Curso de Estadística para el Programa Universitario de Medio Ambiente

Leticia Gracia Medrano
IIMAS Depto. Probabilidad y Estadística
lety@sigma.iimas.unam.mx

1. Análisis Exploratorio de Datos

1.1. Introducción

El análisis de datos multivariados tiene que ver con el estudio de las asociaciones entre conjuntos de mediciones. Se analizan las relaciones entre dos o más conjuntos de mediciones que se hacen a cada objeto o individuo, en una o varias muestras. Los datos los acomodaremos en un arreglo de la forma:

$$X = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1j} & \cdots & X_{1p} \\ X_{21} & X_{21} & \cdots & X_{2j} & \cdots & X_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{i1} & X_{i2} & \cdots & X_{ij} & \cdots & X_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{nj} & \cdots & X_{np} \end{bmatrix}$$

Entonces cada renglón representa a un vector $\bar{X}_i \in \Re^p$ $i = 1, \ldots, n.$, esto puede verse como que tenemos n vectores renglón de dimensión p.

Estos renglones corresponden a: personas, organizaciones, eventos, países, etc. Las variables, son las características, o las propiedades que son medidas a los objetos. Se tiene un universo de observaciones potenciales, pero de ese universo sólo se observará un subconjunto. A este subconjunto se le aplicará un modelo de medición (las condiciones e instrumento para medir) del que resultarán los datos. Se buscan asociaciones entre las variables a través de los distintos modelos multivariados.

2. Calidad en los datos

1. Inspección visual.- Par ver si hay datos fuera de los rangos establecidos, conocer el máximo y mínimo de cada variable. Verificar que las codificaciones sean consistentes en toda la base.

2. Distribución de frecuencias.- De las variables de mayor interés.

3. Gráficas de dispersión.- Identificar grupos u observaciones discrepantes.

4. Verificar métodos de recolección de los datos.- Detectar posibles fuentes de sesgo.

5. Observaciones faltantes.- Tratar de rastrearlas ir a registros originales, razones de su omisión. Definir que se hará con estas observaciones, se puede usar algún valor de reemplazo o imputación o seleccionar cuáles si se desechan. Los valores faltantes generan sesgo este tema es de suma importancia

OJO Cuidado con el número de dígitos a guardar, puede perderse precisión o al revés desperdiciar espacio. Tener control sobre los estándares de medición. Un grupo de datos de poca calidad no merece un análisis muy detallado.

3. Fases del análisis

- Manipulación inicial de los datos. Reunir los datos en forma conveniente.
- Análisis preliminar. Se intenta aclarar la forma de los datos y ver que dirección debe tomar el análisis.
- Análisis definitivo que dará las bases para las conclusiones.
- Presentación de conclusiones

4. Escalas de Medición

Escala de medición	Operaciones	Cambios permitidos	Ejemplo	Valores
Nominal	Pertenencia a categoría	de nombre	Sexo Estado civil	Masc., Fem. Casado Divor
Ordinal	Grado de intensidad	que mantengan orden	Calificaciones sabor	NA, S, B, MB Bno reg malo
Intervalo	Igualdad de intervalos	de escala y origen	Temperatura Tiempo	Enteros, reales
Razón	Igualdad de proporciones	de escala pero no de origen	Concentración sustancias	Enteros, reales
Absoluta	Conteo de elementos	No escala no origen	Número de hijos	Enteros

5. Datos Faltantes

5.1. Datos faltantes completamente al azar

Pueden ser muy variadas las razones por las que existan valores faltantes. Ya sea porque las condiciones climáticas, de seguridad o políticas no permiten recoger la información, porque ese día los instrumentos se descomponen, por que no se encontró a la persona u objeto de la encuesta, aquí se puede pensar que la información se perdió **completamente al azar** (MCAR por su siglas en inglés). Es decir cualdo la probabilidad de que X_i sea **no observada** no está relacionada con el valor mismo de x_i o con el de cualquier otra variable.

Por ejemplo si las personas con un nivel de ingresos alto tienden a no contestar por miedo a ser sujetos "secuestrables", entonces esa observación no se

perdió completamente al azar. MCAR corresponde a pensar que ese dato se perdió con la misma probabilidad que cualquier otro dato. Si la persona no responde acerca de sus ingresos, de la misma manera que no responde a cuántos hijos tiene, entonces se considera MCAR. En este caso los parámetros pueden estimarse sin sesgo.

5.2. Datos faltantes al azar

A diferencia de los datos MCAR, donde la probabilidad de no observar a X_i no depende del valor mismo de x_i o de otras variables. En este caso esa probabilidad no dependerá de x_i luego de controlar o condicionar con otra variable.

Por ejemplo, una persona con depresión puede ser que tienda más a no contestar acerca de su ingreso, la gente con depresión a su vez en general tiene menos ingresos, entonces lo que ocurre es que si hay un tasa alta de no respuesta entre las personas con depresión, la media real puede ser menor que la calculada con los datos existentes, es decir sin tomar en cuenta a los datos faltantes. Ahora si entre las personas con depresión la probabilidad de no contestar acerca de su ingreso no está relacionada con su nivel de ingreso, entonces los datos se consideran faltantes al azar, (MAR). Esto No significa que estos faltantes no produzcan sesgo y que se pueda uno olvidar del problema.

5.3. Datos Faltantes no al azar

Cuando no son MCAR ni MAR entonces se dice que son datos faltantes no al azar (MNAR). Ejemplo: Si se estudia una cierta enfermedad y las persona que padecen esa enfermedad son las que tienen una mayor probabilidad a no contestar a si la padecen, entonces los datos son faltantes no al azar, MNAR. Claramente el estimador de la proporción que padece esa enfermedad será menor que la proporción que se obtendría con los datos completos. Lo mismo ocurre en el caso de las personas con menor ingreso son las que tienden a no contestar su nivel de ingreso. Esta falta de datos no al azar es un problema, la única manera de obtener un estimador insesgado es modelar la esa ausencia de datos y los valores mismos de las ausencias, esta tarea no es para nada simple.

6. Tratamiento de datos faltantes

6.1. Omisión total

Si los datos son MCAR las estimaciones obtenidas serán insesgadas si no son MCAR serám sesgadas, hay que tener en cuenta que esta pérdida de datos genera pérdida de potencia en las pruebas.

6.2. Omisión parcial

Por ejemplo en el cálculo de las correlaciones se usan las observaciones disponibles, pero entonces cada estimación está soportada por diferentes bases de

datos. Puede ser el caso que se llegue a una matriz de correlaciones estimada NO definida positiva.

No hay que olvidar que hay que analizar a las observaciones NA y tratar de ver si se comportan (en ciertas variables) como la población total o si difieren.

Otra cosa importante es considerar que es lo que se tiene perdido. La situación de perder variables explicactivas es diferente a perder variables respuesta.

6.3. Sustitución

Hot deck, sustituir el caso por alguien semejante, (de donde sacamos a alguien semejante si ya acabó la encuesta, tener la providencia de guardar un montoncito extra para la sustitución??)

6.4. Imputación Simple

Sustituir los valores faltantes por la media (el estimador de mx́ima verosimilitud), pero eso tiene consecuencias sobre la estimación de la varianza. Pero siempre estaremos sustituyendo con el mismo valor.

O se puede sustituir usando una regresión, pero el problema sigue siendo que se sustituye por una media (esta vez condicionada) SPPS permite sumar una variación aleatoria, se subsana en algo este tipo de problema.

O se puede usar el Algoritmo EM. En regresión si se conocieran los NA, estimar los parámetros del modelo sería fácil, y si se conocieran los parámetros del modelo de los datos sería sencillo hacer predicciones insesgadas de las observaciones faltantes. Este algoritmo es iterativo y va haciendo ambas cosas: con los datos existentes se estiman los paámetros del modelo de los datos, enseguida con estos parámetros se hacen estimaciones de los datos faltantes, y de nuevo se re-estiman los paámetros con las datos ya completados. Schafer (1997) hizo un programa NORM disponible en http://www.stat.psu.edu/ jls/misoftwa.html, SPSS tiene un procedimiento que hace imputación utilizando EM.

6.5. Imputación Múltiple

En imputación múltiple se generan valores para hacer la imputación basados en los datos existentes. Suponiendo que se estima y usando x, pero esta imputación se hace varias veces, es decir tendremos varios conjuntos de datos completados. Para hacer esto se usan métodos conocidos Markov Chaim Monte Carlo. El programa NORM tambien su parte llamada data augmentation lo hace. SAS tiene dos procedimientos MI y MIANALYZE.

Schafer, J.L. & Olsden, M. K.. (1998). Multiple imputation for multivariate missing-data problems: A data analyst's perspective. Multivariate Behavioral Research, 33, 545-571

En R esta el paquete MICE, material con referencia en: Van Buuren, S., Groothuis-Oudshoorn, K. (2011) MICE: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software.

http://www.stefvanbuuren.nl/publications/MICEinR-Draft.pdf

7. Observaciones Discrepantes

Estas observaciones también son conocidas como aberrantes, discordantes, contaminantes,sorprendentes, en inglés OUTLIER. Puede definírseles de varias formas, una de ellas es decir que es una observación que se encuentra a una distancia ANORMAL de las demás, y entonces hay que definir lo que es una distancia NORMAL, es decir la observación se encuentra fuera de la nube de datos. Estas observaciones pueden distorsionar la información, también pueden ser una señal de que el modelo de distribución de los datos NO es el adecuado, o reflejar el haber encontrado una situación sorprendente o peculiar. Si la observación causa un impacto en el observador se le llama generalmente **discrepante**. Una observación **contaminante** será cualquiera que no corresponda a la distribución supuesta, y ésta puede no ser percibida por el observador.

Estas observaciones afectan fuertemente al estimador $\bar{X}$ de la media μ, y consecuentemente a los estimadores de $Var(X)$, de las de $Cov(X,Y)$ y de $Corr(X,Y)$. En análisis de regresión interesa identificar a las observaciones **influyentes**, que son aquellas que al omitirlas del análisis los valores de las $\hat{\beta}$'s varían mucho.

Detectar estas observaciones puede ser una tarea bastante complicada, sobre todo cuando se tienen datos altamente multivariados.

En el caso univariado se les puede detectar muy fácilmentea través de gráficos boxplot o tambien al verificar si la media de los datos difiere mucho de la mediana.

8. Gráficas datos univariados

- gráfica de barras y de *pie* son solo para datos categóricos, debe haber espacios entre las barras.
- histograma debe tenerse cuidado con los anchos de barras y con los puntos que se consideran en el eje de las x.
- boxplot permite rápidamente ver observaciones discrepantes.
- q-qplot Permite ver si dos muestras provienen de la misma distribución.
- stem, una versión de los histogramas pero permite ver los datos tal cual.
- series de tiempo

9. Gráficas datos multivariados

- estrellas. Convienen cuando no se tienen muchos atributos, pues con más de 10 o 12 aristas las confundimos en su forma.

- faces, debidas a Chernov, dado que el ojo humano esta muy entrenado para reconocer rostros humanos. A cada elemento de la cara: pelo, ancho cara, largo nariz, tamaño de ojos se le asocia una característica.
- curvas de Andrews, a cada individuo se le asigna una curva de la siguiente manera: $t \in [-\pi, \pi]$ Si p es impar

$$f_i(t) = \frac{X_{i1}}{\sqrt{2}} + X_{i2}\sin(t) + X_{i3}\cos(t) + \ldots + X_{ip}\cos(\frac{(p-1)}{2}t)$$

  Si p es par

$$f_i(t) = \frac{X_{i1}}{\sqrt{2}} + X_{i2}\sin(t) + X_{i3}\cos(t) + \ldots + X_{ip}\sin(\frac{p}{2}t)$$

  Estas tres gráficas no son únicas, pues según ordenemos las variables darán origen a estrellas, curvas o caras distintas.
- bagplot parecida a un boxplot pero en dos dimensiones.
- gráfica de paralelas, se usan sobre todo cuando hay varia mediciones para un solo individuo.
- series de tiempo múltiples

10. Distancias y disimilitudes

Para entrar al tema de análisis de conglomerados es muy importante introducir los conceptos de distancia y disimilitud.

Todos conocemos la llamada distancia euclideana Si $x = (x_1, x_2, \ldots, x_p)$ y $y = (y_1, y_2, \ldots, y_p)$ entonces la distancia entre x y y es:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + \ldots + (x_p - y_p)^2}$$

10.1. Producto punto

Sean $\bar{x} = [x_1, x_2, \ldots, x_p]$ y $\bar{y} = [y_1, y_2, \ldots, y_p]$ dos vectores p-dimensionales. Entonces el producto punto entre los vectores se define como $\bar{x} \cdot \bar{y} = x_1y_1 + x_2y_2 + \ldots + x_ny_n = \sum_{i=1}^{m} x_iy_j$. Al producto punto también se le llama producto interior o escalar.

Ejemplo el producto punto

$$\begin{aligned} x \cdot y &= (4, 1, 2, 3) \cdot (3, 1, 7, 2) \\ &= (4)(3) + (1)(1) + (2)(7) + (3)(2) \\ &= 33. \end{aligned}$$

Se tiene que: $d(x, y)^2 = (x - y) \cdot (x - y)$.

10.2. Distancias entre individuos con datos numéricos

La distancia puede generalizarse introduciendo una matriz$A > 0$ es decir una matriz positiva definida con dimensiones $p \times p$ esto es: $(x-y)A(x-y)$, cuando $A = S^{-1}$ donde **S** es la matriz de varianzas y covarianzas de las variables que conforman a los vectores, esta es conocida como distancia de Mahalanobis .

Hay otras distancias como la *city block* que corresponde a:

$$d(x,y) = |(x_1 - y_1)| + \ldots + |(x_p - y_p)|$$

Y también existe la generalización:

$$d(x,y) = ((x_1 - y_1)^\alpha + \ldots + (x_p - y_p)^\alpha)^{1/\alpha}$$

con α un entero,conocida como distancia Minkowski.

Para que una función entre dos puntos sea considerada distancia debe cumplir con las siguientes propiedades:

$$d(x,y) > 0 \; \forall x \neq y$$

$$d(x,y) = 0 \Leftrightarrow x \equiv y$$

$$d(x,y) \leq d(x,z) + d(z,y) \; \forall x, y, z$$

10.3. Distancias entre individuos con datos binarios

A continuación se presentan varias medidas de disimilitud. Si se tienen dos sujetos, i y j y si se denota por

$a =$ todas las coincidencias ++

$b =$ todas las no coincidencias +-

$c =$ todas las no coincidencias -+

$d =$ todas las coincidencias -+

p es el total de características $a + b + c + d = p$.

disimilitud entre i y j es

i) $d_{ij} = 1 - \frac{coincidencias}{p} = 1 - \frac{a+d}{p} = \frac{b+c}{p}$

esta **disimilitud** (en ocasiones no cumple con las propiedades de las distancias) corresponde a la proporción de variables que no coinciden.

Hay quienes sostienen que no deben tomarse en cuenta las coincidencias en ausencias, entonces: ii) Coeficiente de Jaccard

$d_{ij} = \frac{b+c}{a+b+c}$

d no se toma en cuenta pues la ausencia de cierta característica no ayuda a decir si son o no parecidos.

iii)Coeficiente de Czekanowski

$d_{ij} = \frac{b+c}{2a+b+c}$ si se quiere compensar el echar fuera a d

10.4. Distancias entre individuos con datos cualitativos

$c_{ij} =$ coeficiente de coincidencias de Sneath

$c_{ij} = \frac{\text{número de atributos en los que las unidades coinciden}}{p}$

$d_{ij} = 1 - c_{ij}$

10.5. Distancias entre individuos con datos de distintas escalas

Se puede utilizar el índice de Gower para crear distancias cuando se consideran datos con varios tipos de escalas de medición.

Para cada variable x_k la similitud entre individuo i y el j se escribe como
$\delta_{ijk} = 1$ si puedo comparar a i contra j en la variable k o
$\delta_{ijk} = 0$ en otro caso.
y a $s_{ijk} = 1$ si son iguales o $s_{ijk} = 0$ si son diferentes.
La **similitud** entre i y j esta dada por
$c_{ij} = \frac{\sum_{k=1}^{p} s_{ijk}}{\sum_{k=1}^{p} \delta_{ijk}}$
La disimilitud como $d_{ij} = 1 - c_{ij}$
Si las x_k son cuantitativas
$s_{ijk} = 1 - \frac{|x_{ik} - x_{jk}|}{rango(x_k)}$.
Si x_k es cualitativa
$s_{ijk} = 1$ si los individuos concuerdan en la variable k y $s_{ijk} = 0$ si no.

10.6. Distancias entre variables con datos numéricos

Se puede tener otro enfoque, que tal que la disimilitud que quiere calcularse es entre (**columnas** de X), en este caso el coeficiente de correlación es una buena medida. Puede considerarse que si la correlación es cercana a 1 las variables en cuestión son cercanas. Ahora, antes de construir las medidas hay que decidir si un coeficiente de correlación alto pero negativo significa un acercamiento grande entre las variables, o un total alejamiento.

Una disimilitud podría ser $d(V_1, V_2) = 1 - \rho(V_1, V_2)$ o también esta otra $d(V_1, V_2) = 1 - \rho(V_1, V_2)^2$

10.7. Distancias entre variables con datos cualitativos

Si se requiere construir disimilitud entre variables y se trabaja con datos cualitativos

Se construye una tabla de contingencias tamaño 2×2 donde $a+c+b+d = n$
$d_{kl} = \chi^2 = \frac{(ad-bc)^2(a+b+c+d)}{(a+b)(a+c)(c+d)(b+d)}$
pero depende del tamaño de la muestra $\chi^2 \leq n$.
Otra medida que puede usarse es
$d_{kl} = 1 - \sqrt{\frac{\chi^2}{a+b+c+d}}$.

11. Algunos conceptos de Estadística

La **media poblacional** se define como:

$E[X_i] = \int x_i dF(x_i) = \mu_i$

$$E(\mathbf{x}) = \begin{pmatrix} E(x_1) \\ E(x_2) \\ \vdots \\ E(x_p) \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_p \end{pmatrix}$$

La **varianza poblacional**, cuando existe,se define como:

$Var[X_i] = \int (x_i - \mu_i)^2 f(x_i) dx_1 = \sigma_i^2 = \sigma_{ii}^2$

La **covarianza poblacional** se define como:

$cov(x_i, x_j) = \int \int (x_i - \mu_i)(x_j - \mu_j) f(x_i, x_j) dx_i dx_j = \sigma_{ij}$

Estos valores se presentan dentro de la matriz de varianzas y covarianzas Σ

$$\Sigma = E[(X - E[X])(X - E[X])'] = var(x)$$

$$= \begin{bmatrix} E[(X_1 - \mu_1)(X_1 - \mu_1)] & \cdots & E[(X_1 - \mu_1)(X_p - \mu_p)] \\ E[(X_2 - \mu_2)(X_1 - \mu_1)] & \cdots & E[(X_2 - \mu_2)(X_p - \mu_p)] \\ \vdots & \vdots & \vdots \\ E[(X_p - \mu_p)(X_1 - \mu_1)] & \cdots & E[(X_p - \mu_p)(X_p - \mu_p)] \end{bmatrix}$$

$$= \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1p} \\ \sigma_{21} & \cdots & \sigma_{2p} \\ \vdots & \vdots & \vdots \\ \sigma_{p1} & \cdots & \sigma_{pp} \end{bmatrix}$$

La **media muestral** de la j-ésima variable está dada por
$\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_{ik}$
Denotaremos al conjunto de las medias en un vector de medias muestrales
$\mathbf{x}' = (\bar{x}_1, \ldots, \bar{x}_p)$

La **varianza muestral** de la k-ésima variable se calcula como:
$S_{kk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ik} - \bar{x}_k)^2$

La **covarianza** entre la j-ésima variable y la k-ésima variable esta dada por
$S_{lk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{il} - \bar{x}_l)(x_{ik} - \bar{x}_k)$
La **matriz de covarianzas muestral** denotada por **S**, contiene a las varianzas y covarianzas.

$$\mathbf{S} = \frac{\mathbf{1}}{\mathbf{n-1}} \sum_{\mathbf{i=1}}^{\mathbf{n}} (\mathbf{x_i} - \bar{\mathbf{x}})(\mathbf{x_i} - \bar{\mathbf{x}})' = \begin{bmatrix} {S_1}^2 & \ldots & S_{1p} \\ & \ddots & \vdots \\ & & {S_p}^2 \end{bmatrix}.$$

12. Análisis De Conglomerados

El objetivo de este análisis es formar grupos de observaciones, de manera que todas las unidades en un grupo sean similares entre ellas pero que sean diferentes a aquellas de otros grupos. La parte interesante es definir que es similar. Si por ejemplo el rango de las disimilitudes corre de 0 a C, podríamos definir que dos unidades son consideradas similares si su disimilitud es menor a $\frac{1}{2}C$.

Hay métodos **jerárquicos y no- jerárquicos**. Dentro de los jerárquicos están los aglomerativos y los divisivos.

Los métodos jerárquicos aglomerativos dan origen a un gráfico llamado **dendograma** . Estos método son iterativos y en cada paso debe recalcularse la matriz de distancias, en los casos en que se tienen muchas observaciones esto puede llevar mucho tiempo de cómputo.

A continuación se presentan varios de estos métodos:

- **Liga sencilla o del vecino más cercano**: La distancia entre dos conglomerados, es la mínima de todas las posibles distancias entre objetos que pertenecen a distintos conglomerados. Se unen los dos conglomerados más cercanos. Este método tiende a tener un buen desempeo cuando hay grupos de forma elongada, conocidos como tipo cadena.

- **Liga completa o del vecino más lejano**: La distancia entre dos conglomerados, es la máxima de todas las posibles distancias entre objetos que pertenecen a distintos conglomerados. Se unen los dos conglomerados más cercanos.Se desempea bien cuando los conglomerados son de forma circular.

- **Liga promedio**: La distancia entre dos conglomerados, es el promedio de todas las posibles distancias entre objetos que pertenecen a distintos conglomerados.Se unen los dos conglomerados más cercanos. Funciona bien para conglomerados tanto tipo cadena como circulares.

- **Liga centroide**: La distancia entre dos conglomerados está definida como la diferencia entre las medias (centroide) de cada conglomerado.Se unen los dos conglomerados más cercanos.

- **Liga mediana**: La distancia entre dos conglomerados está definida como la diferencia ponderada entre las medias (centroide) de cada conglomerado, la ponderación está dada por el tamao (número de unidades del conglomerado) de los conglomerados.Se unen los dos conglomerados más cercanos. Se espera un mejor rendimiento que el del centroide cuando los grupos varían mucho en tamao.

- **Método de Ward**: Se unen los dos conglomerdos que arrojen la menor varianza intragrupo (witih group) (con respecto al centroide de cada conglomerado).

Sólo las dos primeras ligas son invariantes ante transformaciones monótonas de las disimilitudes d_{ij} En algunos caso se puede trabajar con sólo la matriz de disimilitudes, en otros se requieren los datos originales.

Se construye el dendograma y se hace un corte a la altura máxima de la disimilitud que se fije para considerar a dos sujetos como similares.

En los métodos divisivos R tiene uno llamado DIANA.

Entre los no jerárquicos están:

- **k medias (k means)**: se requiere de antemano dar el número de grupos existente. Es un método iterativo, por lo que requiere de una solución inicial, y de allí se va optimizando una función objetivo. En cada paso se reasignan los elementos del conglomerado de manera tal que la suma de distancias al cuadrado de los puntos al centro de su conglomerado sea mínima.

- **basado en un modelo de mezcla de normales**: éste considera además diferentes estructuras de covarianza, es un método bayesiano, y usa estimación tipo EM.

- **de dos pasos**: se recomienda cuando se tienen muchos datos.

    1. **Formación de Preconglomerados**. La idea es reducir el tamaño de la matriz de distancias entre todos los posibles elementos. Los preconglomerados son conglomerados de los datos originales que se usan después en un método jerárquico. Se lee un caso, el algoritmo decide si se une a los conglomerados ya existentes, o forma un nuevo conglomerado. Una vez acabado el proceso de preaconglomeramiento, cada conglomerado es tratado como un elemento, la nueva matriz de distancias se reduce al número de preconglomerados.
    2. **Agrupación de preconglomerados**. Se usa un método jerárquico estándar usando como elementos a los preconglomerados.

En general cada método puede llevar a una agrupación diferente. Se recomienda usar varios métodos para identificar los individuos que brincan de un grupo a otro según el método usado, tambien para definir los tamaños de los grupos.

13. Análisis de discriminante

A diferencia del análisis de conglomerados aquí se busca construir una regla de asignación de individuos a distintos grupos **ya dados**, entonces la muestra que se tiene puede considerarse como de entrenamiento, y cuando lleguen nuevos individuos ya se tiene la manera de asignarlos a uno de los grupos.

Veremos el análisis de **discriminante lineal** y el **discriminante cuadrático**. Para su derivación se supondrá que los datos son normales.

El primer caso, el de discriminante lineal considera que los grupos tienen distintas medias pero **comparten la misma matriz de varianzas y covarianzas**. Se verá que esto llevará a construir hiperplanos que corten al espacio de observaciones en los distintos grupos. En el segundo caso se considera que cada grupo tiene su propia matriz de varianza y covarianza, aquí los cortes en vez de hacerse de manera plana se hacen con curvas. Si se tiene un número g de grupos entonces se requieren $g-1$ hiperplanos (curvas) para partir al espacio.

13.1. Análisis de discriminante lineal

cuando hay dos grupos

Suponemos que la población π_i es normal multivartiada de dimensión p con $i = 1, 2$

$$f_i(x) = (2\pi)^{-p/2}|\Sigma_i|^{-1/2}\exp[-1/2(x-\mu_i)'\Sigma_i^{-1}(x-\mu_i)]$$

$$\frac{f_1(x)}{f_2(x)} = |\Sigma_1|^{-1/2}|\Sigma_2|^{1/2}\exp[-1/2\{x'(\Sigma_1^{-1}-\Sigma_2^{-1})x-2x'(\Sigma_1^{-1}\mu_1-\Sigma_2^{-1}\mu_2)+\mu_1'\Sigma_1^{-1}\mu_1-\mu_2'\Sigma_2^{-1}\mu_2\}]$$

En el caso de que $\Sigma_1 = \Sigma_2 = \Sigma$ ocurre que

$$|\Sigma_1|^{-1/2}|\Sigma_2|^{1/2} = 1,$$

$$x'(\Sigma_1^{-1} - \Sigma_2^{-1})x = 0,$$

$-2x'(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2) = -2x'(\Sigma^{-1})(\mu_1 - \mu_2)$ y

$$\text{y}\mu_1'\Sigma_1^{-1}\mu_1 - \mu_2'\Sigma_2^{-1}\mu_2 = (\mu_1+\mu_2)'(\Sigma_1^{-1})(\mu_1-\mu_2)$$

Entonces

$$\frac{f_1(x)}{f_2(x)} = \exp[-1/2\{(-2x+(\mu_1+\mu_2))'(\Sigma^{-1})(\mu_1-\mu_2)$$

Cuando $\frac{f_1(x)}{f_2(x)} \geq k$ donde $k \geq 1$ concuerda que debemos asignar a x a la población π_1.

Si $k = 1$ esta regla también puede escribirse como:

$$x'(\Sigma^{-1})(\mu_1-\mu_2) \geq 1/2(\mu_1+\mu_2)'(\Sigma^{-1})(\mu_1-\mu_2)$$

Esta expresión es una función lineal en términos de x y de allí su nombre.

NOTA A esta misma expresión llegó Fisher **sin suponer la normalidad**, sólo hallando la dirección en la que la proyección de los centros de ambos grupos se hace lo más alejada posible. Entonces cuando los datos se alejan mucho de la normalidad hace que la función discriminante no funcione muy bien.

¿Y si hay más grupos?

Se tienen varias posibles reglas de clasificación por parejas, algunas de ellas resultan **redundantes**, en realiidad con g grupos sólo $g-1$ son necesarias.

13.2. Análisis de discriminante cuadrático

Cuando $\Sigma_1 \neq \Sigma_2$ el término $x'(\Sigma_1^{-1} - \Sigma_2^{-1})x$ que está dentro la exponencial no se cancela y este es un término **cuadrático** y da lugar a que la regla de asignación involucre a una **curva**.

Una vez creada la regla de asignación, conviene construir medidas para ver que tan buena es, interesa entonces conocer cuántos de la muestra quedan bien o mal clasificados, también se usan medidas de desempeño tipo *jack knife*, y tipo validación cruzada.

Además de los discriminantes lineal y cuadrático, se puede utilizar la regresión logística para construir una regla de asignación cuando se trata de dos grupos sólamente o también se puede usar algún método bayesiano.

13.3. Análisis de discriminante Logístico

En un anális de regresión logística se tiene un modelo de la forma:

$$P(i \in \pi_1|\mathbf{x}_i) = \frac{\exp(\mathbf{x_i'}\beta)}{1+\exp(\mathbf{x_i'}\beta)} = \frac{\exp(\beta_o + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}{1+\exp(\beta_o + \beta_1 x_{i1} + \ldots + \beta_p x_{ip})}$$

y

$$P(i \in \pi_2|\mathbf{x}_i) = \frac{1}{1+\exp(\mathbf{x_i'}\beta)}$$

Una vez que se estiman las βs, se calculan las probabilidades y se asigna a la poblacion π_1 si $P(i \in \pi_1|\mathbf{x}_i) > 0{,}5$ y ala población π_2 en caso contrario.

13.4. Análisis de discriminante basado en funciones de probabilidad

Cuando es posible tener funciones de probabilidad f_1 y f_2, entonces la regla discriminante para x es: asignar a la población π_1 si $f_1(x) > f_2(x)$, ahora si se conocen las probabilidades *apriori* q_1 la probabilidad de que x provenga de la población π_1 y q_2, la de provenir de π_2 con $q_1 + q_2 = 1$ entonces asignar x a la población π_1 si

$$\frac{f_1(x)}{f_2(x)} > \frac{q_2}{q_1}$$

Es decir la observación se asigna a la población que tenga la verosimilitud más alta.

La probabilidad de clasificación errónea sería:

$$q_1 P(2|1) + q_2 P(1|2)$$

donde $P(2|1)$ y $P(1|2)$ son las probabilidades de clasificación errónea de cada población.

Usando teorema de Bayes la probabilidad *aposteriori* de que un individuo con valores observados x_o provenga de la población π_i es

$$q(\pi_i|x_o) = \frac{q_i f_i(x_o)}{q_1 f_1(x_o) + q_2 f_2(x_o)},$$

entonces se debe asignar a x_o a la población que tenga la probabilidad *aposteriori* más alta.

NOTA

Si de las $f_i(x)$ no se sabe gran cosa éstas deben ser **estimadas**, por ejemplo con estimadores **no paramétricos** tipo kernel, tipo *spline*, suavizamientos,etc.

No es raro encontrar en ciertos estudios se usa primero un análisis de conglomerados y después uno de discriminante para ver que tan bien quedaron formados los grupos.

14. Comandos en R

Hist, boxplot, stem, script.., barplot, pieplot, parallel, stars, faces, Andrews curves, bagplot. Dist, Mclust, hclust, Diana, clara, kmeans,agnes

dicrim