Introducción

Análisis Multivariable

Santiago Alférez

Agosto de 2020

Análisis Estadístico de Datos
MACC

Universidad del Rosario

Contenidos

Introducción

Introducción y algunas ideas

- La mayoría de problemas involucran varias medidas de múltiples variables.
- Extenderemos algunos métodos y veremos otros nuevos, involucrando álgebra matricial, cálculo de varias variables y, probabilidad y estadística.
- Particularmente, muchos métodos que veremos se basan en la distribución normal multivariable.
- Utilizaremos R, Rstudio y Rmarkdown.
- Existen muchísimas aplicaciones del análisis estadístico de datos.

Etapas donde se aplican los métodos multivariables

Algunas etapas de investigaciones científicas donde se aplican los métodos de análisis estadístico de datos son:

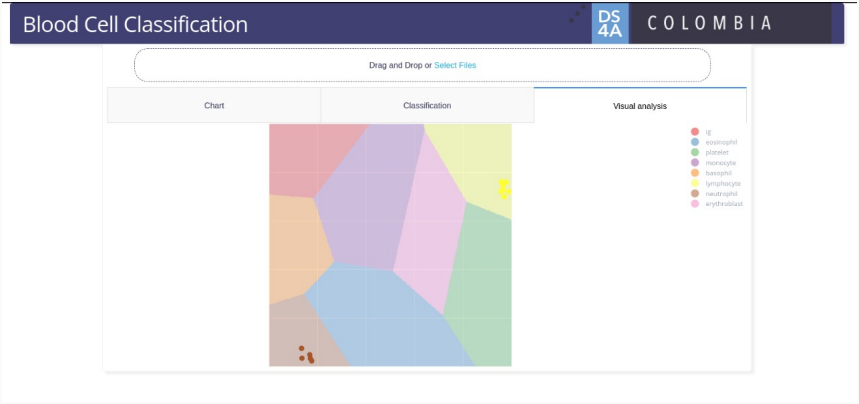
Objetivos científicos

- Reducción de datos o simplificación estructural
- Ordenamiento y agrupamiento
- Investigación acerca de la dependencia entre variables
- Predicción
- Construcción y prueba de hipótesis

Un ejemplo reciente sobre varios objetivos

login: team14; password: t34m14

pbcellrecognition.herokuapp.com



Organización y nomenclatura

Matrices

Usaremos la notación x_{ij} como medida de la k-ésima variable del j-ésimo dato u observación.

Notación de conjunto de datos

	Variable 1	Variable 2	$\cdots$	Variable k	$\cdots$	Variable p
Item 1:	x_{11}	x_{12}	$\cdots$	x_{1k}	$\cdots$	x_{1p}
Item 2:	x_{21}	x_{22}	$\cdots$	x_{2k}	$\cdots$	x_{2p}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item j :	x_{j1}	x_{j2}	$\cdots$	x_{jk}	$\cdots$	x_{jp}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
Item n :	x_{n1}	x_{n2}	$\cdots$	x_{nk}	$\cdots$	x_{np}

Matrices

Usaremos la notación x_{ij} como medida de la k-ésima variable del j-ésimo dato u observación.

Notación de matriz

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Estadística Descriptiva Multivariable

Estadística descriptiva multivariable

- Generalmente, los conjuntos de datos son grandes y es complejo visualizar y extraer información.
- Para obtener información, es útil calcular medidas que resumen los datos.
- Conocemos algunas, cómo la media (arimética) que suministra una medida de ubicación (valor central).
- Otra medida es el promedio de los cuadrados de la distancias de todos los números respecto a la media. Esta suministra una medida de dispersión o variación.
- Principalmente, trabajaremos con estadísticas que miden la ubicación, la variación y la asociación lineal.

Variable k

x_{1k}

x_{2k}

$\vdots$

x_{jk}

$\vdots$

x_{nk}

Media muestral

$$\bar{x}_k = \frac{1}{n}\sum_{j=1}^{n} x_{jk} \quad k = 1, 2, \ldots, p$$

Varianza muestral

$$s_k^2 = \frac{1}{n}\sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \quad k = 1, 2, \ldots, p$$

Observe que en este caso la varianza está dividida por n no por $n - 1$. ¿Qué implicaciones tiene?

Notación matricial de la varianza

Aunque estamos acostumbrados a la notación s^2 para la varianza, por razones que veremos, se puede considerar la varianza como la diagonal de una matriz:

$$s_k^2 = s_{kk} = \frac{1}{n}\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2 \quad k = 1, 2, \ldots, p$$

La raíz cuadrada de la varianza muestral $\sqrt{s_{kk}}$ es la **desviación estándar**

Si hay varias variables ¿Cuál es la varianza?

Supongamos que tenemos n pares de medidas sobre dos variables:

$$\begin{bmatrix} x_{11} \\ x_{12} \end{bmatrix}, \begin{bmatrix} x_{21} \\ x_{22} \end{bmatrix}, \ldots, \begin{bmatrix} x_{n1} \\ x_{n2} \end{bmatrix}$$

Lo anterior se puede expresar cómo dos variables x_{j1} y x_{j2} sobre n experimentos $(j = 1, 2, \ldots, n)$.

La covarianza muestral

$$s_{12} = \frac{1}{n} \sum_{j=1}^{n} (x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2)$$

Mide la asociación lineal entre las dos variables.

$$s_{12} = \frac{1}{n} \sum_{j=1}^{n} (x_{j1} - \bar{x}_1)(x_{j2} - \bar{x}_2)$$

Consideraciones sobre la covarianza

- Es el promedio del producto entre las desviaciones de sus respectivas medias.
- Si se observan valores grandes en ambas variables, y pequeños valores también se presenta de forma conjunta, entonces s_{12} será positiva.
- Si se presentan valores grandes de una variable con valores pequeños de la otra variable, entonces s_{12} será negativa.
- Si no hay asociación entre los valores de las dos variables, s_{12} será aproximadamente cero.

Si tenemos p variables:

$$\begin{bmatrix} x_{11} & \cdots & x_{1i} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & \cdots & x_{2i} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & & \vdots & & \vdots & & \vdots \\ x_{j1} & \cdots & x_{ji} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & & \vdots & & \vdots & & \vdots \\ x_{n1} & \cdots & x_{ni} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

Covarianza muestral

$$s_{ik} = \frac{1}{n} \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)$$

$$i = 1, 2, \ldots, p, \quad k = 1, 2, \ldots, p$$

- Mide la asociación entre las i-ésima y k-ésima variables.
- Cuando $i = k$ es la varianza muestral.
- Es simétrica $s_{ik} = s_{ki}$.

Si ahora normalizamos para no depender de las unidades:

Coeficiente de correlación muestral

$$r_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = \frac{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k)}{\sqrt{\sum_{j=1}^{n}(x_{ji} - \bar{x}_i)^2}\sqrt{\sum_{j=1}^{n}(x_{jk} - \bar{x}_k)^2}}$$

para $i = 1, 2, \ldots, p$ y $k = 1, 2, \ldots, p$

- Es simétrica, $r_{ik} = r_{ki}$ para todo i y k.
- Es una medida de asociación lineal entre dos variables, pero no depende de las unidades de medida.
- Es una versión estandarizada de la covarianza, donde el producto de las raíces cuadradas de las varianzas producen la estandarización.

La correlación es más sencilla de interpretar que la covarianza, porque su magnitud es limitada

Propiedades de la correlación

- r está entre -1 y $+1$.
- r mide la asociación lineal
  - Si $r = 0$ no hay asociación.
  - Si $r < 0$ una variable es más grande que su promedio cuando la otra es más pequeña que su promedio.
  - Si $r > 0$ una variable es más grande cuando la otra es más grande (que el promedio) y la misma tendencia cuando ambas son pequeñas.
- El valor de r_{ik} es invariante si las medidas sobre la i-ésima variable y la k-ésima variables cambian de forma lineal:

$$y_{jl} = ax_{ji} + b \text{ y } y_{jk} = cx_{jk} + d.$$

Suma cuadrada de las desviaciones respecto a la media

$$w_{kk} = \sum_{j=1}^{n} (x_{jk} - \bar{x}_k)^2 \quad k = 1, 2, \ldots, p$$

Suma del producto de las desviaciones respecto a la media

$$w_{ik} = \sum_{j=1}^{n} (x_{ji} - \bar{x}_i)(x_{jk} - \bar{x}_k) \quad i = 1, 2, \ldots, p, \quad k = 1, 2, \ldots, p$$

Estadística descriptiva matricial

Media Muestral

$$\overline{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix}$$

Covarianzas muestrales

$$\mathbf{S}_n = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ s_{p1} & s_{p2} & \cdots & s_{pp} \end{bmatrix}$$

Correlaciones muestrales

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

Ejemplo

Considere los siete pares de medidas (x_1, x_2) siguientes:

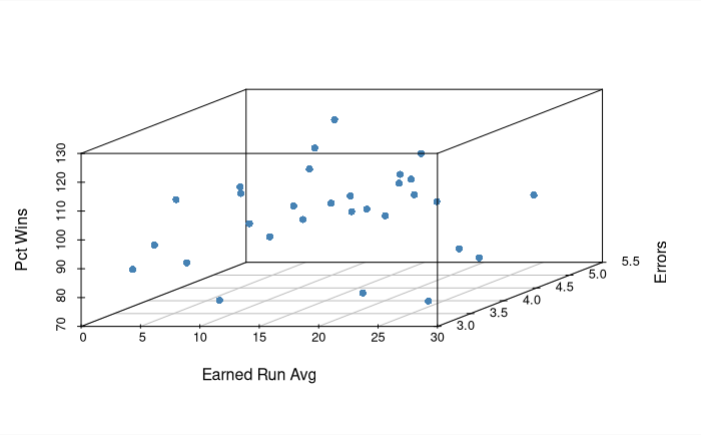
x_1	3	4	2	6	8	2	5
x_2	5	5.5	4	7	10	5	7.5

a. Dibuje el diagrama de dispersión

b. Calcule las medias muestrales, las varianzas muestrales para ambas variables y la covarianza.

Visualizaciones (EDA)

Gráficos para análisis exploratorio



Continuamos en el notebook ...

Ejercicios

Ejercicio EDM

Ejercicio

Un periódico lista los siguientes precios para carros usados, para un compacto extranjero con edad x_1 medido en años y un precio de venta x_2 medido en miles de dólares:

x_1	1	2	3	3	4	5	6	8	9	11
x_2	18.95	19.00	17.95	15.54	14.00	12.95	8.94	7.49	6.00	3.99

a. Construya un diagrama de dispersión de los datos

b. Infiera el signo de la covarianza muestral s_{12} a partir del diagrama de dispersión

c. Calcule las medias muestrales $\bar{x_1}$ y $\bar{x_2}$ y las varianzas muestrales s_{11} y s_{22}. Calcule la covarianza muestral s_{12} y el coeficiente de correlación muestral r_{12}. Interprete esas cantidades.

d. Escriba el vector de medias $\bar{\mathbf{x}}$, la matriz de covarianza muestral $\mathbf{S_n}$ y la matriz de correlación muestral $\mathbf{R}$