = \\
&= [\alpha \quad \beta \quad \gamma] \begin{bmatrix} s_x^2 & s_{xy} & s_{xz} \\ s_{xy} & s_y^2 & s_{yz} \\ s_{xz} & s_{yz} & s_z^2 \end{bmatrix} \begin{bmatrix} \alpha \\ \beta \\ \gamma \end{bmatrix} = \mu
\end{aligned}$$

Ahora bien, si la función que se quiere minimizar equivale a μ, de los tres valores propios posibles $\mu_1\ \mu_2\ \mu_3$ debemos seleccionar el mínimo. Convengamos en que este sea μ_3. En este punto podemos enunciar la solución al problema que nos

hemos planteado: el plano (que pase por el origen de coordenadas) de regresión ortogonal que mejor se ajusta a una nube de puntos (cuyo centro de gravedad coincida con el origen) tiene por coeficientes los elementos del vector propio de la matriz de covarianzas ($\alpha 3\ \beta_3\ \gamma_3$) correspondientes al valor propio mínimo μ_3.

3. COMBINACIÓN LINEAL ÓPTIMA DE (x, y, z). VARIANZA EXPLICADA

Se ha dicho que la regresión ortogonal es un instrumento del análisis de interdependencias. La utilización de esta técnica se realiza en forma de reducción de rango: se trata de sustituir las tres variables (x, y, z) por dos, siendo éstas combinación lineal de aquellas tres.

El vector $(\alpha_3 \quad \beta_3 \quad \gamma_3)$ nos informa de la dirección perpendicular al plano de regresión ortogonal. Los otros dos vectores propios de la matriz de covarianzas (simétrica real), correspondientes a los valores propios $\mu_1 \quad \mu_2$ (supondremos $\mu_1 > \mu_2 > \mu_3$),

$$(\alpha_1 \quad \beta_1 \quad \gamma_1) \qquad (\alpha_2 \quad \beta_2 \quad \gamma_2)$$

conforman con aquel una matriz ortogonal $\mathbf{P}$. Esta matriz $\mathbf{P}$ permite pasar a un nuevo sistema de ejes coordenados, mediante una rotación que pivota sobre el origen. Las nuevas coordenadas se obtendrán del siguiente modo:

$$\begin{aligned} F_1 &= \alpha_1 x + \beta_1 y + \gamma_1 z \\ F_2 &= \alpha_2 x + \beta_2 y + \gamma_2 z \\ e &= \alpha_3 x + \beta_3 y + \gamma_3 z \end{aligned}$$

Esta tercera coordenada mide la distancia de los puntos de la nube al plano de regresión ortogonal. Las dos primeras coordenadas corresponden a las proyecciones de los puntos de la nube sobre el citado plano.

Pues bien, nuestra técnica de reducción de rango propone precisamente a F_1, F_2 como combinaciones lineales simplificadoras de la variable tridimensional, en tanto que la combinación lineal e quedaría como elemento residual.

Teniendo en cuenta que

$$\overline{x} = \overline{y} = \overline{z} = 0$$

se prueba rapidamente que la media de F_1 es cero. Análogamente se tiene que la media de F_2 es cero y que la media de los residuos es igualmente cero.

La varianza de la combinación lineal F_1 (supondremos que todos los vectores propios están normalizados) resulta ser:

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}(F_{1i})^2 &= \frac{1}{m}\sum_{i=1}^{m}(\alpha_1 x_i + \beta_1 y_i + \gamma_1 z_i)^2 = \\
&= (\alpha_1)^2(s_x)^2 + (\beta_1)^2(s_y)^2 + (\gamma_1)^2(s_z)^2 + \\
&+ 2\alpha_1\beta_1 s_{xy} + 2\alpha_1\gamma_1 s_{xz} + 2\beta_1\gamma_1 s_{yz} = \\
&= \begin{bmatrix}\alpha_1 & \beta_1 & \gamma_1\end{bmatrix}\begin{bmatrix} s_x^2 & x_{xy} & x_{xz} \\ s_{xy} & s_y^2 & s_{yz} \\ x_{xz} & s_{yz} & x_z^2 \end{bmatrix}\begin{bmatrix}\alpha_1 \\ \beta_1 \\ \gamma_1\end{bmatrix} = \mu_1
\end{aligned}$$

es decir, el mayor valor propio.

De igual modo se puede establecer que la varianza de la combinación lineal F_2 es el valor propio μ_2; y también que la varianza residual es el valor propio μ_3.

Los valores de F_1 están incorrelados con los residuos. En efecto, la covarianza entre ambas variables será:

$$\begin{aligned}
\frac{1}{m}\sum_{i=1}^{m}F_{1i}e_i &= \frac{1}{m}\sum_{i=1}^{m}(\alpha_1 x_i + \beta_1 y_i + \gamma_1 z_i)(\alpha_3 x_i + \beta_3 y_i + \gamma_3 z_i) = \\
&= \alpha_1\alpha_3(s_x)^2 + \beta_1\beta_3(s_y)^2 + \gamma_1\gamma_3(s_z)^2 + (\alpha_1\beta_3 + \alpha_3\beta_1)s_{xy} + \\
&+ (\alpha_1\gamma_3 + \alpha 3\gamma_1)s_{xz} + (\beta_1\gamma_3 + \beta_3\gamma_1)s_{yz} = \\
&= \begin{bmatrix}\alpha_1 & \beta_1 & \gamma_1\end{bmatrix}\begin{bmatrix} s_x^2 & x_{xy} & x_{xz} \\ s_{xy} & s_y^2 & s_{yz} \\ x_{xz} & s_{yz} & x_z^2 \end{bmatrix}\begin{bmatrix}\alpha_3 \\ \beta_3 \\ \gamma_3\end{bmatrix} = \\
&= \begin{bmatrix}\alpha_1 & \beta_1 & \gamma_1\end{bmatrix}\begin{bmatrix}\alpha_3 \\ \beta_3 \\ \gamma_3\end{bmatrix}\mu_3 = 0
\end{aligned}$$

Del mismo modo se establece la incorrelación entre la variable F_2 y los residuos, o entre las variables F_1 y F_2.

Designaremos combinación lineal óptima a F_1 por ser la de máxima varianza; a F_2 la calificaremos de combinación lineal subóptima; por último, la combinación lineal e se denominará residual.

La parte de varianza explicada por el par (F_1, F_2), es decir, por el plano de regresión ortogonal, se calculará mediante:

$$\frac{\mu_1 + \mu_2}{\mu_1 + \mu_2 + \mu_3}$$

Los planos formados por (F_2, e) o por (F_3, e) también contienen información interesante; la parte de varianza explicada será en cada uno de los casos:

$$\frac{\mu_1 + \mu_3}{\mu_1 + \mu_2 + \mu_3} \qquad \frac{\mu_2 + \mu_3}{\mu_1 + \mu_2 + \mu_3}$$

4. UN EJEMPLO

Consideremos la base de datos[1],

i	x	y	z
1	−16.9680	−7.5500	−15.3420
2	−18.3820	−4.7200	−13.3400
3	−18.3820	−8.4920	−12.0080
.....			
166	16.2610	13.4450	17.0120
167	20.5030	12.0290	13.0100
168	16.2610	12.5020	17.3450
169	20.5030	11.0860	13.3430

conformada por 169 puntos del espacio $\mathbb{R}^3$ que son parte de la estructura de un avión (figura 1).

[1]Los datos están disponibles como fichero "avion.dat" en: anonymous ftp **from** porthos.bio.ub.es **directory**: /pub/multicua.

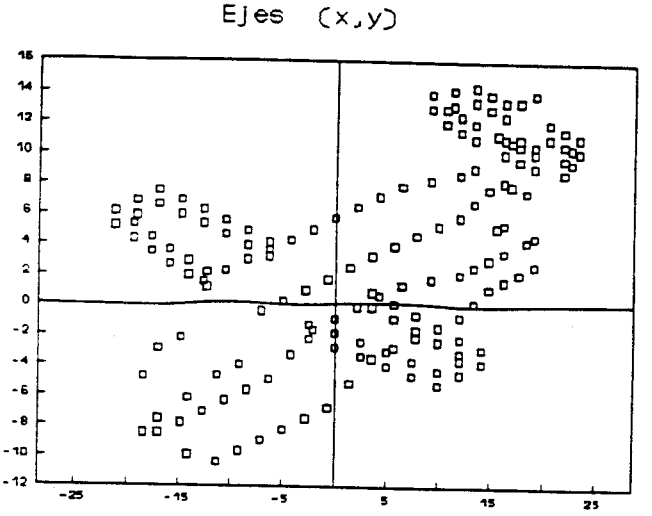

Figura 1

El análisis de regresión ortogonal ofrece los siguientes resultados:

- *vector de medias:*

$$\begin{bmatrix} 3.723 & 3.185 & 2.827 \end{bmatrix}$$

- *matriz de covarianzas:*

$$\begin{bmatrix} 163.7 & 39.4 & 63.5 \\ 39.4 & 40.0 & 59.8 \\ 63.5 & 59.8 & 137.2 \end{bmatrix}$$

- *raíces propias:*

$$\begin{bmatrix} 239.1 & 90.9 & 10.9 \end{bmatrix}$$

- *vectores propios (en filas):*

$$\begin{bmatrix} -.703 & -.329 & -.631 \\ .708 & -.236 & -.666 \\ -.070 & .915 & -.398 \end{bmatrix}$$

– *coordenadas sobre nueva base:*

i	F_1	F_2	e
1	29.533	−.014	−1.133
2	28.334	−3.014	.758
3	28.733	−3.013	−3.223
.....			
167	−21.132	−2.992	2.857
167	−21.125	3.010	2.858
168	−21.033	−2.991	1.862
169	−21.025	3.010	1.863

El plano principal (F_1, F_2), que pone de manifiesto la estructura básica del avión, es decir, fuselaje y alas (figura 2) explica el 96.8% de la varianza, en efecto:

$$\frac{239.1 + 90.9}{239.1 + 90.9 + 10.9} = 0.968$$

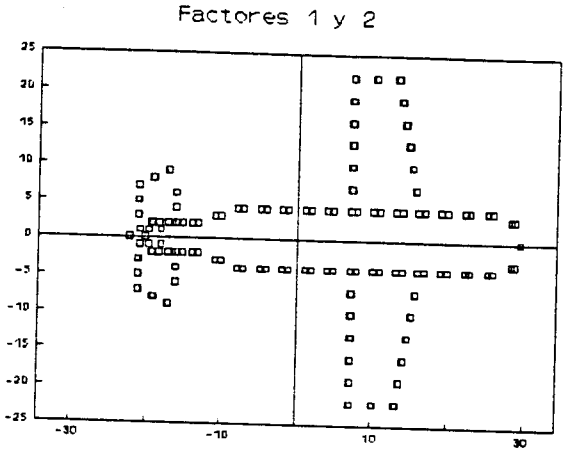

Figura 2

El plano de la figura 3 (F_1, e) muestra el perfil del avión y explica el 73.3% de la varianza.

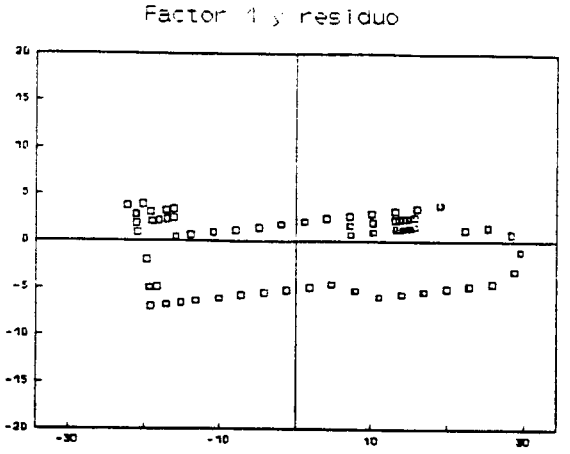

Figura 3

5. REGRESIÓN ORTOGONAL Y ANÁLISIS EN COMPONENTES PRINCIPALES

La técnica expuesta coincide plenamente con el análisis de componentes principales. Conviene, si acaso, aclarar que nuestro trabajo se ha realizado sobre datos centrados y por tanto se ha analizado la matriz de covarianzas, en tanto que el ACP habitual (conocido como normado) se suele realizar por motivos de igualación de escalas, sobre datos tipificados, y consiguientemente se basa en el análisis de la matriz de correlación.

El resumen de operaciones necesarias para realizar un ACP normado sobre una base de datos p–dimensional sería:

- tipificación de los datos: $t_1, t_2, \ldots, t_p$
- diagonalización de la matriz de correlación (valores y vectores propios):

$$\mu_1 > \mu_2 > \mu_3 > \cdots > \mu_p$$

$$
\begin{array}{lcl}
v_1 & : & (\alpha_1 \quad \beta_1 \quad \gamma_1 \quad \cdots\cdots) \\
v_2 & : & (\alpha_2 \quad \beta_2 \quad \gamma_2 \quad \cdots\cdots) \\
v_3 & : & (\alpha_3 \quad \beta_3 \quad \gamma_3 \quad \cdots\cdots) \\
 & & \cdots\cdots\cdots\cdots\cdots\cdots \\
v_p & : & (\alpha_p \quad \beta_p \quad \gamma_p \quad \cdots\cdots)
\end{array}
$$

- se puede, entonces, reducir el rango; supongamos que se decide sustituir las p variables por 3 combinaciones lineales de aquellas:

$$
\begin{array}{rcl}
F_1 & = & \alpha_1 t_1 + \beta_1 t_2 + \gamma_1 t_3 + \cdots \\
F_2 & = & \alpha_2 t_1 + \beta_2 t_2 + \gamma_2 t_3 + \cdots \\
F_3 & = & \alpha_3 t_1 + \beta_3 t_2 + \gamma_3 t_3 + \cdots
\end{array}
$$

- la parte de varianza explicada por estas tres combinaciones lineales se calcularía mediante:

$$
\frac{\mu_1 + \mu_2 + \mu_3}{\mu_1 + \mu_2 + \mu_3 + \cdots + \mu_p}
$$

- aunque los vectores propios nos informan sobre la relación entre las variables y las combinaciones lineales, es muy interesante conocer las correlaciones entre unas y otras:

$$
\begin{array}{lll}
\text{corr}\ (t_1, F_1) & \text{corr}\ (t_1, F_2) & \text{corr}\ (t_1, F_3) \\
\text{corr}\ (t_2, F_1) & \text{corr}\ (t_2, F_2) & \text{corr}\ (t_2, F_3) \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & & \\
\text{corr}\ (t_p, F_1) & \text{corr}\ (t_p, F_2) & \text{corr}\ (t_p, F_3)
\end{array}
$$

- como las combinaciones lineales están incorreladas entre sí, por simple suma de los cuadrados de los coeficientes de correlación se llega a saber la parte de varianza explicada de cada variable:

$$
\begin{array}{ll}
t_1 : & \text{corr}^2(t_1, F_1) + \text{corr}^2(t_1, F_2) + \text{corr}^2(t_1, F_3) \\
t_2 : & \text{corr}^2(t_2, F_1) + \text{corr}^2(t_2, F_2) + \text{corr}^2(t_2, F_3) \\
\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots & \\
t_p : & \text{corr}^2(t_p, F_1) + \text{corr}_2(t_p, F_2) + \text{corr}^2(t_p, F_3)
\end{array}
$$

BIBLIOGRAFÍA

[1] **Cuadras, C.M.** (1991) "Ejemplos y aplicaciones insólitas en regresión y correlación". *Qüestiió*, **15**, 367–382.

[2] **Martin-Guzman, Martin Pliego** (1987). *Curso básico de Estadística Económica.* Editorial AC.