

Técnicas multivariadas

Análisis en componentes principales

Mario J. P. López

Departamento de Matemáticas
Universidad El Bosque

2020

Escalafón de la competitividad de los dpts de Colombia 2017

Dpto	F.E.	I.L.	B.S.C.H.	C.T.I.	I.G.P.
Amazonas	30.3	21.3	20.0	26.4	50.5
Antioquia	75.0	66.0	76.6	60.2	79.3
Arauca	35.4	36.3	45.5	4.9	61.5
Atlántico	71.7	74.6	77.1	36.5	76.3
Bog.-Cund.	91.9	78.6	87.3	95.4	83.6
⋮	⋮	⋮	⋮	⋮	⋮

- F.E.: Fortaleza de la Economía
- I.L.: Infraestructura y logística
- B.S.C.H.: Bienestar Social y Capital Humano
- C.T.I.: Ciencia, Tecnología e Innovación
- I.G.P.: Institucionalidad y Gestión Pública

Escalafón de la competitividad de los dpts de Colombia 2017

Podemos estar interesados en

- contruir un indicador global de competitividad
- representar las asociaciones entre variables y departamentos gráficamente

Objetivos del ACP

Con el ACP se pretende explicar la estructura de varianzas - covarianzas de un conjunto de variables a través de unas pocas combinaciones lineales (componentes principales) de esas mismas variables. En particular se pretende lograr

- ① una reducción de los datos,
- ② la interpretación de las componentes,
- ③ la construcción de indicadores.

ACP poblacional

- Dado el v.a. $\mathbf{X} = (X_1, X_2, \dots, X_p)^\top$ con $\text{Cov}(\mathbf{X}) = \Sigma$
- Considere las combinaciones lineales

$$Y_1 = \mathbf{a}_1^\top \mathbf{X} = a_{11}X_1 + a_{12}X_2 + \dots + a_{1p}X_p$$

$$Y_2 = \mathbf{a}_2^\top \mathbf{X} = a_{21}X_1 + a_{22}X_2 + \dots + a_{2p}X_p$$

$$\vdots$$

$$Y_p = \mathbf{a}_p^\top \mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \dots + a_{pp}X_p$$

con

$$\text{Var}(Y_j) = \mathbf{a}_j^\top \Sigma \mathbf{a}_j, \quad j = 1, \dots, p$$

$$\text{Cov}(Y_j, Y_k) = \mathbf{a}_j^\top \Sigma \mathbf{a}_k, \quad j, k = 1, \dots, p$$

ACP poblacional

- Las c.p.'s son aquellas combinaciones lineales Y_1, Y_2, \dots, Y_p incorreladas con máxima varianza.
- Dado que la varianza $\mathbf{a}_j^T \Sigma \mathbf{a}_j$ puede incrementar conforme se escojan \mathbf{a}_j “más grandes” se impone la restricción $\mathbf{a}_j^T \mathbf{a}_j = 1$.
- La 1ra c.p. es la combinación lineal Y_1 que maximiza $\mathbf{a}_1^T \Sigma \mathbf{a}_1$ sujeto a $\mathbf{a}_1^T \mathbf{a}_1 = 1$
- La 2da c.p. es la combinación lineal Y_2 que maximiza $\mathbf{a}_2^T \Sigma \mathbf{a}_2$ sujeto a $\mathbf{a}_2^T \mathbf{a}_2 = 1$ y $\text{Cov}(Y_1, Y_2) = 0$
- La j -ésima c.p. es la combinación lineal Y_j que maximiza $\mathbf{a}_j^T \Sigma \mathbf{a}_j$ sujeto a $\mathbf{a}_j^T \mathbf{a}_j = 1$ y $\text{Cov}(Y_j, Y_k) = 0$ para $k < j$

Resultado

Dado el v.a. $\mathbf{X} = (X_1, \dots, X_p)^\top$ y las parejas $(\lambda_j, \mathbf{e}_j)$, $j = 1, \dots, p$, de valores y vectores propios asociados de $\Sigma = \text{Cov}(\mathbf{X})$, con $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, entonces la j -ésima componente principal está dada por

$$Y_j = \mathbf{e}_j^\top \mathbf{X} = e_{j1}X_1 + e_{j2}X_2 + \dots + e_{jp}X_p, \quad j = 1, \dots, p$$

Con esta elección

$$\text{Var}(Y_j) = \mathbf{e}_j^\top \Sigma \mathbf{e}_j, \quad j = 1, 2, \dots, p$$

$$\text{Cov}(Y_j, Y_k) = \mathbf{e}_j^\top \Sigma \mathbf{e}_k = 0, \quad j \neq k$$

Resultado

Para las c.p.'s $Y_j = \mathbf{e}_j^\top \mathbf{X}$, $j = 1, 2, \dots, p$,

$$\text{tr}(\Sigma) = \sum_{j=1}^p \sigma_j^2 = \sum_{j=1}^p \text{Var}(X_j) = \sum_{j=1}^p \lambda_j$$

De esta forma:

- $\sum_{j=1}^p \lambda_j$ es la varianza total y
- $\lambda_j / \sum_{j=1}^p \lambda_j$ es la proporción de la varianza total explicada por la j -ésima c.p.

Resultado

Para las c.p.'s $Y_j = \mathbf{e}_j^\top \mathbf{X}$, $j = 1, 2, \dots, p$,

$$\rho(Y_j, X_k) = \frac{e_{jk} \sqrt{\lambda_j}}{\sigma_k}, \quad j, k = 1, 2, \dots, p,$$

donde $\rho(Y_j, X_k) = \text{Cor}(Y_j, X_k)$.

Resultado

Si $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma)$ entonces $\mathbf{Z} = (\mathbf{X} - \boldsymbol{\mu}) \sim N(\mathbf{0}, \Sigma)$, luego las c.p.'s de \mathbf{Z} son tal que $Y_j = \mathbf{e}_j^\top \mathbf{Z}$, $j = 1, 2, \dots, p$, y

$$\begin{aligned}(\mathbf{X} - \boldsymbol{\mu})^\top \Sigma^{-1} (\mathbf{X} - \boldsymbol{\mu}) &= \mathbf{Z}^\top \Sigma^{-1} \mathbf{Z} \\&= \sum_{j=1}^p \frac{1}{\lambda_j} \left(\mathbf{e}_j^\top \mathbf{Z} \right)^2 \\&= \sum_{j=1}^p \frac{1}{\lambda_j} Y_j^2\end{aligned}$$

Resultado

Dada la variable estandarizada $\mathbf{Z} = (V^{1/2})^{-1}(\mathbf{X} - \boldsymbol{\mu})$, con

$$V = \text{Diag} \{ \sigma_1^2, \sigma_2^2, \dots, \sigma_p^2 \}$$

y $\text{Var}(\mathbf{Z}) = \text{Cor}(\mathbf{Z}) = \mathcal{P}$, las c.p.'s de \mathbf{Z} son tal que

$$Y_j = \mathbf{e}_j^\top \mathbf{Z} = \mathbf{e}_j^\top (V^{-1/2})^{-1} (\mathbf{X} - \boldsymbol{\mu}), \quad j = 1, 2, \dots, p$$

con

$$\sum_{j=1}^p \lambda_j = \sum_{j=1}^p \text{Var}(Z_j) = p$$

y

$$\rho(Y_j, Z_k) = e_{jk} \sqrt{\lambda_j}, \quad j, k = 1, 2, \dots, p.$$

Ejemplo

Considere

$$\Sigma = \begin{bmatrix} 1 & 4 \\ 4 & 100 \end{bmatrix} \Rightarrow \mathcal{P} = \begin{bmatrix} 1 & 0.4 \\ 0.4 & 1 \end{bmatrix}$$

con valores y vectores propios para Σ dados por

$$\lambda_1 = 100.16 \Rightarrow \mathbf{e}_1^T = [0.040, 0.999]$$

$$\lambda_2 = 0.84 \Rightarrow \mathbf{e}_2^T = [0.999, -0.040]$$

y para \mathcal{P} dados por

$$\lambda_1 = 1.4 \Rightarrow \mathbf{e}_1^T = [0.707, 0.707]$$

$$\lambda_2 = 0.6 \Rightarrow \mathbf{e}_2^T = [0.707, -0.707]$$

Ejemplo

Las respectivas componentes principales para Σ son

$$Y_1 = 0.040X_1 + 0.999X_2$$

$$Y_2 = 0.999X_1 - 0.040X_2$$

y para \mathcal{P}

$$Y_1 = 0.707Z_1 + 0.707Z_2 = 0.707 \frac{X_1 - \mu_1}{1} + 0.707 \frac{X_2 - \mu_2}{10}$$

$$= 0.707(X_1 - \mu_1) + 0.0707(X_2 - \mu_2)$$

$$Y_2 = 0.707Z_1 - 0.707Z_2 = 0.707 \frac{X_1 - \mu_1}{1} - 0.707 \frac{X_2 - \mu_2}{10}$$

$$= 0.707(X_1 - \mu_1) - 0.0707(X_2 - \mu_2)$$

Ejemplo

Para Σ , la 1ra c.p. explica

$$\frac{\lambda_1}{\lambda_1 + \lambda_2} = \frac{100.16}{100.16 + 0.84} = 0.992$$

con

$$\rho(Y_1, X_1) = \frac{e_{11}\sqrt{\lambda_1}}{\sigma_1} = \frac{0.040\sqrt{100.16}}{1} = 0.4$$

$$\rho(Y_1, X_2) = \frac{e_{12}\sqrt{\lambda_1}}{\sigma_2} = \frac{0.999\sqrt{100.16}}{10} = 0.1$$

Ejemplo

Para \mathcal{P} , la 1ra c.p. explica

$$\frac{\lambda_1}{p} = \frac{1.4}{2} = 0.7$$

con

$$\rho(Y_1, Z_1) = e_{11}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837$$

$$\rho(Y_1, Z_2) = e_{12}\sqrt{\lambda_1} = 0.707\sqrt{1.4} = 0.837$$

Muestra

Considere una m.a. $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n$ de una población con media $\boldsymbol{\mu}$ y matriz de covarianzas $\boldsymbol{\Sigma}$, con

$$\bar{\mathbf{x}} = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{pmatrix}, \quad S = \begin{pmatrix} s_1^2 & s_{12} & \cdots & s_{1p} \\ s_{21} & s_2^2 & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_p^2 \end{pmatrix}$$

y

$$R = \begin{pmatrix} 1 & r_{12} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & 1 \end{pmatrix}$$

Objetivo

El objetivo es construir componentes principales muestrales

$$y_j = \mathbf{a}_j^\top \mathbf{x} = a_{j1}x_1 + a_{j2}x_2 + \cdots + a_{jp}x_p, \quad j = 1, \dots, p,$$

teniendo en cuenta que

$$\bar{\mathbf{y}} = \mathbf{a}^\top \bar{\mathbf{x}}$$

$$\text{Var}(y_j) = \mathbf{a}_j^\top \mathbf{S} \mathbf{a}_j$$

$$\text{Cov}(y_j, y_k) = \mathbf{a}_j^\top \mathbf{S} \mathbf{a}_k$$

Componentes principales

- La 1ra c.p. es la combinación lineal y_1 que maximiza $\mathbf{a}_1^\top \mathbf{S} \mathbf{a}_1$ sujeto a $\mathbf{a}_1^\top \mathbf{a}_1 = 1$
- La 2da c.p. es la combinación lineal y_2 que maximiza $\mathbf{a}_2^\top \mathbf{S} \mathbf{a}_2$ sujeto a $\mathbf{a}_2^\top \mathbf{a}_2 = 1$ y $\text{Cov}(y_1, y_2) = 0$
- La j -ésima c.p. es la combinación lineal y_j que maximiza $\mathbf{a}_j^\top \mathbf{S} \mathbf{a}_j$ sujeto a $\mathbf{a}_j^\top \mathbf{a}_j = 1$ y $\text{Cov}(y_j, y_k) = 0$ para $k < j$

Resultado

Dadas las parejas $(\lambda_j, \mathbf{e}_j)$, $j = 1, \dots, p$, de valores y vectores propios asociados de S , con $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_p \geq 0$, entonces la j -ésima componente principal está dada por

$$y_j = \mathbf{e}_j^\top \mathbf{x} = e_{j1}x_1 + e_{j2}x_2 + \dots + e_{jp}x_p, \quad j = 1, \dots, p$$

Con esta elección

$$\text{Var}(y_j) = \mathbf{e}_j^\top S \mathbf{e}_j, \quad j = 1, 2, \dots, p$$

$$\text{Cov}(y_j, y_k) = \mathbf{e}_j^\top S \mathbf{e}_k = 0, \quad j \neq k$$

Resultado

Para las c.p.'s $y_j = \mathbf{e}_j^\top \mathbf{x}$, $j = 1, 2, \dots, p$,

$$\text{tr}(\mathbf{S}) = \sum_{j=1}^p s_j^2 = \sum_{j=1}^p \lambda_j = \sum_{j=1}^p y_j$$

De esta forma:

- $\sum_{j=1}^p \lambda_j$ es la varianza muestral total y
- $\lambda_j / \sum_{j=1}^p \lambda_j$ es la proporción de la varianza muestral total explicada por la j -ésima c.p.

Resultado

Para las c.p.'s $y_j = \mathbf{e}_j^\top \mathbf{x}$, $j = 1, 2, \dots, p$,

$$r(y_j, x_k) = \frac{e_{jk} \sqrt{\lambda_j}}{s_k}, \quad j, k = 1, 2, \dots, p,$$

Resultado

Dada la variable estandarizada $\mathbf{z} = (V^{1/2})^{-1} (\mathbf{x} - \bar{\mathbf{x}})$, con

$$V = \text{Diag} \{s_1^2, s_2^2, \dots, s_p^2\}$$

y $S_z = \mathbf{R}$, las c.p.'s de \mathbf{z} son tal que

$$y_j = \mathbf{e}_j^\top \mathbf{z} = \mathbf{e}_j^\top (V^{-1/2})^{-1} (\mathbf{x} - \bar{\mathbf{x}}), \quad j = 1, 2, \dots, p$$

con

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

y

$$r(y_j, z_k) = e_{jk} \sqrt{\lambda_j}, \quad j, k = 1, 2, \dots, p.$$

Ejemplo (competitividad)

Estimadores de μ, Σ y \mathcal{P}

$$\bar{\mathbf{x}} = \begin{pmatrix} 48.6 \\ 51.3 \\ 56.7 \\ 25.5 \\ 60.9 \end{pmatrix}, \quad S = \begin{pmatrix} 327.0 & 315.1 & 345.7 & 297.4 & 176.5 \\ & 394.7 & 391.2 & 264.5 & 177.9 \\ & & 467.4 & 292.0 & 225.0 \\ & & & 420.8 & 176.0 \\ & & & & 146.7 \end{pmatrix}$$

$$R = \begin{pmatrix} 1 & 0.877 & 0.884 & 0.802 & 0.806 \\ & 1 & 0.911 & 0.649 & 0.739 \\ & & 1 & 0.658 & 0.859 \\ & & & 1 & 0.708 \\ & & & & 1 \end{pmatrix}$$

Ejemplo (competitividad)

Valores y vectores propios de S

$$\lambda = (1470.6, 180.3, 57.4, 29.8, 18.6)^T$$

$$\mathbf{e} = (\mathbf{e}_1 | \mathbf{e}_2 | \mathbf{e}_3 | \mathbf{e}_4 | \mathbf{e}_5)$$

$$= \begin{pmatrix} 0.453 & 0.049 & 0.166 & 0.873 & 0.050 \\ 0.483 & -0.372 & 0.599 & -0.365 & 0.370 \\ 0.536 & -0.399 & -0.388 & -0.147 & -0.618 \\ 0.445 & 0.837 & 0.068 & -0.283 & -0.129 \\ 0.276 & -0.004 & -0.677 & -0.053 & 0.680 \end{pmatrix}$$

Ejemplo (competitividad)

Componentes principales

Primeras dos c.p.'s

$$y_1 = 0.453x_1 + 0.483x_2 + 0.536x_3 + 0.445x_4 + 0.276x_5$$

$$y_2 = 0.049x_1 - 0.372x_2 - 0.399x_3 + 0.837x_4 - 0.004x_5$$

Proporción de variabilidad explicada por las dos primeras c.p.'s

$$\frac{\lambda_1}{\sum_{j=1}^5 \lambda_j} = \frac{1470.6}{1756.7} = 0.837$$

$$\frac{\lambda_2}{\sum_{j=1}^5 \lambda_j} = \frac{180.3}{1756.7} = 0.103$$

Correlaciones

Correlaciones con el 1er c.p.

$$r(y_1, x_1) = \frac{e_{11}\sqrt{\lambda_1}}{s_1} = \frac{0.453\sqrt{1470.6}}{18.1} = 0.961$$

$$r(y_1, x_2) = \frac{e_{12}\sqrt{\lambda_1}}{s_2} = \frac{0.483\sqrt{1470.6}}{19.9} = 0.931$$

$$r(y_1, x_3) = \frac{e_{13}\sqrt{\lambda_1}}{s_3} = \frac{0.536\sqrt{1470.6}}{21.6} = 0.951$$

$$r(y_1, x_4) = \frac{e_{14}\sqrt{\lambda_1}}{s_4} = \frac{0.445\sqrt{1470.6}}{20.5} = 0.832$$

$$r(y_1, x_5) = \frac{e_{15}\sqrt{\lambda_1}}{s_5} = \frac{0.276\sqrt{1470.6}}{12.1} = 0.873$$

Correlaciones

Correlaciones con el 2do c.p.

$$r(y_2, x_1) = \frac{e_{21}\sqrt{\lambda_2}}{s_1} = \frac{0.049\sqrt{180.3}}{18.1} = 0.036$$

$$r(y_2, x_2) = \frac{e_{22}\sqrt{\lambda_2}}{s_2} = \frac{-0.372\sqrt{180.3}}{19.9} = -0.252$$

$$r(y_2, x_3) = \frac{e_{23}\sqrt{\lambda_2}}{s_3} = \frac{-0.399\sqrt{180.3}}{21.6} = -0.248$$

$$r(y_2, x_4) = \frac{e_{24}\sqrt{\lambda_2}}{s_4} = \frac{0.837\sqrt{180.3}}{20.5} = 0.548$$

$$r(y_2, x_5) = \frac{e_{25}\sqrt{\lambda_2}}{s_5} = \frac{-0.004\sqrt{180.3}}{12.1} = -0.005$$

Ejemplo (competitividad)

Valores y vectores propios de R

$$\lambda = (4.167, 0.426, 0.265, 0.086, 0.056)^T$$

$$\mathbf{e} = (\mathbf{e}_1 | \mathbf{e}_2 | \mathbf{e}_3 | \mathbf{e}_4 | \mathbf{e}_5)$$

$$= \begin{pmatrix} 0.470 & 0.048 & -0.268 & -0.830 & 0.125 \\ 0.450 & -0.402 & -0.463 & 0.451 & 0.466 \\ 0.465 & -0.380 & 0.081 & 0.096 & -0.790 \\ 0.407 & 0.831 & -0.165 & 0.311 & -0.140 \\ 0.442 & -0.006 & 0.824 & 0.036 & 0.352 \end{pmatrix}$$

Ejemplo (competitividad)

Componentes principales

Primeras dos c.p.'s

$$y_1 = 0.470z_1 + 0.450z_2 + 0.465z_3 + 0.407z_4 + 0.442z_5$$

$$y_2 = 0.048z_1 - 0.402z_2 - 0.380z_3 + 0.831z_4 - 0.006z_5$$

Proporción de variabilidad explicada por las dos primeras c.p.'s

$$\frac{\lambda_1}{p} = \frac{4.167}{5} = 0.833$$

$$\frac{\lambda_2}{p} = \frac{0.426}{5} = 0.085$$

Correlaciones

Correlaciones con el 1er c.p.

$$r(y_1, z_1) = e_{11} \sqrt{\lambda_1} = 0.470 \sqrt{4.167} = 0.959$$

$$r(y_1, z_2) = e_{12} \sqrt{\lambda_1} = 0.450 \sqrt{4.167} = 0.919$$

$$r(y_1, z_3) = e_{13} \sqrt{\lambda_1} = 0.465 \sqrt{4.167} = 0.949$$

$$r(y_1, z_4) = e_{14} \sqrt{\lambda_1} = 0.407 \sqrt{4.167} = 0.830$$

$$r(y_1, z_5) = e_{15} \sqrt{\lambda_1} = 0.442 \sqrt{4.167} = 0.902$$

Correlaciones

Correlaciones con el 2do c.p.

$$r(y_2, z_1) = e_{21}\sqrt{\lambda_2} = 0.048\sqrt{0.426} = 0.032$$

$$r(y_2, z_2) = e_{22}\sqrt{\lambda_2} = -0.402\sqrt{0.426} = -0.263$$

$$r(y_2, z_3) = e_{23}\sqrt{\lambda_2} = -0.380\sqrt{0.426} = -0.248$$

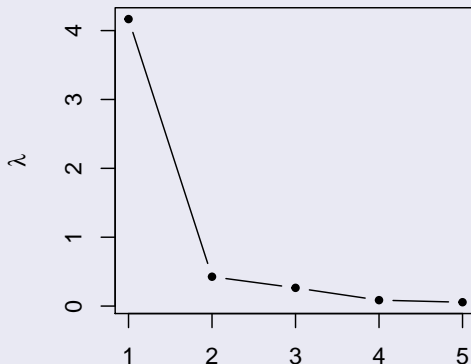
$$r(y_2, z_4) = e_{24}\sqrt{\lambda_2} = 0.831\sqrt{0.426} = 0.543$$

$$r(y_2, z_5) = e_{25}\sqrt{\lambda_2} = -0.006\sqrt{0.426} = -0.004$$

Ejemplo (competitividad)

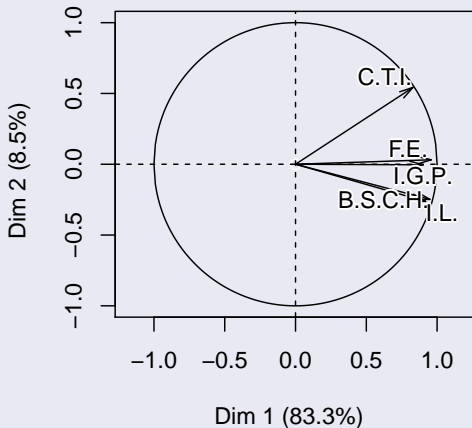
Número de componentes

Scree plot



Variables

Círculo de correlaciones



Individuos

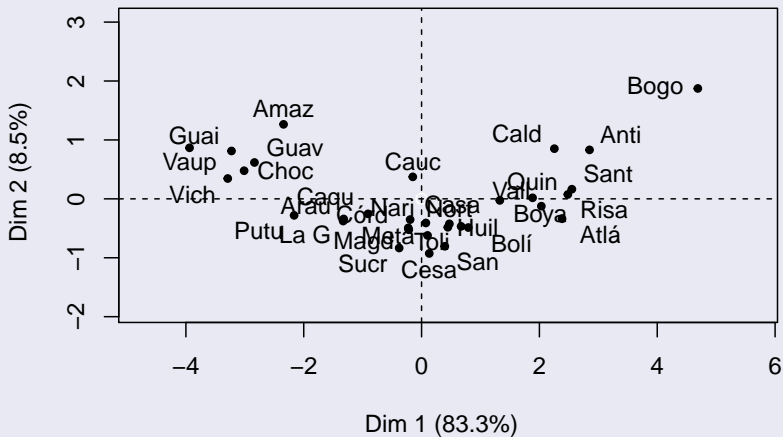
Para cada individuo $i = 1, 2, \dots, n$

$$y_{1i} = 0.470z_{1i} + 0.450z_{2i} + 0.465z_{3i} + 0.407z_{4i} + 0.442z_{5i}$$

$$y_{2i} = 0.048z_{1i} - 0.402z_{2i} - 0.380z_{3i} + 0.831z_{4i} - 0.006z_{5i}$$

Dpto	z_1	z_2	z_3	z_4	z_5	y_1	y_2
Amaz.	-1.013	-1.510	-1.698	0.042	-0.856	-2.343	1.264
Antio.	1.459	0.740	0.920	1.690	1.522	2.851	0.831
Arau.	-0.731	-0.755	-0.518	-1.006	0.053	-1.331	-0.377
Atlán.	1.276	1.173	0.943	0.535	1.275	2.384	-0.338
Bogo.	2.393	1.375	1.415	3.406	1.877	4.689	1.874
\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots	\vdots




Individuos



Indicador global de competitividad

La 1ra c.p. se puede usar como un indicador global de competitividad

$$\begin{aligned} I &= \sum_{j=1}^p e_{1j} \frac{x_j - \bar{x}_j}{s_j} \\ &= 0.470 \frac{x_1 - 48.6}{18.1} + 0.450 \frac{x_2 - 51.3}{19.9} + 0.465 \frac{x_3 - 56.7}{21.6} + \\ &\quad 0.407 \frac{x_4 - 25.5}{20.5} + 0.442 \frac{x_5 - 60.9}{12.1} \\ &= -6.374 + 0.026x_1 + 0.022x_2 + 0.021x_3 + 0.020x_4 + 0.037x_5 \end{aligned}$$

-  Johnson, R., Wichern, D.
Applied Multivariate Statistical Analysis.
6th ed. Pearson, 2007.
-  Mardia, K., Kent, J., Bibby, J.
Multivariate Analysis.
Academic Press, 1995.
-  Rencher, A.
Methods of Multivariate Analysis.
2nd ed. Willey, 2002.