Capítulo 1

Tipificar Datos

Para tipificar (*"estandarizar"*) los datos en la forma habitual puede invocarse a la función **standard**.

```
standard(X)
```

Esta función toma la matriz de datos X y opera por columnas restando la media y dividiendo por el desvío: $\frac{x_{ij} - \bar{x}_j}{s_j}$.

Ademas de la operación se agregan dos propiedades útiles:

- Si se quiere tipificar datos que están en una estructura de "data.frame", el resultado tipificado también pertenecerá a la misma *clase* y conservará los *atributos*.
- La forma particular de tratamiento de los valores faltantes que es: *'pairwise.complete.obs'* que se describe a continuación desde la ayuda del **R**.

help(var)

...

If 'use' is 'all.obs', then the presence of missing observations will produce an error. If 'use' is 'complete.obs' then missing values are handled by casewise deletion. **Finally, if 'use' has the value 'pairwise.complete.obs' then the correlation between each pair of variables is computed using all complete pairs of observations on those variables.** This can result in covariance or correlation matrices which are not positive semidefinite.

The denominator n - 1 is used which gives an unbiased estimator of the (co)variance for i.i.d. observations. These functions return 'NA' when there is only one observation ...

1.1. Uso

```
>DatosST <- standard(DATOS[,3:9])
```

Capítulo 2

Análisis de Grupos

El Análisis de Grupos, como se ve en el curso de Multivariado I, consta del uso de un *script* que utiliza básicamente las tres funciones descritas en las siguientes secciones. Este archivo se llama "cluster.R" y no está en la biblioteca "multivar", pues debe correrse en forma interactiva haciendo las modificaciones pertinentes para cada caso.
Recuerdese que el tipo de análisis que se desarrolla es *jerárquico agregativo* y que se utilizan un conjunto de "reglas de parada" o "pseudoindicadores" para decidir el número de grupos. El **R** ya posee un conjunto de funciones para realizar Análisis de Grupos de distintos tipos basado en el libro de Kaufman y Rousseeuw (1990).

2.1. agnes

Esta función pertenece al paquete *cluster* y computa un análisis de grupos aglomerativo jerárquico de datos, se describe el capitulo 5 de Kaufman y Rousseeuw (1990)[1].

Entrada

$agnes(x, diss = inherits(x, "dist"), metric = "euclidean", stand = FALSE,$
$method = "average", par.method, keep.diss = n < 100, keep.data = !diss)$

Los parámetros que pueden pasarse son los siguientes:

x : matriz de datos o "data frame", o matriz de disimilaridad, dependiendo del valor del argumento 'diss'. Las variables deben ser cuantitativas. El caso habitual será el primero.

[1]Tomado de la ayuda del **R**.

diss Argumento que toma un valor lógico: si es TRUE (valor por defecto) se asume que "x" es una matriz de disimilaridad. Si es FALSE "x"es tratada como una matriz de observaciones por variables. Generalmente se pondrá **FALSE**.

metric : "character string" que especifica la metrica a ser usada para calcular las disimilaridades entre observaciones. Las opciones disponibles son"euclidean" y "manhattan". Si "x" ya es una matriz de disimilaridad este argumento será ignorado.

stand : Argumento lógico, si es TRUE entoces la "x" son tipificadas (en la forma habitual) antes de calcular las disimilaridades. Si "x" ya es una matriz de disimilaridad este argumento será ignorado.

method : "character string" que define el metodo de agrupamiento. Los seis métodos implementados son:

- "single" (vecino mas cercano),
- "complete" (vecino mas lejano),
- "ward" (método de Ward),
- "average" (group average method),
- "weighted" (weighted average linkage)
- "flexible" generalización del anterior ... uses (a constant version of) the Lance-Williams formula ...

par.method : Si el "method == "flexible"" ...

eep.diss, keep.data : logicals indicating if the dissimilarities and/or input data 'x' should be kept in the result. Setting these to 'FALSE' can give much smaller results and hence even save memory allocation_time_.

Salida

La salida de la función es un objeto de clase "agnes" que es una *lista* con los siguientes componentes:

order un vector que da el orden de entrada a la jerarquia o de agrupamiento

height vector con las distancias en la unión de los grupos en las sucesivas etapas.

ac Coeficiente de aglomeración que mide la estructura de grupo de los datos. Para cada observación i, se denota $m(i)$ su disimilaridad con el primer grupo con el cual se une, dividido por la disimilaridad en el último paso del algoritmo. El "ac" es el promedio de todos los $1 - m(i)$. Como "ac" crece

con el número de observaciones, esta medida no debiera ser utilizada para comparar datos de tamaño muy diferente.

merge matriz de $(n-1) \times 2$ con n numero de observaciones. La fila i describe la unión de grupos al nivel i de la jerarquia. Si un número j en la fila es negativo la observación (individuo) j es unido a ese nivel. Si j es positivo la unión es con el grupo formado en el paso j tdel algoritmo.

diss un objeto de clase "dissimilarity" que representa la matriz de disimilaridad de los datos.

order.lab un vector similar a "order" pero conteniendo las etiquetas de las observaciones en vez de sus números, si los datos originales estaban etiquetados.

call como (con que argumentos) fue la llamada a la función "agnes".

method que método se dio como argumento de entrada.

una matriz que contiene los datos originales o tipificados dependiendo de la opción "stand" de la función "agnes". Si la entrada fue una matriz de disimilaridades este componente no estará disponible.

De esta salida lo mas importante, desde la forma de ver las cosas en este curso, será el elemento 4("merge") de esta lista de salida. Ese elemento 4 será el insumo para el calculo de los "pseudo-indicadores".

Uso

```
## Ejecuta la funcion agnes sobre los datos tipificados y los carga en AGRUPO
### en general aca los unico que se modificara sera:
   #  method = 'ward', 'single','complete','average'.

> AGRUPO<-agnes(DATOSst, metric = "euclidean", stand = FALSE, method = "ward")
```

2.2. Indicadores

Esta función permite generar e imprimir los pseudoindicadores o reglas de detención que ayudarán a determinar el número de grupos con el que se desarrollará el análisis. La versión actual es modificada sobre código de Oscar Gutierrez y Andrés Castrillejo (2003).

$$indicadores(agnes4, datos, imprime = p)$$

Entrada

Los parámetros que pueden pasarse son los siguientes:

- Parámetros Obligatorios

agnes4 UNIONES (parte 4 o $names= 'merge') de un *objeto de clase agnes*

datos nombre de los datos sobre los que se ejecutó *agnes*. Una matriz con variables cuantitativas, también puede ser un subgrupo de la forma x[,3:12].

- Parámetros Opcionales (ya tienen un valor por defecto)

imprime Cantidad de lineas de la historia que IMPRIME (por defecto TODAS)

Salida

Se imprime una tabla con:

history Historia de las uniones a cada nivel de la jerarquía

Freq Frecuencia del grupo fromado a ese nivel.

Rcuad $pseudo - R^2$

psF $pseudo - F$

psT $pseudo - t^2$

Uso

```
## Ejecuta la funcion indicadores y guarda en IND
### aca se pueden tocar todos los parametros:
### AGRUPO[4] _ es el elemento 4 o \$merge del objeto generado a traves de agnes
### DATOSst   _ es el conjunto de datos sobre los que ejecute agnes
### p _ es el numero de pasos de la jerarquia que queremos ver impresos

> IND <- indicadores(AGRUPO[4],DATOSst,imprime=20)
```

2.3. cutree

Esta función pertenece al paquete *stats* y genera un árbol de datos. Por ejemplo, tomando resultados desde "hclust" corta un árbol en varios grupos de acuerdo a un número especificado de grupos deseados o en base a las distancias [2].

$$cutree(tree, k = NULL, h = NULL)$$

Entrada

tree : Se produce por "hclust". "cutree()" espera una lista con los componentes "merge", "height" y "labels" cada una con el contenido adecuado. En este caso se utilizará el elemento "merge"([4]) de la lista de salida de "agnes".

k : un entero escalar o vector con el número deseado de grupos.

h : escalar o vector con las alturas en donde el árbol debe ser cortado.

Al menos uno "k" o "h" debe ser especificado. En este caso se utilizará "k".

Salida

Retorna un vector con las membresías de cada individuo al número de grupos especificados.

Uso

```
##  ejecuta la funcion cutree y guarda la membresia de cada individuo en cl
### AGRUPO[4] _ es el elemento 4 o \$merge del objeto agnes
### k es el numero de grupos decidido a partir de indicadores

cl <- as.factor(cutree(AGRUPO[4],k) )


### se genera la variable cl ya como factor, dado q es categ\'orica.
```

2.4. Ejemplo

La idea aca es desarrollar un ejemplo completo....PA'l FUTURO

[2] Tomado de la ayuda del **R**.