Análisis Multivariante con Stata

Diploma Avanzado en Metodología de Investigación en Ciencias de la Salud

EVES. Valencia, Oct 2013

Francisco García García
fgarcia@cipf.es

Indice

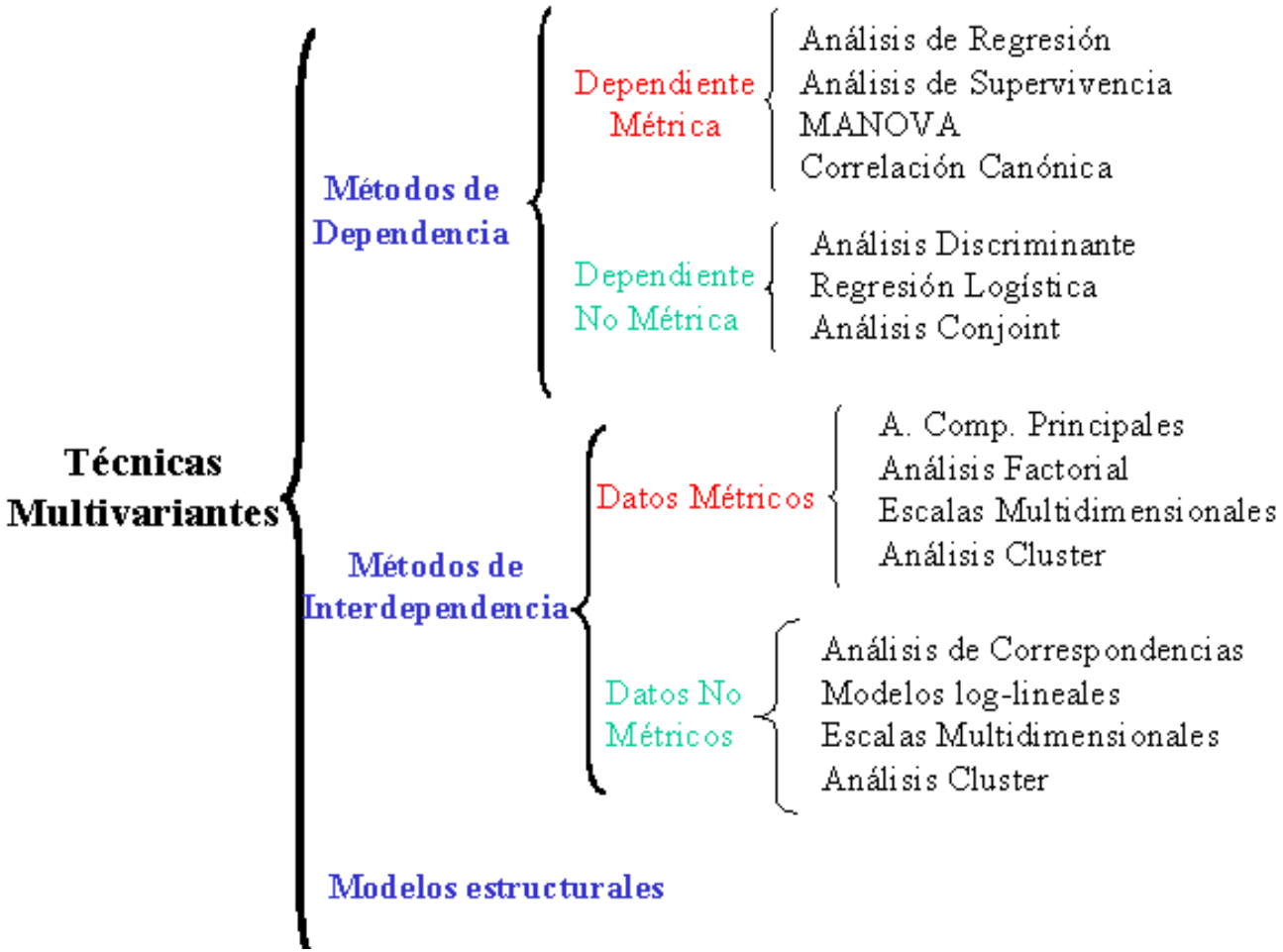
Algunas preguntas...

1. ¿De qué forma se pueden agrupar a los pacientes que ingresan en un hospital según los recursos que consumen?
2. ¿Qué criterios pueden ayudar a diagnosticar si una obstrucción de vías biliares está provocada por un tumor maligno o es de naturaleza benigna?
3. ¿Cómo se puede obtener un indicador de necesidad de servicios sanitarios en distintas unidades geográficas?

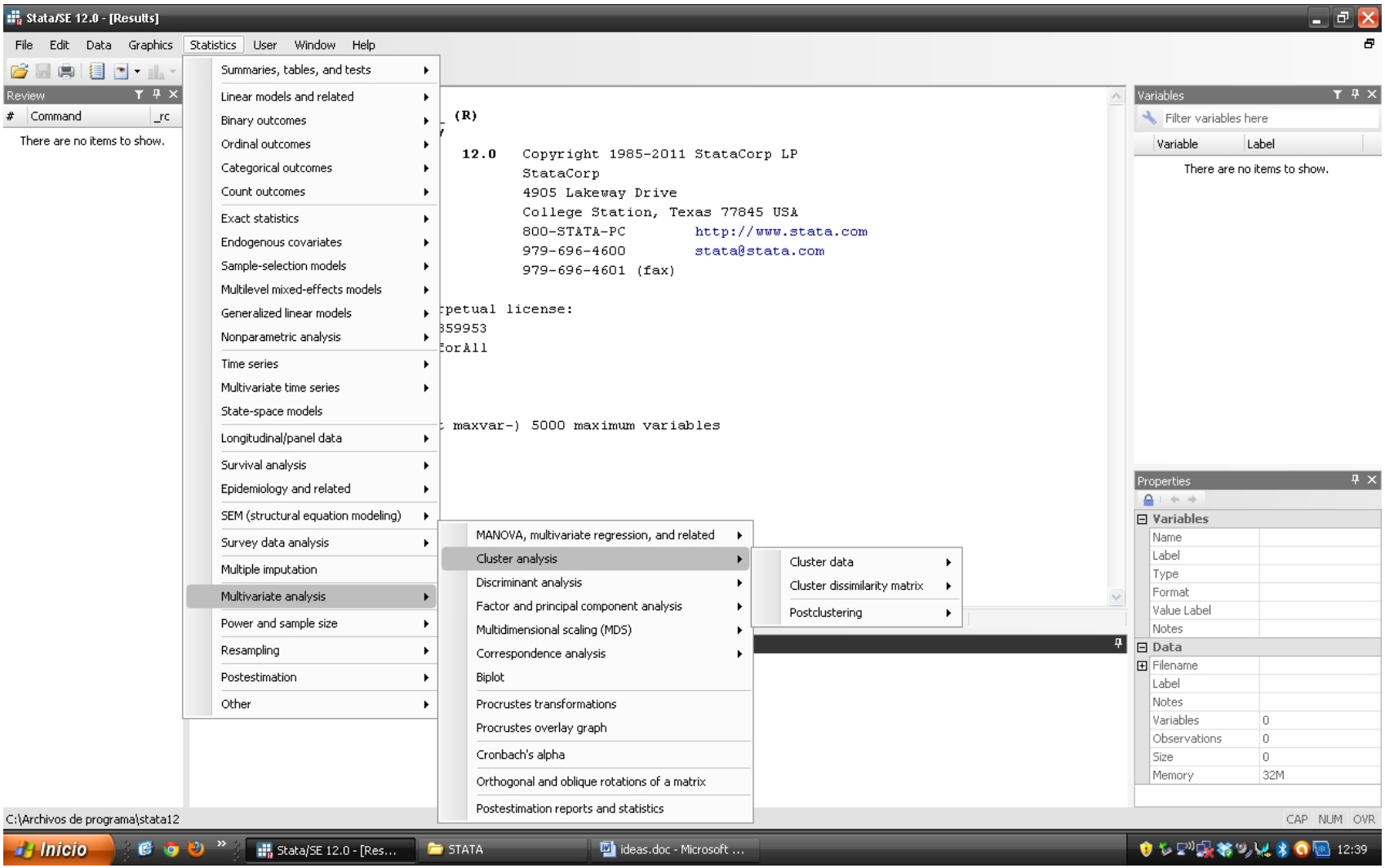
¿Qué tienen en común las preguntas anteriores?

- Estas cuestiones tienen en común que sus respuestas se basan en el **análisis conjunto de muchas variables**.
- El **Análisis Multivariante** presenta métodos que analizan conjuntamente **varias variables**, medidas sobre un **grupo de individuos u objetos**.
- Los resultados del Análisis Multivariante proporcionará una información interesante para la **toma de decisiones** del investigador.

Clasificación de métodos multivariantes:



Análisis Multivariante con Stata



Nos centraremos en los siguientes métodos:

1. **Análisis Clúster**.
2. **Análisis Discriminante**.
3. **Análisis Factorial**.

Análisis Clúster

- Su **objetivo** es formar grupos de objetos (individuos) homogéneos respecto a una variedad de atributos que pueden ser tanto cualitativos como cuantitativos, de forma que las observaciones pertenecientes a un grupo sean muy similares entre sí y muy disimilares del resto.

- A diferencia del Análisis Discriminante se desconoce el número y la composición de dichos grupos. El Análisis de Clustering busca la formación de grupos mientras que el Análisis Discriminante predice la pertenencia a grupos ya prefijados.

Ejemplos:

1. Detección de subgrupos de pacientes con cáncer de mama en función de variables clínicas y genéticas.

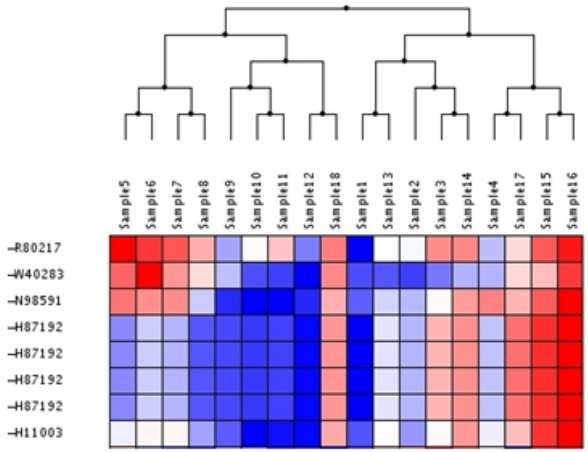
2. Agrupar diferentes frutas y verduras por sus características nutricionales (energía, proteínas, lípidos, glúcidos...)

Los tipos de clustering variarán en función de 3 criterios:

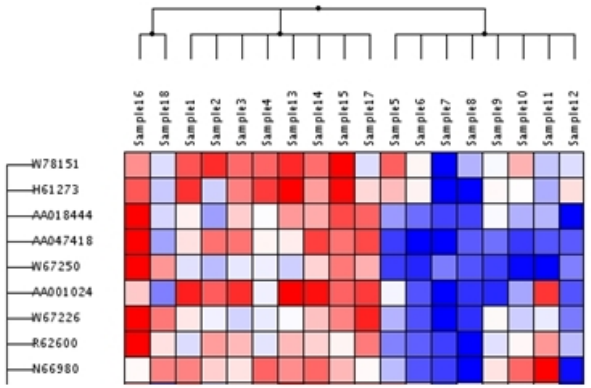
1. ¿Jerárquico o no jerárquico?
2. ¿Qué método utilizamos para medir la similitud de los grupos?
3. ¿Qué medida de distancia usaremos?

1. Jerárquico vs. no jerárquico

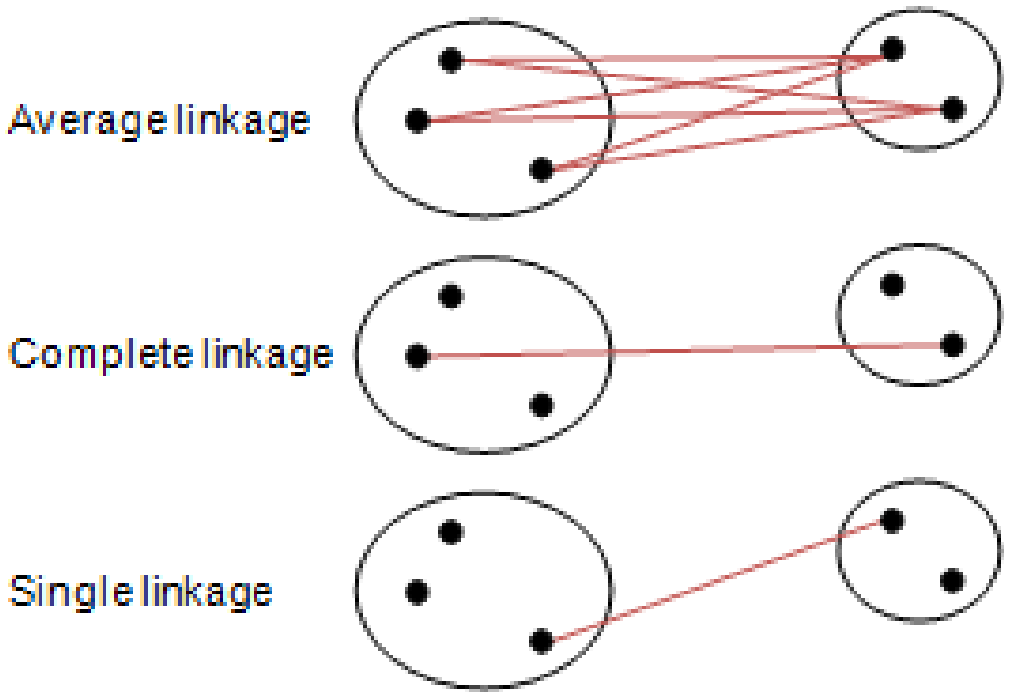
Cluster jerárquico



Cluster NO jerárquico



2. Método para medir la similitud entre grupos.



2. Método para medir la similitud entre grupos.

Single linkage
- Nearest-neighbor method
- Minimum method
- Hierarchical analysis
- Space-contracting method
- Elementary linkage analysis
- Connectedness method

Complete linkage
- Furthest-neighbor method
- Maximum method
- Compact method
- Space-distorting method
- Space-dilating method
- Rank-order typal analysis
- Diameter analysis

Average linkage
- Arithmetic-average clustering
- Unweighted pair-group method using arithmetic averages
- UPGMA
- Unweighted clustering
- Group-average method
- Unweighted group mean
- Unweighted pair-group method

Weighted-average linkage
- Weighted pair-group method using arithmetic averages
- WPGMA
- Weighted group-average method

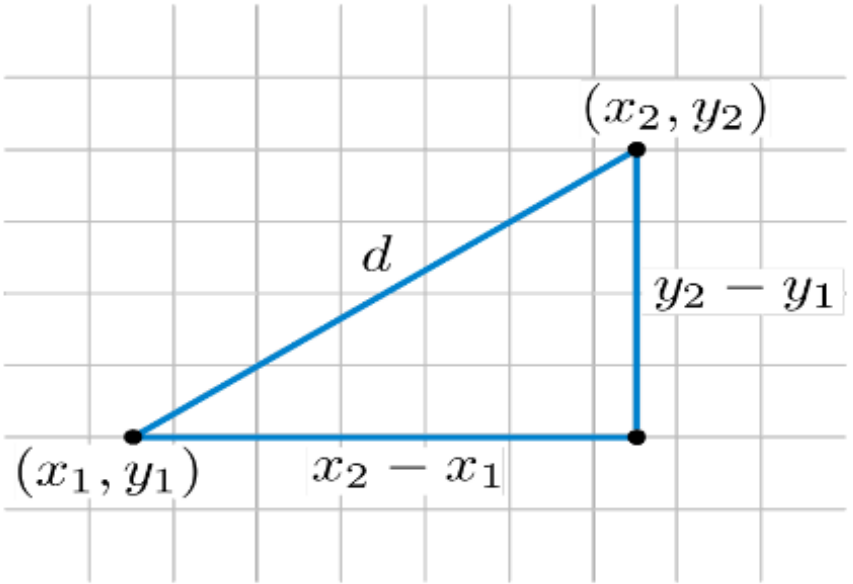
Centroid linkage
- Unweighted centroid method
- Unweighted pair-group centroid method
- UPGMC
- Nearest-centroid sorting

Median linkage
- Gower's method
- Weighted centroid method
- Weighted pair-group centroid method
- WPGMC
- Weighted pair method
- Weighted group method

Ward's method
- Minimum-variance method
- Error-sum-of-squares method
- Hierarchical grouping to minimize tr(W)
- HGROUP

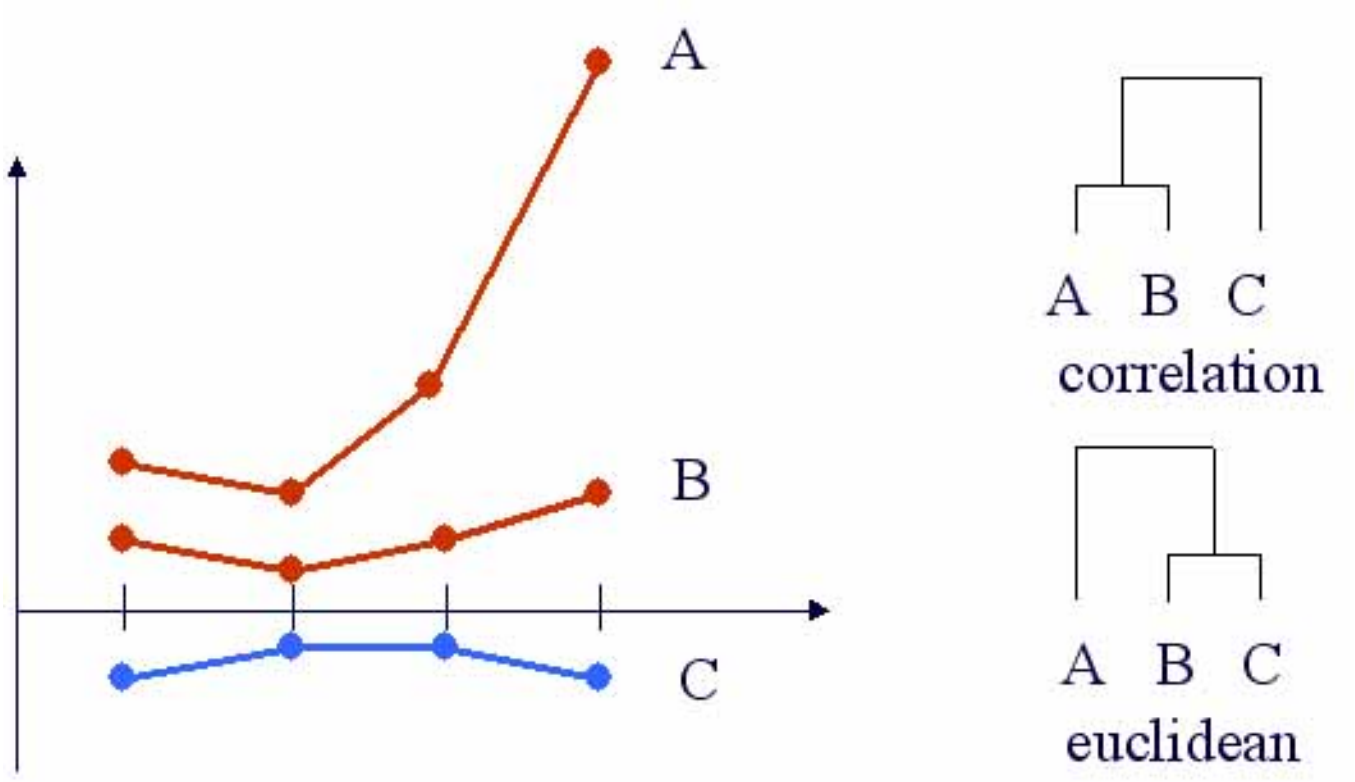
3. Tipo de medida de distancia utilizada.

Distancia euclídea

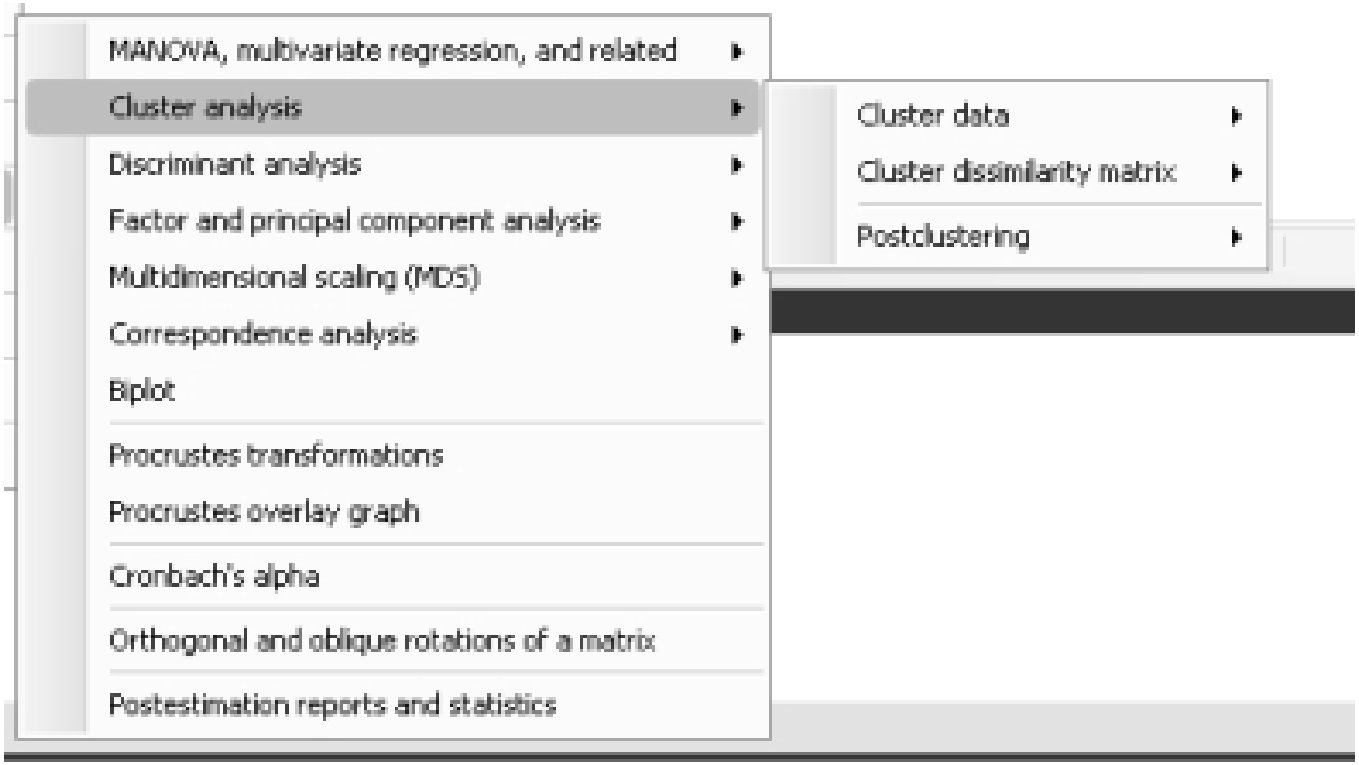


$$d_E(P_1, P_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

3. Tipo de medida de distancia utilizada.



Stata tiene implementado diversos métodos de análisis de clustering:



Clustering en Stata

También en Stata, escogeremos el método de clustering que se ajuste a las contestaciones de las preguntas que ya comentamos:

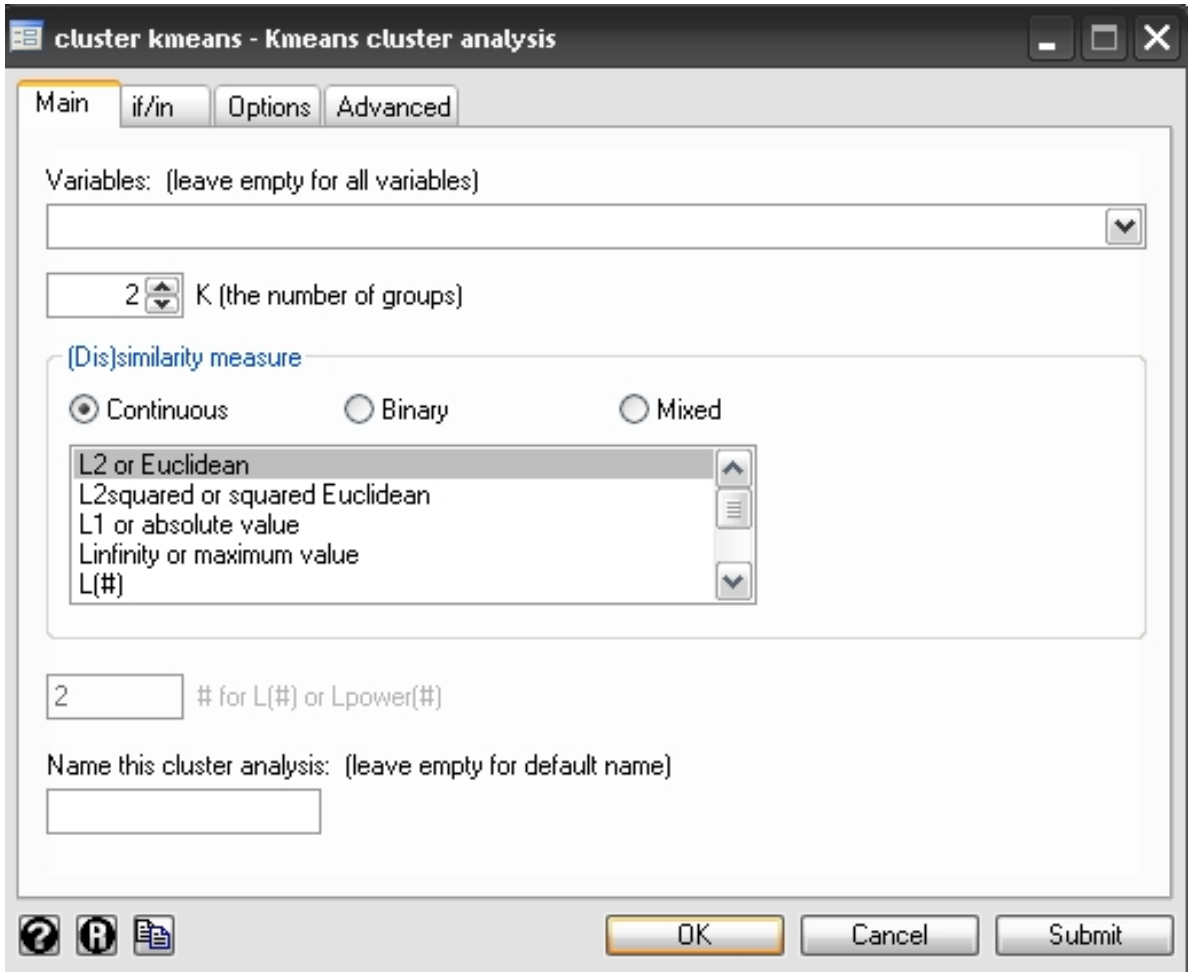
1. ¿Jerárquico o no jerárquico?
2. ¿Qué método utilizamos para medir la similitud de los grupos?
3. ¿Qué medida de distancia usaremos?

Clustering en Stata

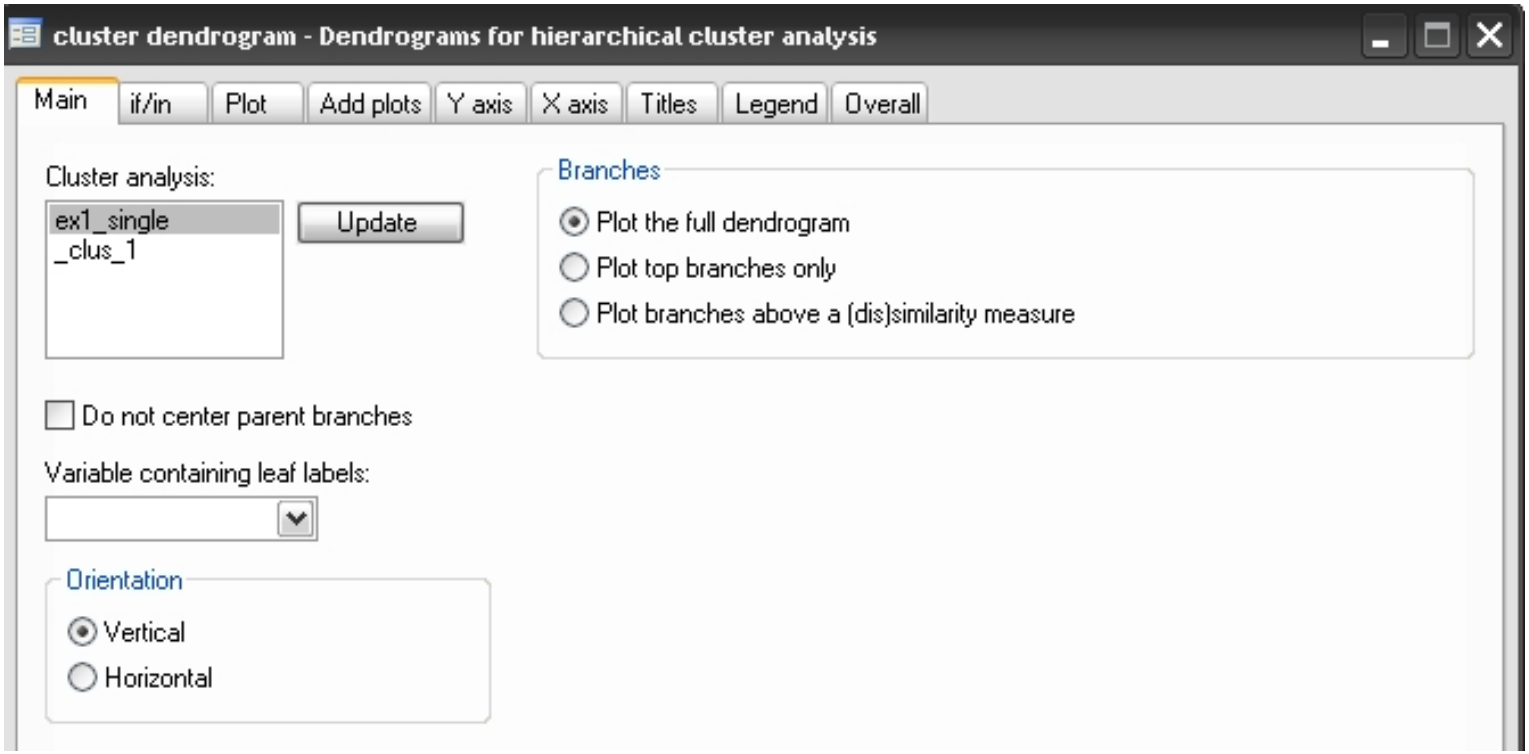
Independientemente del método elegido, seguiremos varios pasos:

1. Generamos el objeto clúster.
2. Visualizamos el dendrograma o árbol de clúster (para algunos métodos no está implementado).
3. Extraemos información de los grupos detectados.

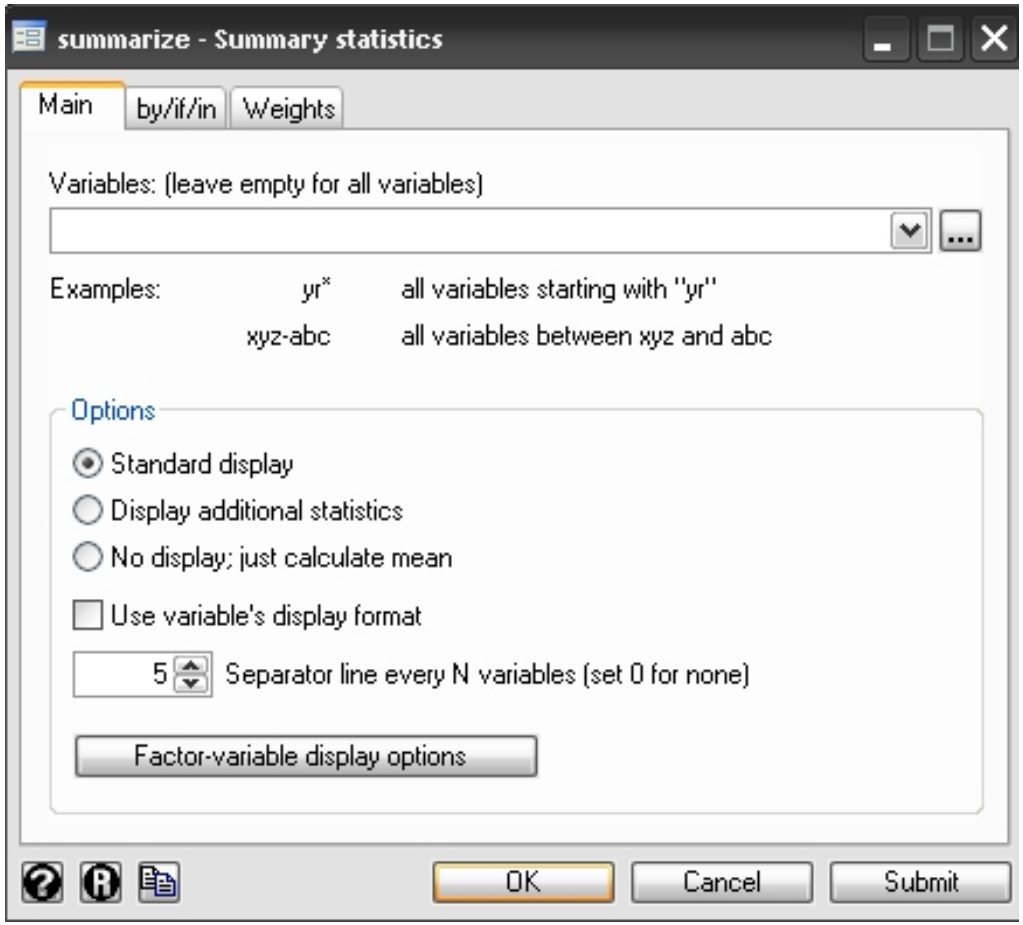
1. Generamos un objeto clúster.



2. Visualizamos el clustering mediante un dendrograma.



3. Extraemos información sobre los grupos detectados



Ejemplo de análisis de clustering con Stata:

- Empezamos trabajando con el dataset **homework.dta** que describe la realización de tareas domésticas de un grupo de 30 personas.
- La **matriz de datos** tiene una dimensión de 30 filas y 61 columnas. Cada fila es una persona y cada variable representa si esa persona realiza o no cada una de las 60 tareas valoradas (**variables binarias**).
- El **objetivo** es determinar grupos parecidos de personas según las actividades domésticas que realicen.
- Para este primer ejemplo contamos con una **información extra**. El investigador nos proporciona previamente a qué grupo pertenece cada individuo, según sus criterios (está recogido en la variable 61 del dataset). Esta referencia será interesante para evaluar como está trabajando nuestro clustering. Por supuesto, no será habitual que dispongamos de esta información al comienzo del estudio!

Ejemplo de análisis de clustering con Stata:

- **Leemos los datos desde Stata.** Tenemos varias posibilidades:

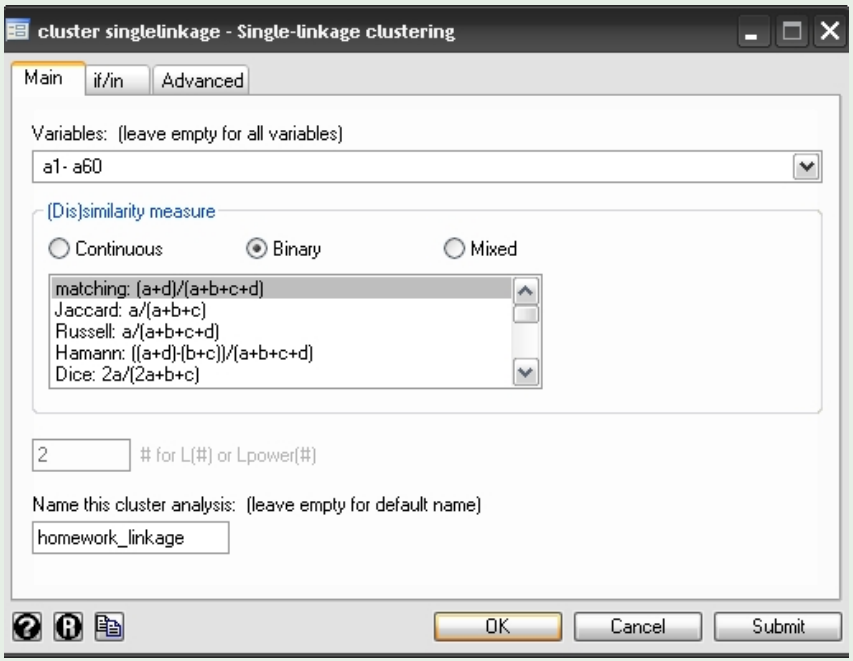
  1. Insertando en la ventana de comandos:
  use http://www.stata-press.com/data/r12/homework.dta
  2. Recuperando el dataset del repositorio de datos de Stata: *File / Example Datasets / Stata 12 Manual Datasets / Multivariate Statistics Reference Manual*
  3. Otra opción es directamente desde *File / Open* y seleccionamos el fichero de datos en la carpeta donde lo tengamos almacenado.

- En cualquier **análisis estadístico** que hagamos, siempre habrá que realizar un **descriptivo** que nos permita conocer nuestros datos:

  Statistics / Summaries, tables and tests.
  Graphics / Scatterplot matrix.

Ejemplo de análisis de clustering con Stata:

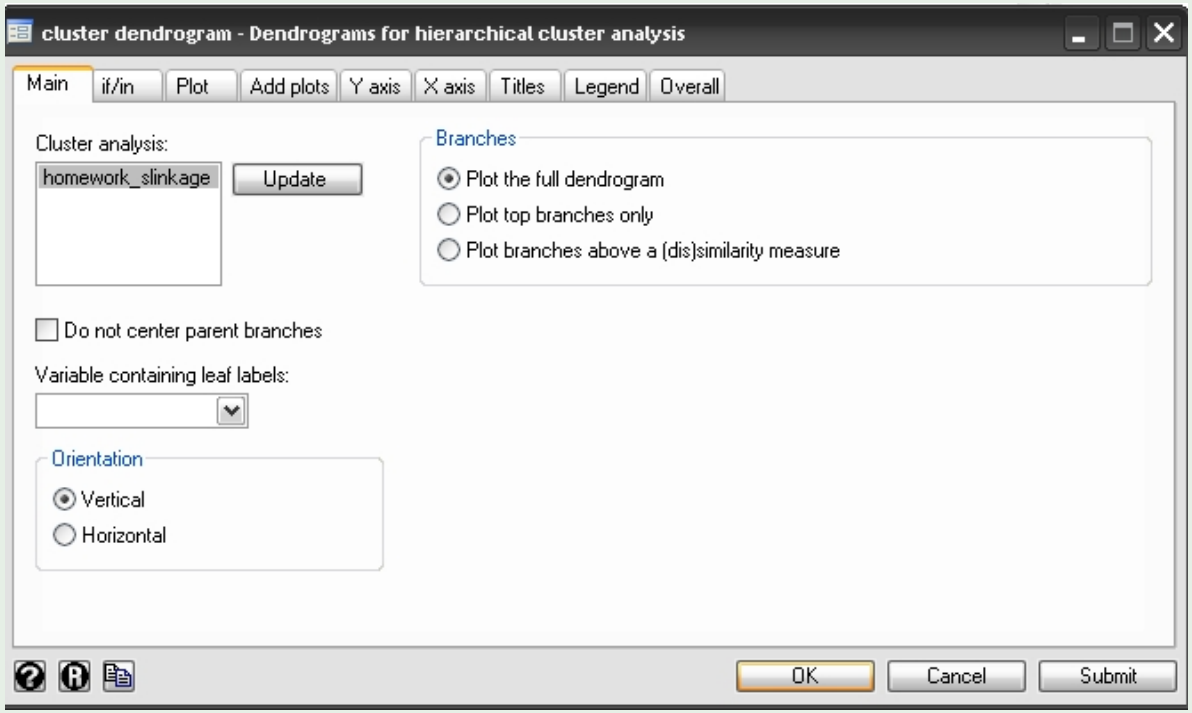
Paso 1. Generamos el objeto cluster:
Statistics / Multivariate Analysis / Cluster Analysis / Cluster Data



Si vamos al editor, observaremos que tenemos 3 nuevas variables: id, ord, hgt (id, orden y altura). Son variables que proporcionan info sobre la construcción del clustering.

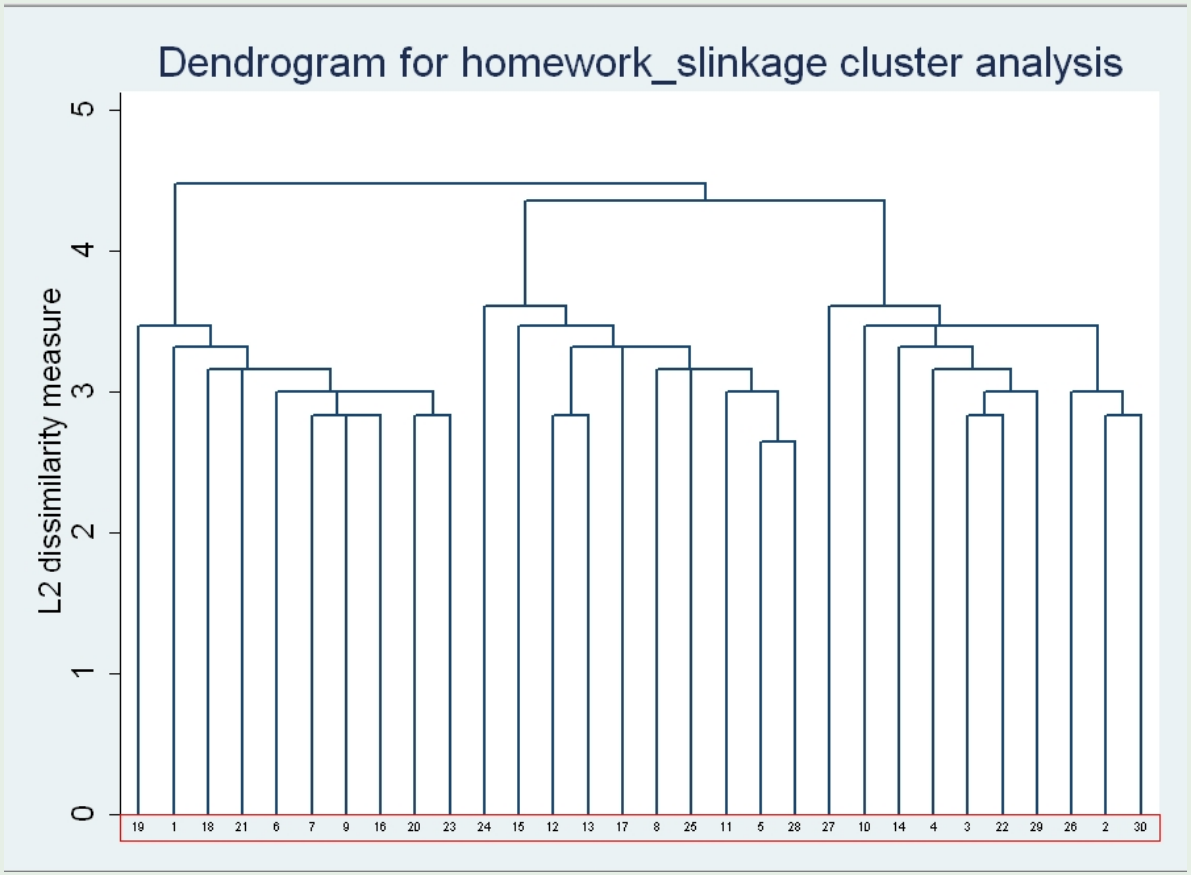
Ejemplo de análisis de clustering con Stata:

Paso 2. Representamos el dendrograma o árbol clúster: Statistics / Multivariate Analysis / Cluster Analysis / Postclustering / Dendrograms



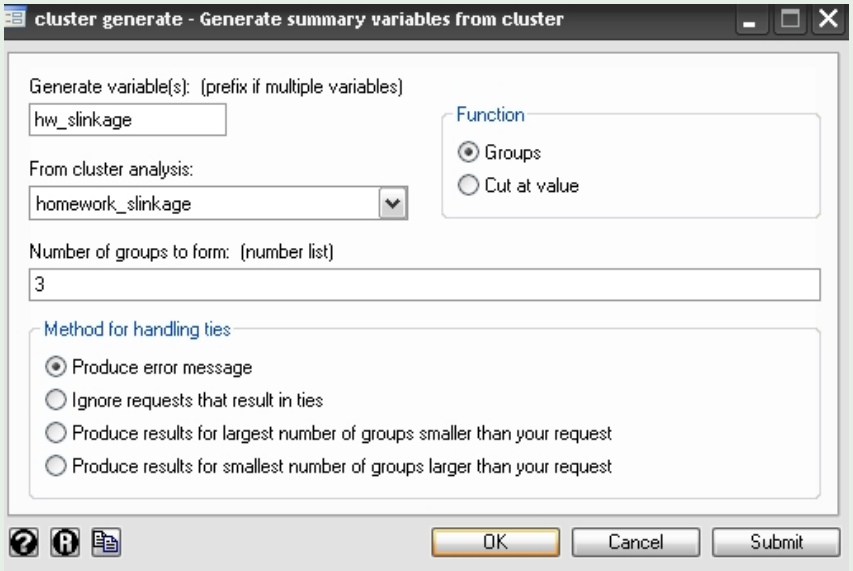
Ejemplo de análisis de clustering con Stata:

Paso 2. Representamos el dendrograma o árbol clúster:
Statistics / Multivariate Analysis / Cluster Analysis / Postclustering / Dendrograms



Ejemplo de análisis de clustering con Stata:

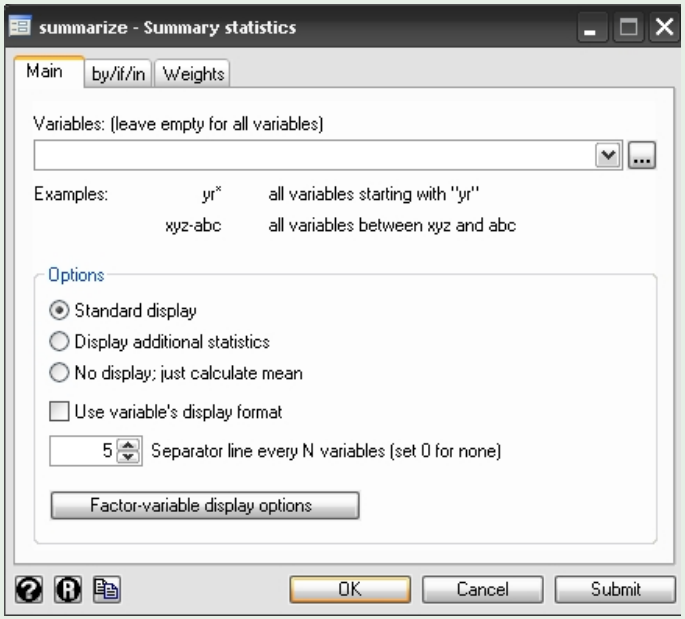
Paso 3. Extracción de información de los grupos generados: Statistics / Multivariate Analysis / Cluster Analysis / Postclustering / Summary variables from cluster analysis



Tras la visualización del dendrograma, podemos decidir el número de clusters con el que queremos trabajar. Hemos escogido 3 y conoceremos que sujetos pertenecen a cada uno de ellos. Esta información quedará incorporada en una nueva variable que se creará.

Ejemplo de análisis de clustering con Stata:

Paso 3. Extracción de información de los grupos generados: Statistics / Summaries, tables and tests / Summary and descriptive statistics / Summary statistics



En la pestaña **by if in** , indicamos la variable que establece los grupos.

Ejemplo de análisis de clustering con Stata:

Paso 3. Extracción de información de los grupos generados: Statistics / Summaries, tables and tests / Summary and descriptive statistics / Summary statistics

```
. by hw_slinkage, sort : summarize
```

-> hw_slinkage = 1

Variable	Obs	Mean	Std. Dev.	Min	Max
a1	10	0	0	0	0
a2	10	0	0	0	0
a3	10	0	0	0	0
a4	10	0	0	0	0
a5	10	0	0	0	0
a6	10	.3	.4830459	0	1
a7	10	.1	.3162278	0	1
a8	10	0	0	0	0
a9	10	0	0	0	0
a10	10	.4	.5163978	0	1
a11	10	0	0	0	0
a12	10	0	0	0	0
a13	10	.7	.4830459	0	1
a14	10	.3	.4830459	0	1
a15	10	.4	.5163978	0	1

--more--

Ejemplo de análisis de clustering con Stata:

Paso 3. Extracción de información de los grupos generados: Statistics / Summaries, tables and tests /Table / Table of summary statistics

```
. table truegrp hw_slinkage
```

truegrp	hw_slinkage 1	2	3
1			**10**
2	**10**		
3		**10**	

Como teníamos la asignación "real" de sujetos a sus grupos iniciales, podemos evaluar como ha realizado nuestro clustering la agrupación de individuos: los resultados son coincidentes!

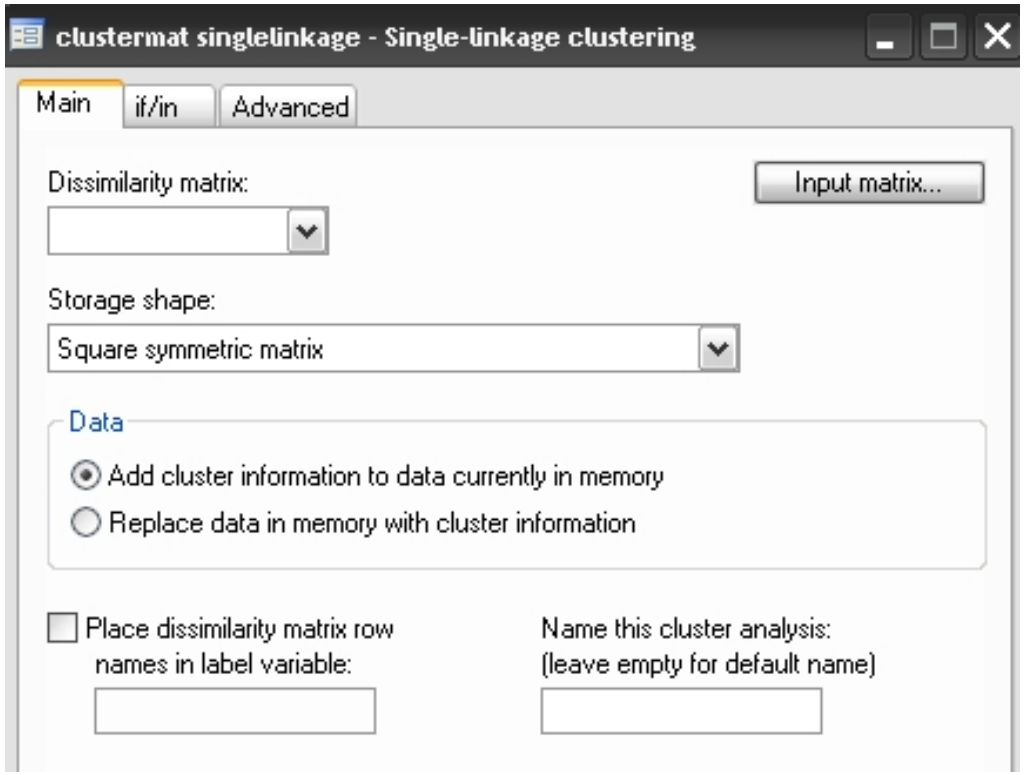
Ejemplo de análisis de clustering con Stata:

Paso 3. Extracción de información de los grupos generados:
Statistics / Multivariate Analysis / Cluster Analysis / Postclustering / Detailed listing of clusters

```
. cluster list homework_linkage
homework_linkage  (type: hierarchical,  method: single,  similarity: matching)
      vars: homework_linkage_id (id variable)
            homework_linkage_ord (order variable)
            homework_linkage_hgt (height variable)
     other: cmd: cluster singlelinkage a1- a60, measure(matching) name(homework_linkage)
            varlist: a1 a2 a3 a4 a5 a6 a7 a8 a9 a10 a11 a12 a13 a14 a15 a16 a17 a18 a19 a20 a21 a22 a23 a24 a25 a26 a27 a28 a29
                 a30 a31 a32 a33 a34 a35 a36 a37 a38 a39 a40 a41 a42 a43 a44 a45 a46 a47 a48 a49 a50 a51 a52 a53 a54 a55 a56 a5
                 a58 a59 a60
            range: 1 0
```

Nos informa de las características del clustering utilizado.

Matrices de similitud/disimilitud



El análisis de clustering también se puede realizar sobre una matriz de similitud / disimilitud.

Ejercicio 1: clustering jerárquico

- Trabajaremos con el dataset **labtech.dta** disponible en *File / Example Datasets / Stata 12 Manual Datasets / Multivariate Statistics Reference Manual*
- Los **datos** representan un parámetro clínico medido en 50 pacientes. Para cada sujeto se determinaron 4 cuantificaciones a lo largo del día y las 4 fueron efectuadas por el mismo técnico.
- De modo que nuestra matriz dispone de 50 filas (pacientes) y 5 columnas (una para cada una de las 4 mediciones realizadas y una quinta columna con el nombre del técnico).
- El **objetivo** del estudio es determinar si hay grupos homogéneos de pacientes considerando las medidas realizadas del parámetro clínico.
- Decidimos arbitrariamente elegir el **clustering single-linkage con la distancia euclídea** que aparece por defecto.

Ejercicio 1: clustering jerárquico

Plan de trabajo:

1. Realiza una **descripción de los datos**. Explora gráficamente la relación entre las 4 mediciones mediante gráficos de dispersión.

2. Realiza el **análisis de clustering de las muestras** obteniendo el correspondiente dendrograma.

3. ¿Hay alguna agrupación clara en el árbol del cluster? ¿Detectas algo extraño en el dendrograma? ¿Alguna explicación?.

   El análisis de clustering también constituye una buena herramienta de exploración de los datos. **Pista:** cuando hagas el dendrograma cambia la variable que se utiliza para dar nombre a las muestras y elige "labtech" que es el nombre de los técnicos de laboratorio.

Ejercicio 2: clustering no jerárquico

- Trabajaremos con el dataset **physed.dta** disponible en *File / Example Datasets / Stata 12 Manual Datasets / Multivariate Statistics Reference Manual*
- Los **datos** recogen la información de velocidad, fuerza y flexibilidad de 80 personas que participan en un programa de rehabilitación.
- Tenemos como **objetivo** determinar 4 grupos homogéneos según los atributos físicos medidos, para optimizar la formación y asignación de recursos dirigidos a estos sujetos.

Ejercicio 2: clustering no jerárquico

Plan de trabajo:

1. Leemos los datos en Stata.
2. Realiza un **análisis descriptivo** que nos permita conocer los datos con los que estamos trabajando. Para empezar un resumen de estadísticos descriptivos y un gráfico de dispersión de las 3 variables estaría bien.
   ¿Hay algún tipo de relación entre las variables? ¿A partir de este descriptivo detectas la existencia de grupos de pacientes?.
3. Realiza el **análisis de clustering no jerárquico** de las muestras utilizando el método **kmeans**. Intenta realizar el correspondiente dendrograma. ¿Algún problema?
4. Tras la realización del análisis de clúster, vamos a conocer mejor los grupos detectados. Contesta la siguientes preguntas:

Ejercicio 2: clustering no jerárquico

Plan de trabajo:

- ¿Cuántas personas están incluidas en cada uno de los 4 grupos?
- Nos gustaría conocer la media, mínimo y máximo de los atributos evaluados para cada grupo y así confirmaremos si realmente están bien diferenciados. Comenta los resultados.
- Por último, representa de nuevo los datos utilizando los gráficos de dispersión pero esta vez en lugar de que aparezcan puntos, mejor si indicamos que aparezca el número del grupo al que pertenece cada sujeto, así visualizaremos la relación entre las variables incorporando esta información.

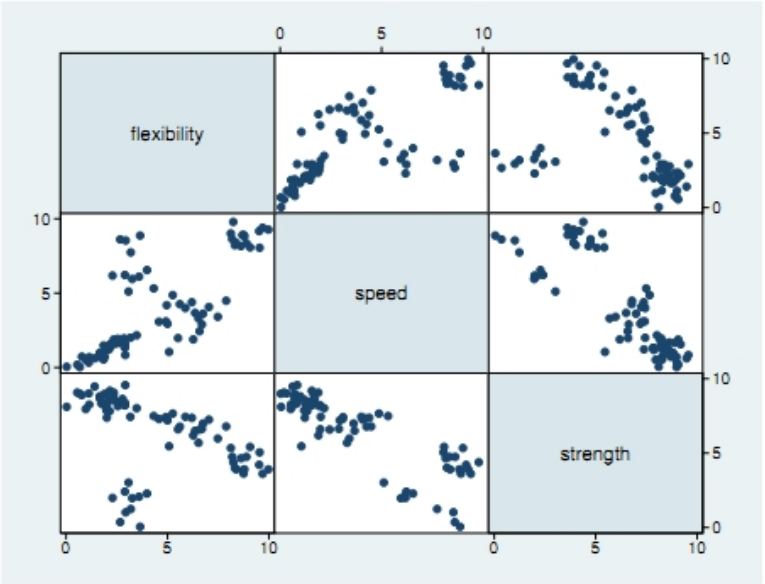
Ejercicio 2: clustering no jerárquico

Resultados:

```
. summarize flex speed strength
```

Variable	Obs	Mean	Std. Dev.	Min	Max
flexibility	80	4.402625	2.788541	.03	9.97
speed	80	3.875875	3.121665	.03	9.79
strength	80	6.439875	2.449293	.05	9.57

```
. graph matrix flex speed strength
```



Ejercicio 2: clustering no jerárquico

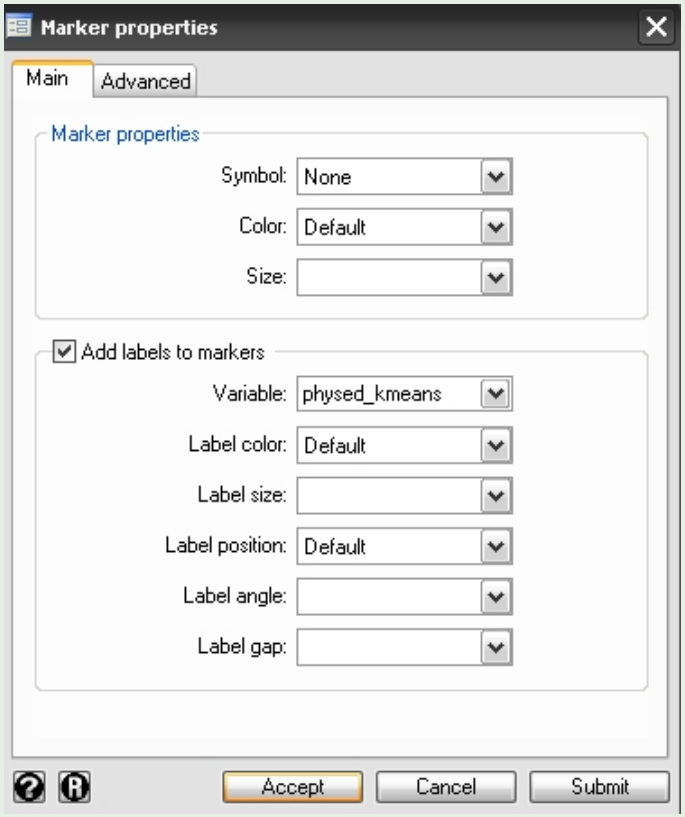
Resultados:

Summary statistics: min, mean, max, N
by categories of: physed_kmeans

physed_kmeans	flexib~y	speed	strength
1	4.32	1.05	5.46
	5.9465	3.4485	6.8325
	7.89	5.32	7.66
	20	20	20
2	.03	.03	7.38
	1.969429	1.144857	8.478857
	3.48	2.17	9.57
	35	35	35
3	8.12	8.05	3.61
	8.852	8.743333	4.358
	9.97	9.79	5.42
	15	15	15
4	2.29	5.11	.05
	3.157	6.988	1.641
	3.99	8.87	3.02
	10	10	10
Total	.03	.03	.05
	4.402625	3.875875	6.439875
	9.97	9.79	9.57
	80	80	80

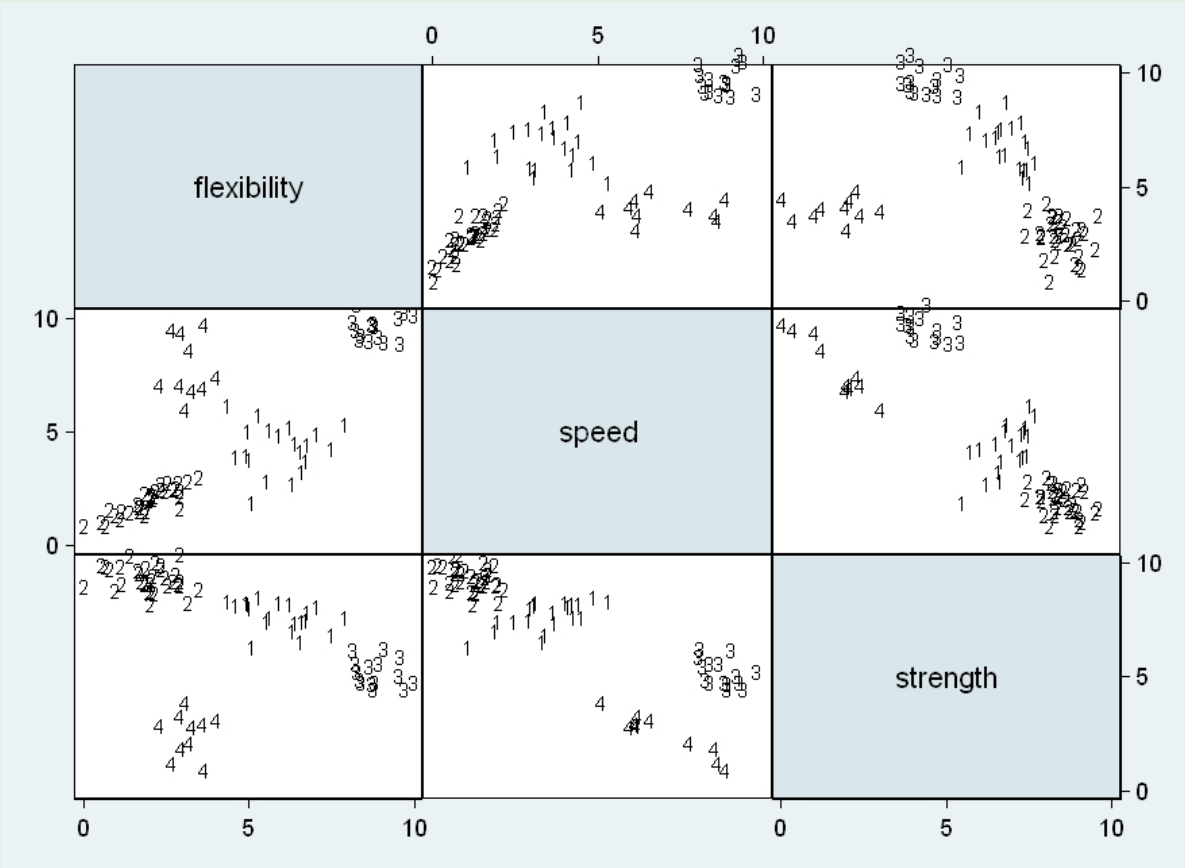
Ejercicio 2: clustering no jerárquico

Resultados:



Ejercicio 2: clustering no jerárquico

Resultados:



Análisis Discriminante

Esta técnica presenta reglas de clasificación óptimas de nuevas observaciones de las que se desconoce su grupo de procedencia basándose en la información proporcionada los valores que en ella toman las variables independientes.

Ejemplos:

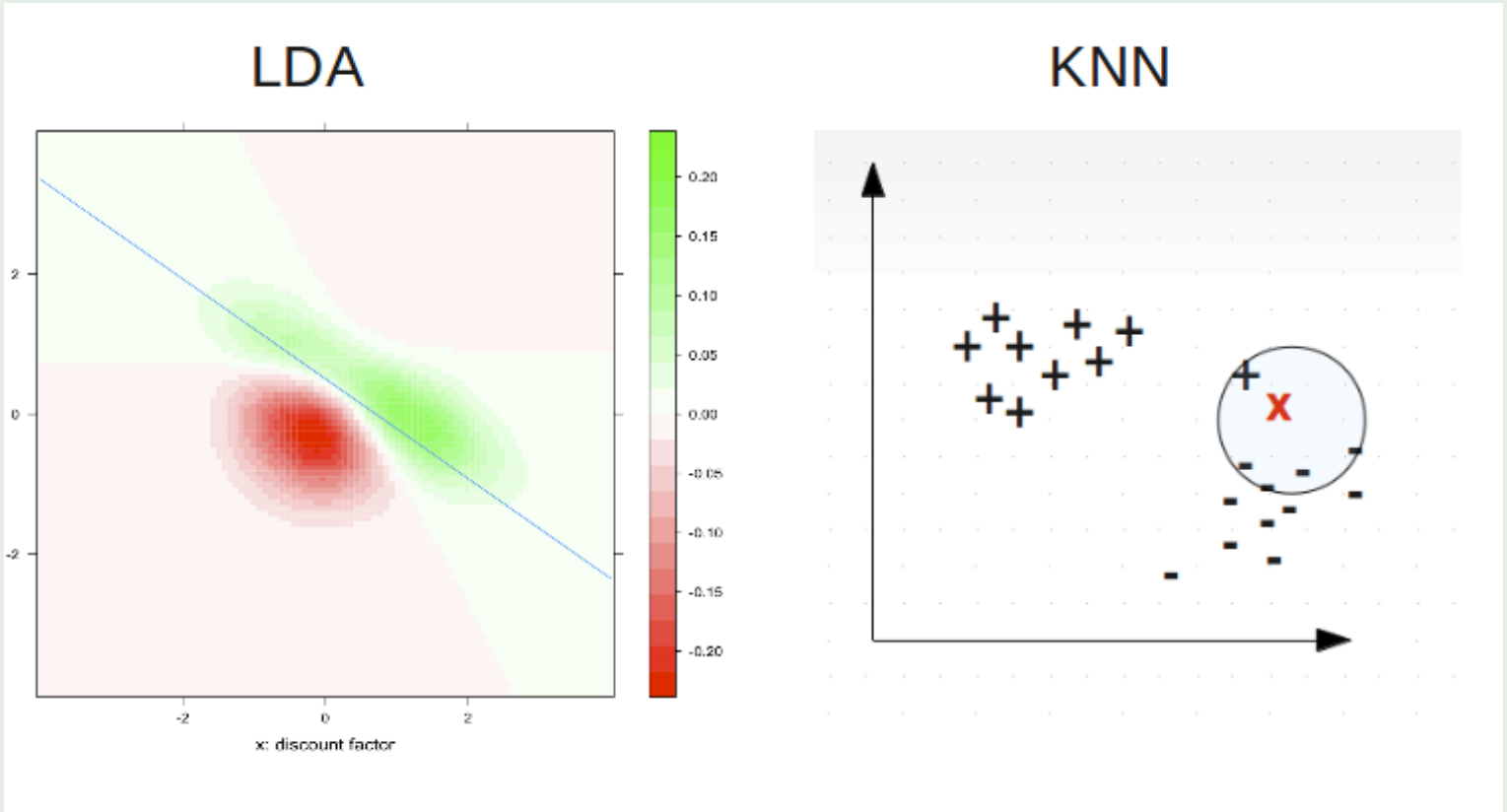
1. Determinar las variables clínicas que permitan discriminar mejor entre pacientes de alto/no alto riesgo en una Unidad de Medicina Intensiva.

2. En planificación y gestión sanitaria también se presentan problemas de discriminación. ¿Cómo se explican las diferencias entre los usuarios de la sanidad pública y la privada?, ¿hasta qué punto son el nivel de renta, la gravedad de los síntomas, etc., responsables de la elección de médico? Las encuestas de salud proporcionan información de base suficiente para contestar a estas preguntas con ayuda del **Análisis Discriminante**.

Tipos de Análisis Discriminante:

1. Linear Discriminant Analysis (LDA).
2. Quadratic Discriminant Analysis (QDA).
3. Logistic Discriminant Analysis.
4. K th-Nearest-Neighbor Discriminant Analysis.

Tipos de Análisis Discriminante:

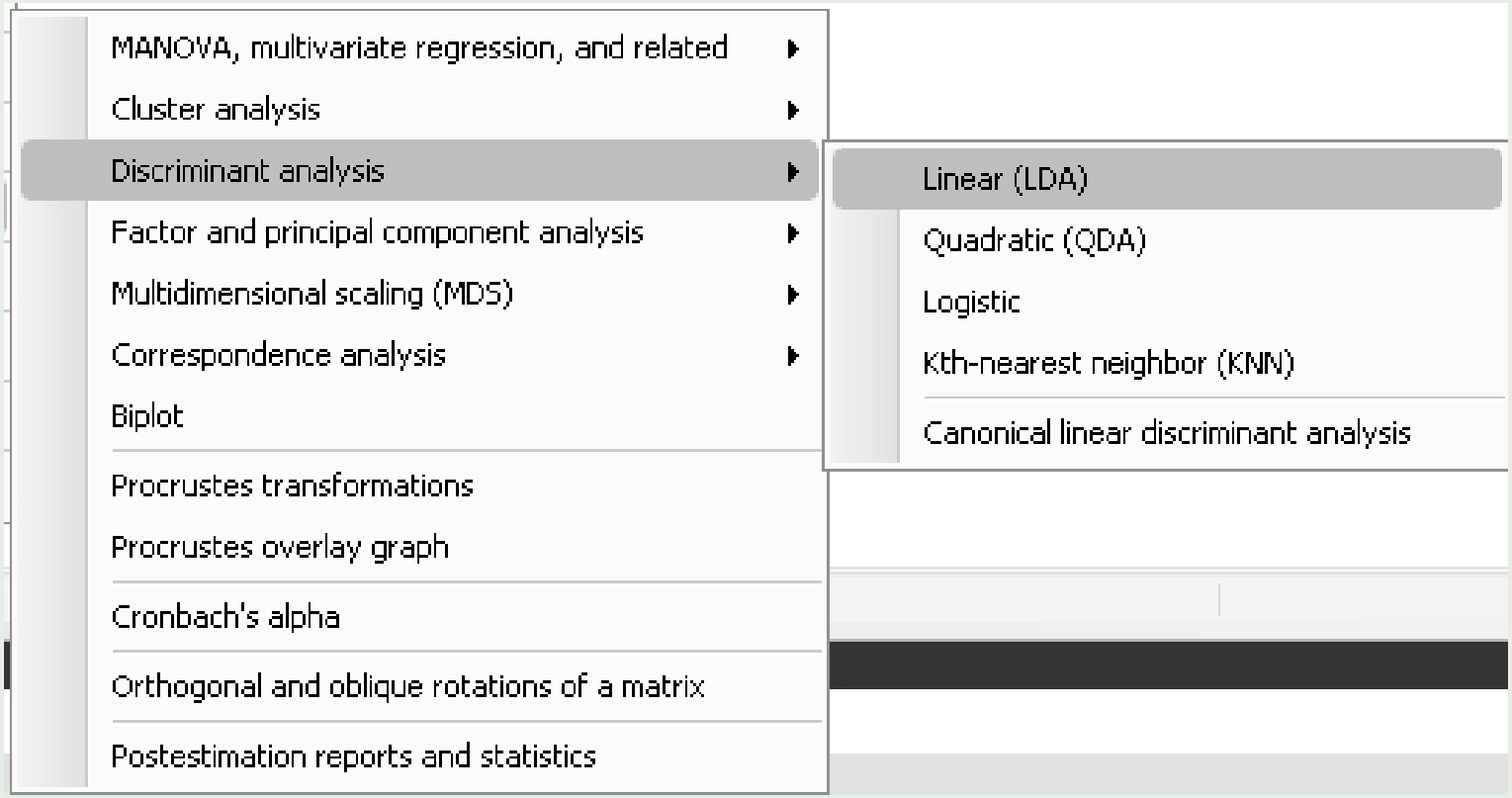


¿Cómo realizamos el Análisis Discriminante en Stata?

Varios pasos:

1. Descripción de los datos.
2. Estimación del modelo.
3. Evaluación del modelo y predicción.

¿Cómo realizamos el Análisis Discriminante en Stata?



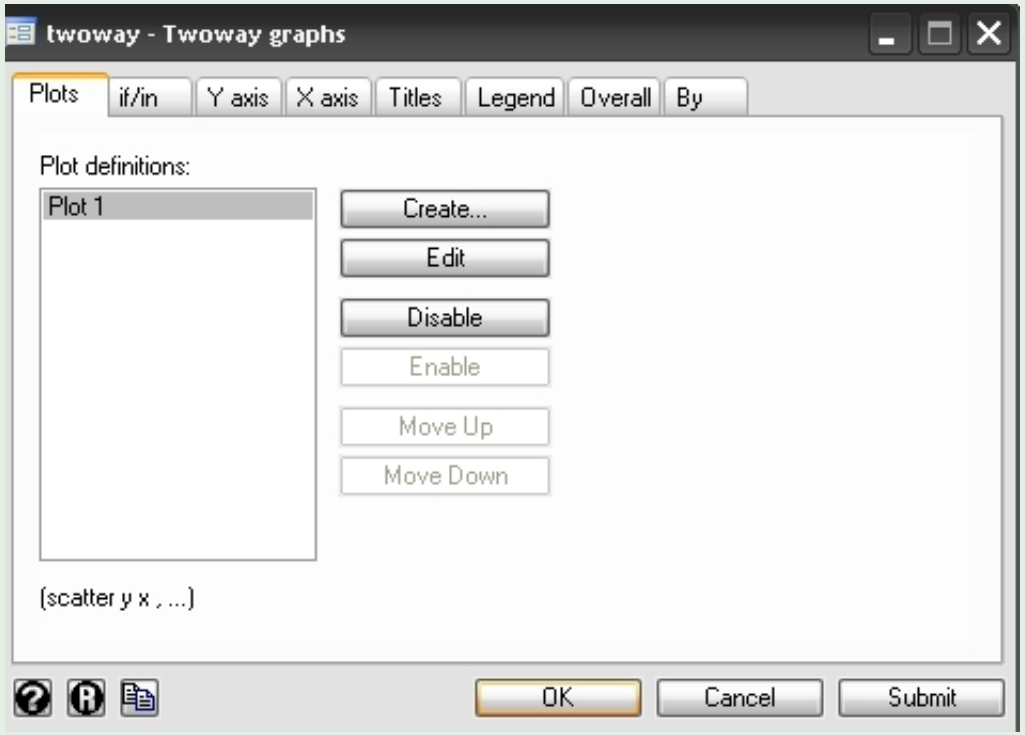
Ejemplo de Análisis Discriminante con Stata:

Empezamos trabajando con el dataset **twogroups.dta** que incluye 30 observaciones y tres variables. La primera establece dos grupos y las variables **x, y** son las que discriminan la pertenencia a un determinado grupo.

	group	y	x
1	1	29	58
2	1	27	44
3	1	42	35
4	1	25	36
5	1	30	43
6	1	33	46
7	1	23	62
8	1	48	55
9	1	27	51
10	1	41	57
11	1	40	42
12	1	47	56
13	1	37	34
14	1	51	27
15	1	37	32
16	2	23	28
17	2	49	15
18	2	32	18
19	2	18	45
20	2	43	17
21	2	26	28
22	2	20	30
23	2	25	25
24	2	33	35
25	2	31	14
26	2	16	40
27	2	33	37
28	2	27	30
29	2	35	9
30	2	21	27

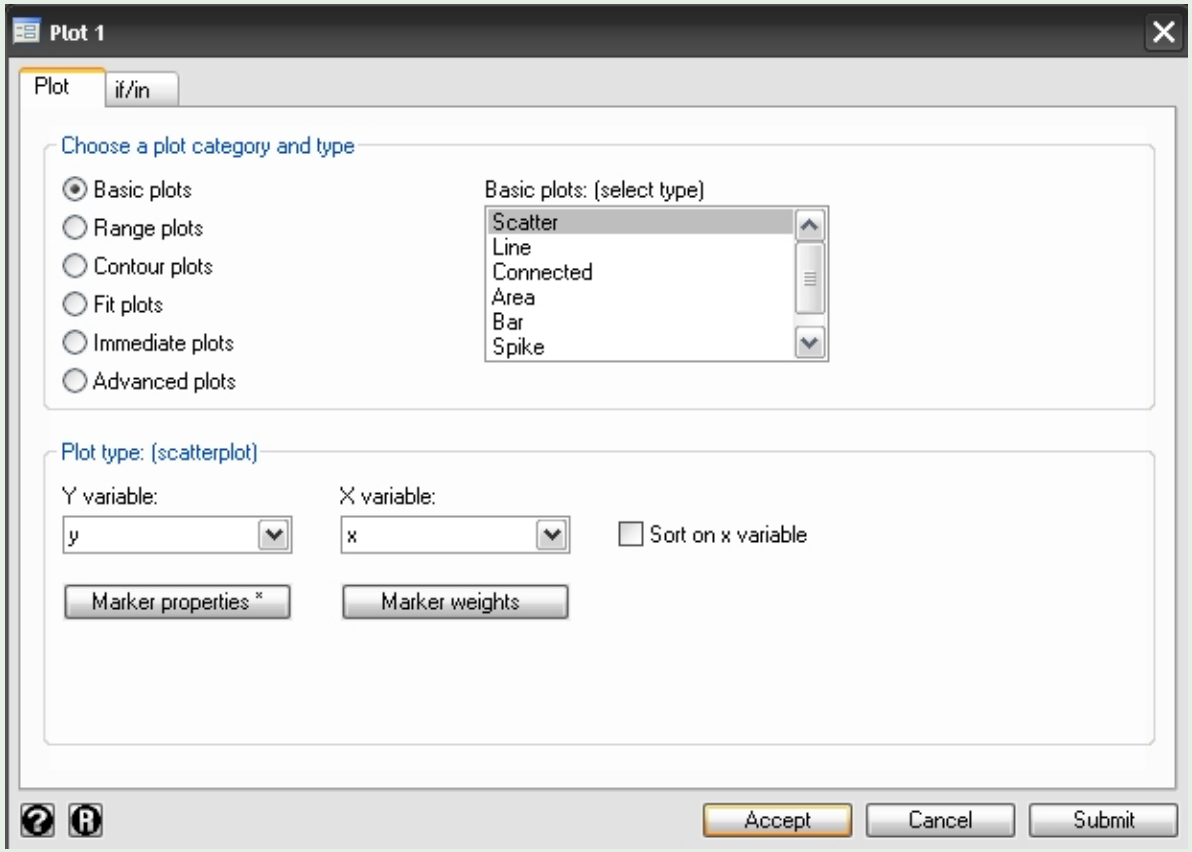
Ejemplo de Análisis Discriminante con Stata:

Paso 1. Descripción de los datos: *Graphs / Twoways graph: scatter*



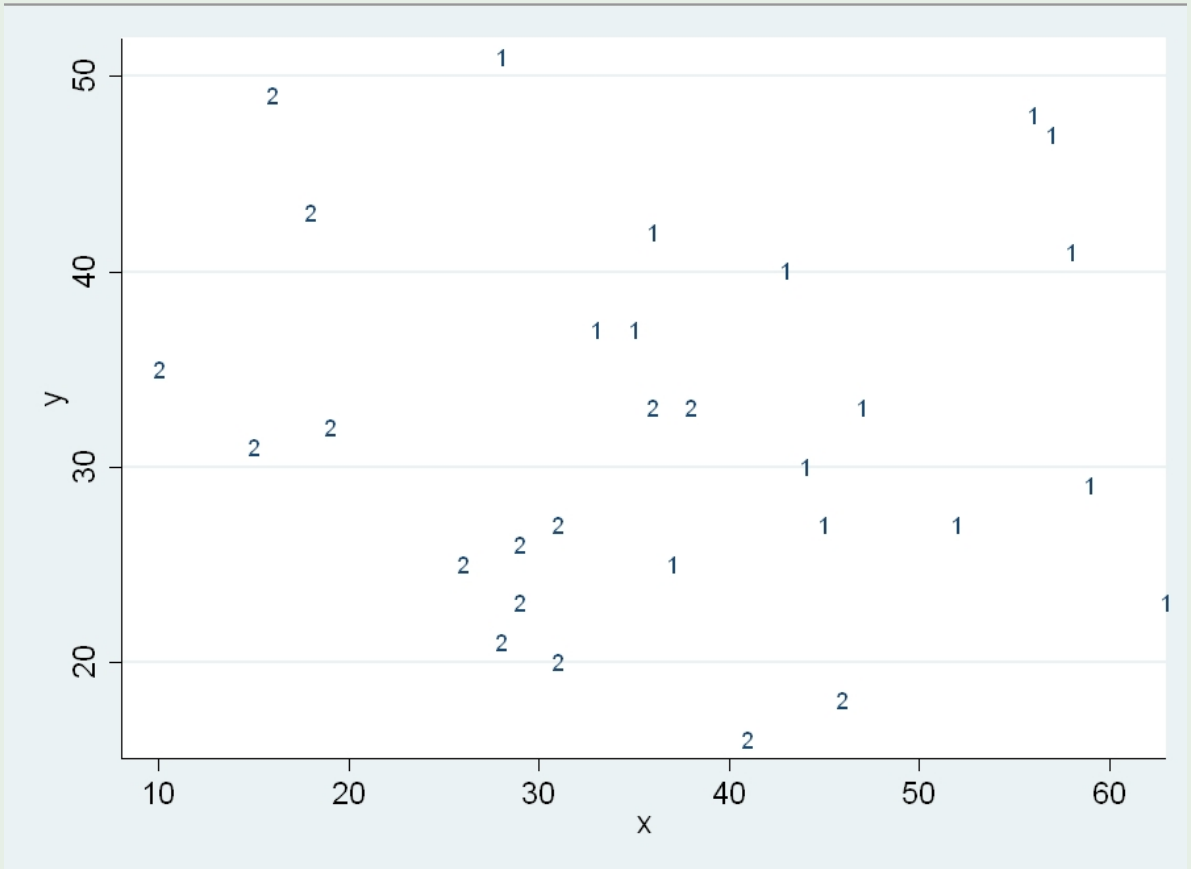
Ejemplo de Análisis Discriminante con Stata:

Paso 1. Descripción de los datos: *Graphs / Twoways graph: scatter*



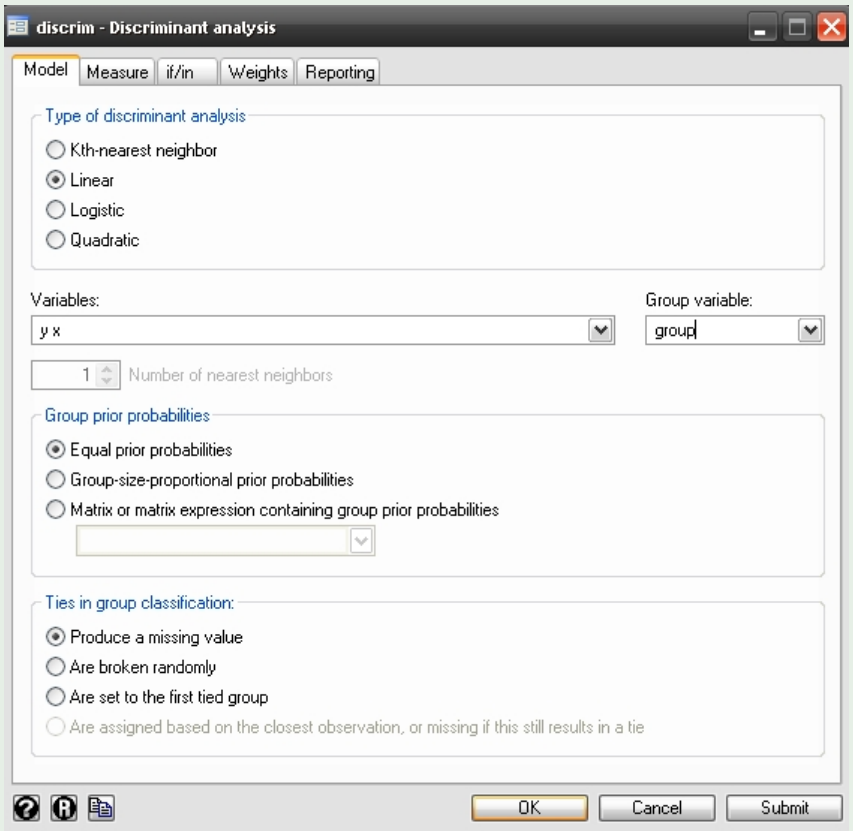
Ejemplo de Análisis Discriminante con Stata:

Paso 1. Descripción de los datos: *Graphs / Twoways graph: scatter*



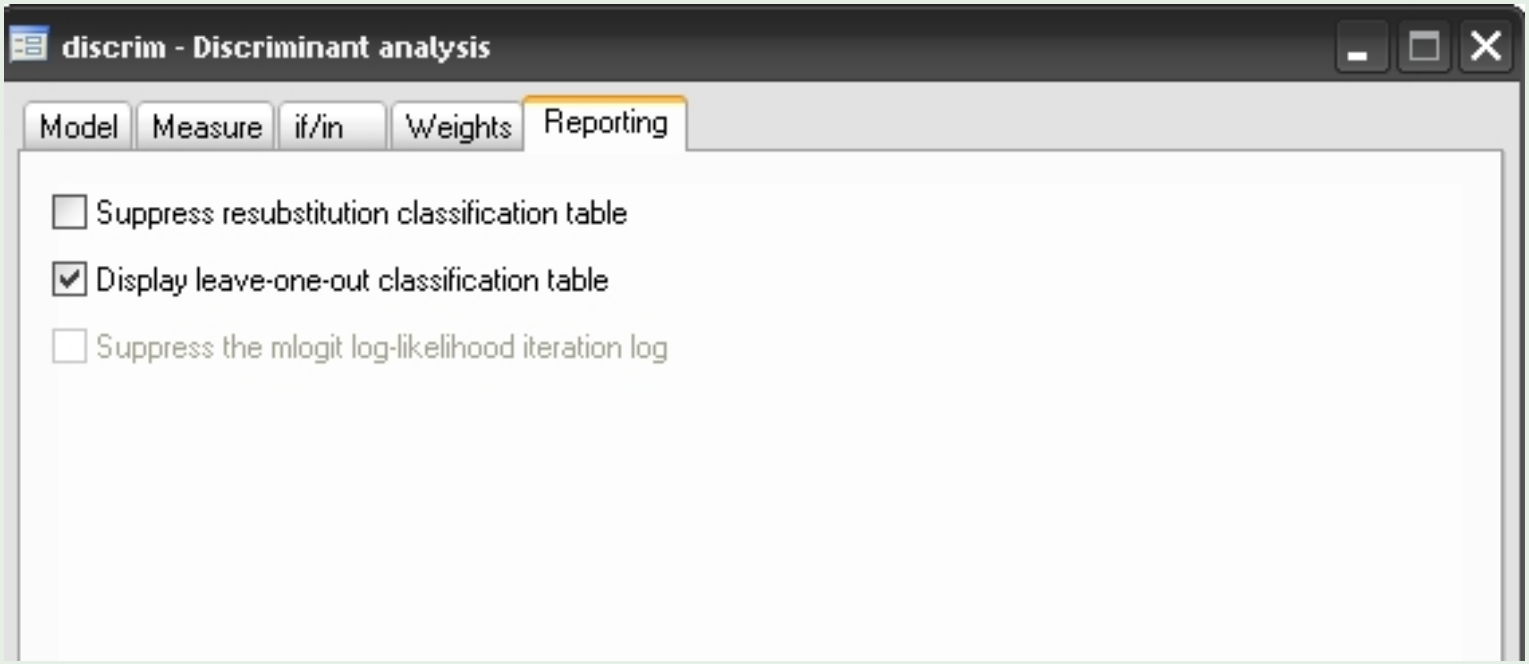
Ejemplo de Análisis Discriminante con Stata:

Paso 2. Estimación del modelo:
Statistics / Multivariate Analysis / Discriminant Analysis / DLA



Ejemplo de Análisis Discriminante con Stata:

Paso 2. Estimación del modelo:
Statistics / Multivariate Analysis / Discriminant Analysis / DLA



Ejemplo de Análisis Discriminante con Stata:

Paso 2. Estimación del modelo:
Statistics / Multivariate Analysis / Discriminant Analysis / DLA

```
Linear discriminant analysis
Resubstitution classification summary
```

Key
Number Percent

True group	Classified 1	2	Total
1	**14** **93.33**	**1** **6.67**	**15** **100.00**
2	**1** **6.67**	**14** **93.33**	**15** **100.00**
Total	**15** **50.00**	**15** **50.00**	**30** **100.00**
Priors	**0.5000**	**0.5000**	

Ejemplo de Análisis Discriminante con Stata:

Paso 2. Estimación del modelo:
Statistics / Multivariate Analysis / Discriminant Analysis / DLA

Leave-one-out classification summary

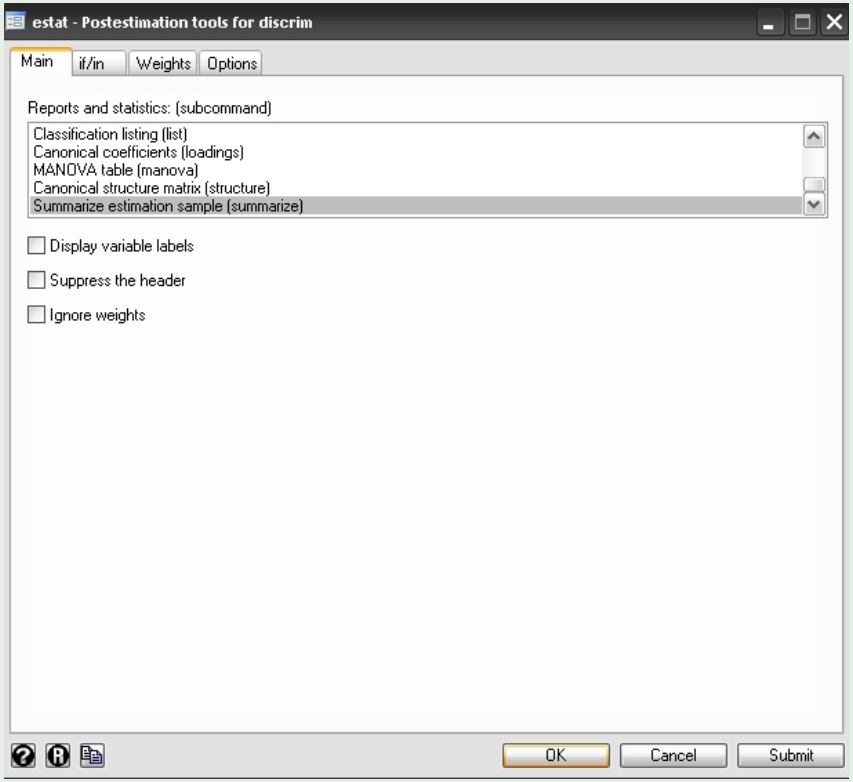
Key

Number
Percent

True group	Classified 1	2	Total
1	13	2	15
	86.67	13.33	100.00
2	2	13	15
	13.33	86.67	100.00
Total	15	15	30
	50.00	50.00	100.00
Priors	0.5000	0.5000	

Ejemplo de Análisis Discriminante con Stata:

Paso 3. Evaluación y predicción del modelo: *Postestimation / Reports and statistics*



Descripción de los datos por grupos.

Ejemplo de Análisis Discriminante con Stata:

Paso 3. Evaluación y predicción del modelo: *Postestimation / Reports and statistics*

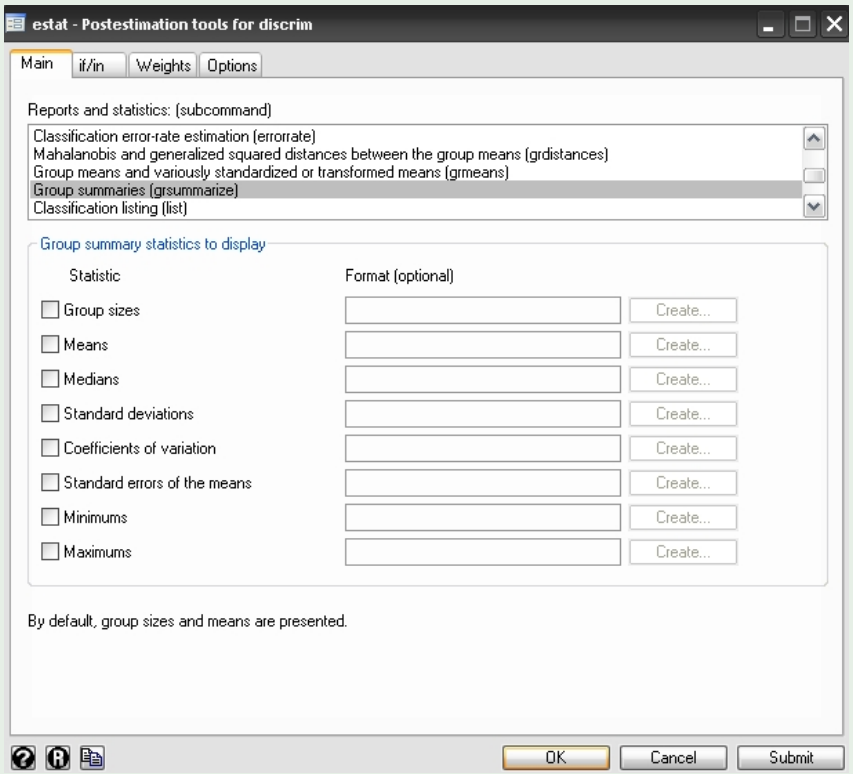
```
. estat summarize

Estimation sample discrim                       Number of obs =      30
```

Variable	Mean	Std. Dev.	Min	Max
groupvar				
group	1.5	.5085476	1	2
variables				
y	32.3	9.574211	16	51
x	35.86667	14.11464	9	62

Descripción de los datos por grupos.

Ejemplo de Análisis Discriminante con Stata:

Paso 3. Evaluación y predicción del modelo: *Postestimation / Reports and statistics*



Descripción de los datos por grupos.

Ejemplo de Análisis Discriminante con Stata:

Paso 3. Evaluación y predicción del modelo: *Postestimation / Reports and statistics*

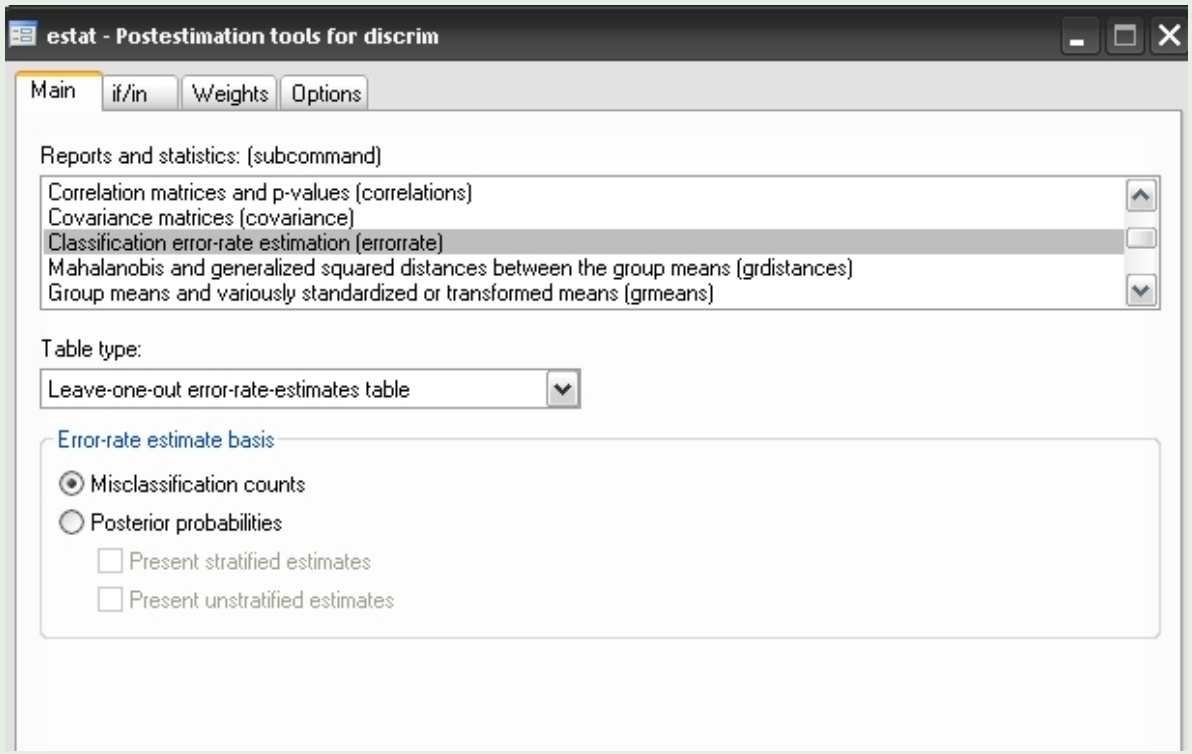
```
. estat grsummarize

Estimation sample discrim lda
Summarized by group
```

Mean	group 1	2	Total
y	35.8	28.8	32.3
x	45.2	26.53333	35.86667
N	15	15	30

Descripción de los datos por grupos.

Ejemplo de Análisis Discriminante con Stata:

Paso 3. Evaluación y predicción del modelo: *Postestimation / Reports and statistics*



Evaluación del modelo: tasa de error por grupos.

Ejemplo de Análisis Discriminante con Stata:

Paso 3. Evaluación y predicción del modelo: *Postestimation / Reports and statistics*

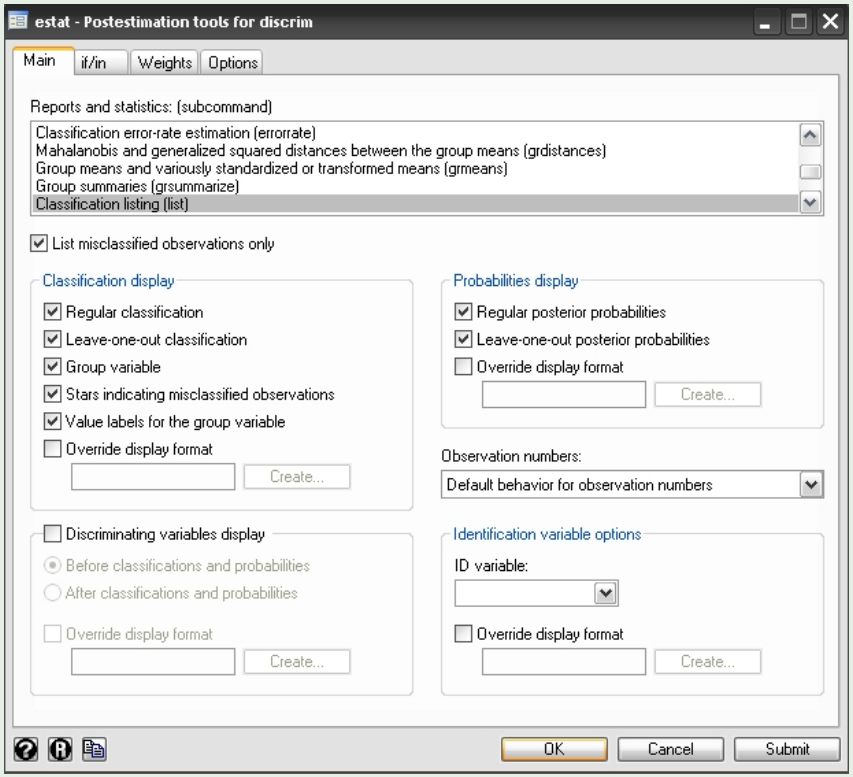
```
. estat errorrate, looclass

Error rate estimated by leave-one-out error count
```

	group		
	1	2	Total
Error rate	.1333333	.1333333	.1333333
Priors	.5	.5	

Evaluación del modelo: tasa de error por grupos.

Ejemplo de Análisis Discriminante con Stata:

Paso 3. Evaluación y predicción del modelo: *Postestimation / Reports and statistics*



Evaluación del modelo: sujetos mal clasificados.

Ejemplo de Análisis Discriminante con Stata:

Paso 3. Evaluación y predicción del modelo: *Postestimation / Reports and statistics*

```
. estat list, misclassified classification(looclass) probabilities(loopr)
```

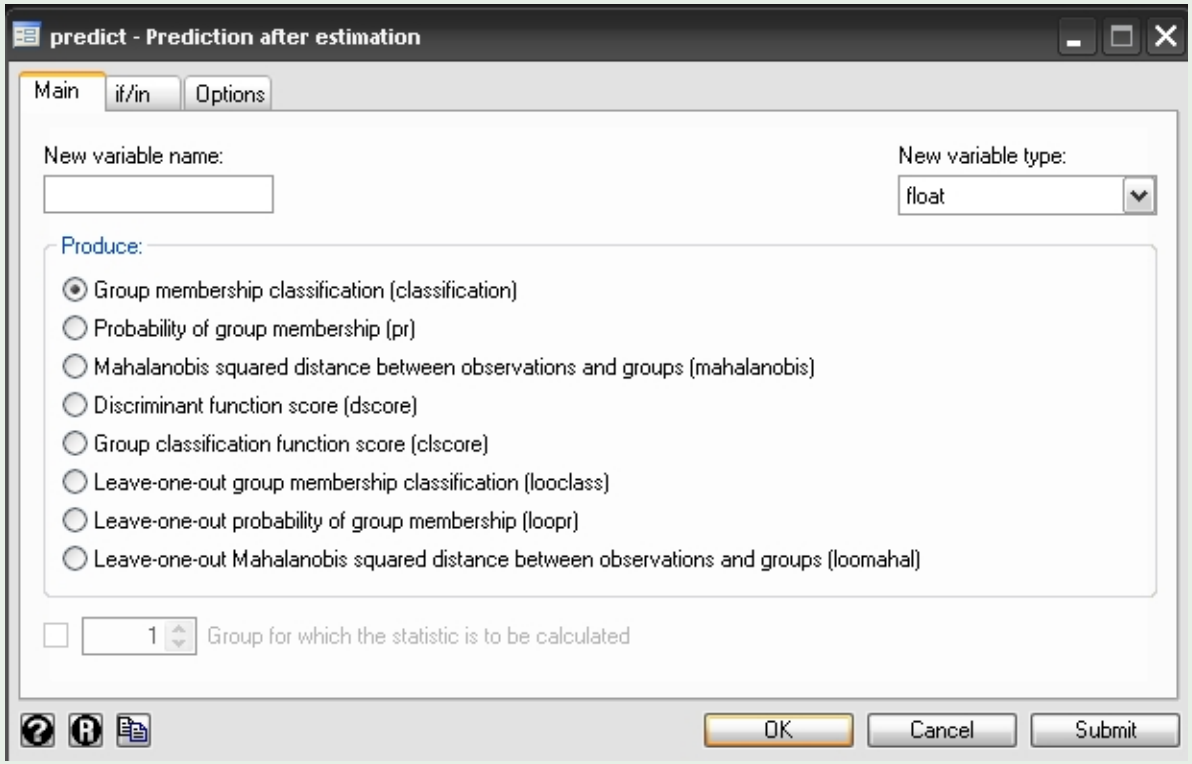
	Classification		Probabilities		LOO Probabilities	
Obs.	True Class.	LOO Cl.	1	2	1	2
4	1	2 *	2 *	0.1801	0.8199	0.1105	0.8895
15	1	1	2 *	0.5127	0.4873	0.4709	0.5291
24	2	2	1 *	0.4841	0.5159	0.5116	0.4884
27	2	1 *	1 *	0.6048	0.3952	0.6447	0.3553

```
* indicates misclassified observations
```

Evaluación del modelo: sujetos mal clasificados.

Ejemplo de Análisis Discriminante con Stata:

Paso 3. Evaluación y predicción del modelo: *Postestimation / Predictions and errors*



Predicción para un nuevo caso o sujeto.

Ejercicio 1: Análisis Discriminante Lineal (DLA)

- Trabajaremos con el dataset **lawnmower2** disponible en *File / Example Datasets / Stata 12 Manual Datasets / Multivariate Statistics Reference Manual*
- Tenemos 24 individuos con información correspondiente a tres variables: **owner** que establece los dos grupos que hay en el dataset (propietario y no propietario), **income** y **lotsize** son las dos variables que discriminan si un sujeto pertenece a un grupo u a otro.

Ejercicio 1: Análisis Discriminante Lineal (DLA)

	owner	income	lotsize
1	owner	90.0	18.4
2	owner	115.5	16.8
3	owner	94.8	21.6
4	owner	91.5	20.8
5	owner	117.0	23.6
6	owner	140.1	19.2
7	owner	138.0	17.6
8	owner	112.8	22.4
9	owner	99.0	20.0
10	owner	123.0	20.8
11	owner	81.0	22.0
12	owner	111.0	20.0
13	nonowner	105.0	19.6
14	nonowner	82.8	20.8
15	nonowner	94.8	17.2
16	nonowner	73.2	20.4
17	nonowner	114.0	17.6
18	nonowner	79.2	17.6
19	nonowner	89.4	16.0
20	nonowner	96.0	18.4
21	nonowner	77.4	16.4
22	nonowner	63.0	18.8
23	nonowner	81.0	14.0
24	nonowner	93.0	14.8

Ejercicio 1: Análisis Discriminante Lineal (DLA)

Plan de trabajo:

1. Realiza una **descripción de los datos**. Explora gráficamente los datos mediante un diagrama de dispersión. ¿Crees que están bien diferenciados los sujetos pertenecientes a cada grupo?

2. Realiza un **análisis de discriminante DLA** y pide la **tabla de clasificación Leave-one-out**. Interpreta esta tabla, ¿es un buen clasificador el modelo que hemos escogido?. Describe con detalle el funcionamiento de las **matrices de confusión**: significado de los valores que están en la diagonal, en los extremos...

3. Vamos a acercarnos un poco más a nuestros datos y al modelo que hemos generado:

   - Desde el menú de *Postestimation* realiza un **descriptivo por grupos** para ver como se comportan las variables.
   - **Por sujetos**: lista aquellos sujetos mal clasificados y comenta los resultados que nos ofrece STATA.
   - **Por grupos**: ¿cuál es el grupo que presenta una probabilidad de error más alta?

Análisis Factorial

Se utiliza para analizar interrelaciones entre un número elevado de variables cuantitativas explicando dichas interrelaciones en términos de un número menor de variables que se denominan **factores** o **componentes principales**.

Análisis Factorial vs. Análisis de Componentes Principales:

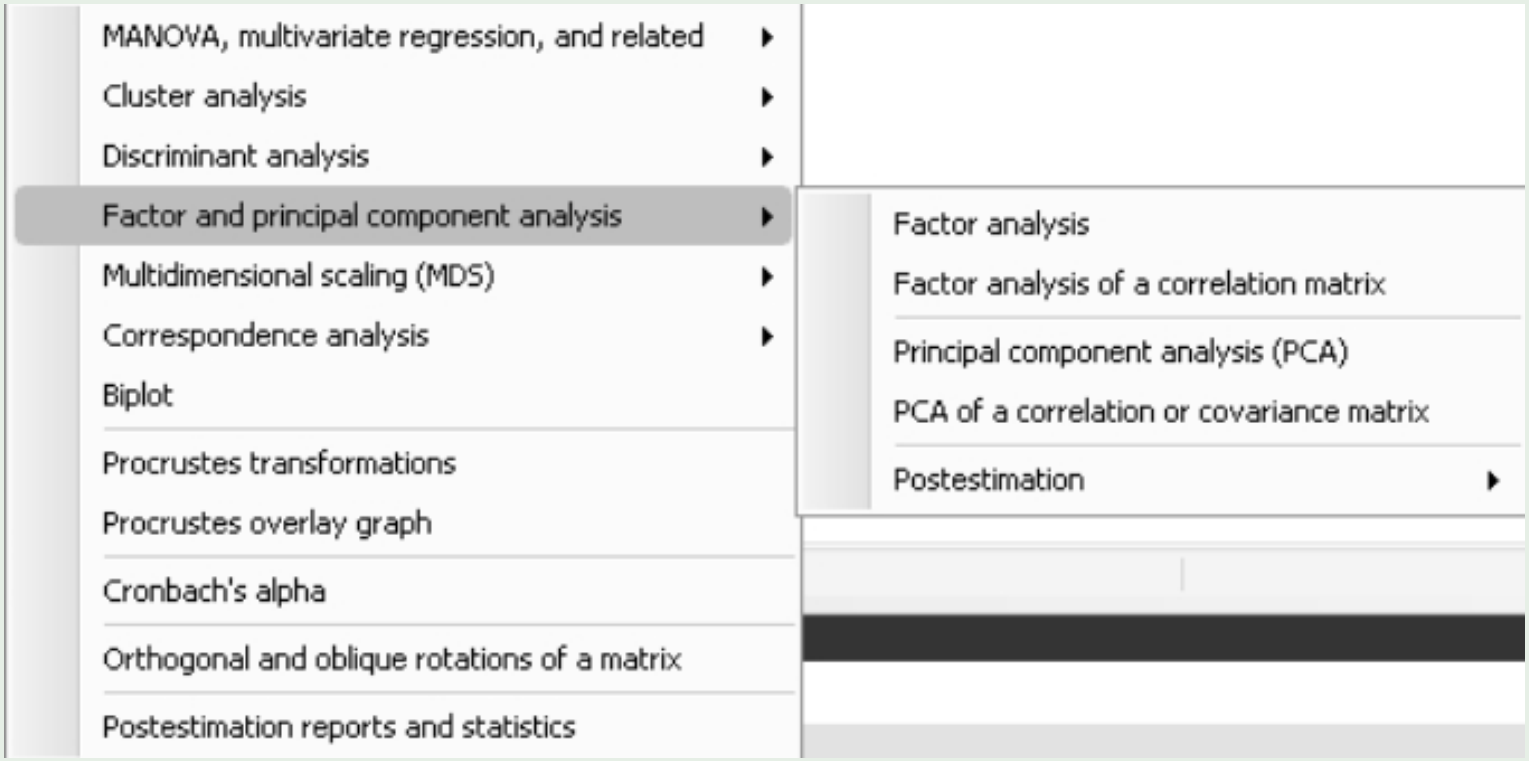
- El Análisis Factorial y el Análisis de Componentes Principales están muy relacionados. Algunos autores consideran el segundo como una etapa del primero.
- El Análisis de Componentes Principales trata de hallar componentes (factores) que sucesivamente expliquen la mayor parte de la varianza total. Por su parte el Análisis Factorial busca factores que expliquen la mayor parte de la varianza común.
- El Análisis Factorial supone que existe un factor común subyacente a todas las variables, el Análisis de Componentes Principales no hace tal asunción.

Ejemplos:

1. Si un psicólogo quiere determinar los factores que caracterizan la inteligencia de un individuo a partir de sus respuestas a un test de inteligencia, utilizaría para resolver este problema un **Análisis Factorial**.

2. Determinación de indicadores que midan las necesidades en los servicios sociales de una ciudad a partir de datos de indicadores socioeconómicos y demográficos. **Análisis de Componentes Principales**.

Análisis Factorial con Stata



Análisis Factorial con Stata

Varios pasos:

1. Descripción de los datos.
2. Elección y realización del tipo de análisis: Factorial / Componentes Principales.
3. Post-estimación: evaluación de resultados.

Ejemplo de Análisis de Componentes Principales con Stata:

Disponemos del set de datos **audiometric.dta** que incluye mediciones audiométricas de niños de 9 años. Se midieron 4 intensidades diferentes para el oído derecho y el izquierdo. Así por ejemplo la variable **lft1000** hace reference al oído izquierdo en la frecuencia 1000 Hz.

	id	lft500	lft1000	lft2000	lft4000	rght500	rght1000	rght2000	rght4000
1	1	0	5	10	15	0	5	5	15
2	2	-5	0	-10	0	0	5	5	15
3	3	-5	0	15	15	0	0	10	25
4	4	-5	0	-10	-10	-10	-5	-10	10
5	5	-5	-5	-10	10	0	-10	-10	50
6	6	5	5	5	-10	0	5	0	20
7	7	0	0	0	20	5	5	5	10
8	8	-10	-10	-10	-5	-10	-5	0	5
9	9	0	0	0	40	0	0	-10	10

Ejemplo de Análisis de Componentes Principales con Stata:

Paso 1. Descripción de los datos: correlaciones

Statistics / Summaries, tables, and tests / Summary and descriptive statistics / Correlations and covariances

```
. correlate lft* rght*
(obs=100)
```

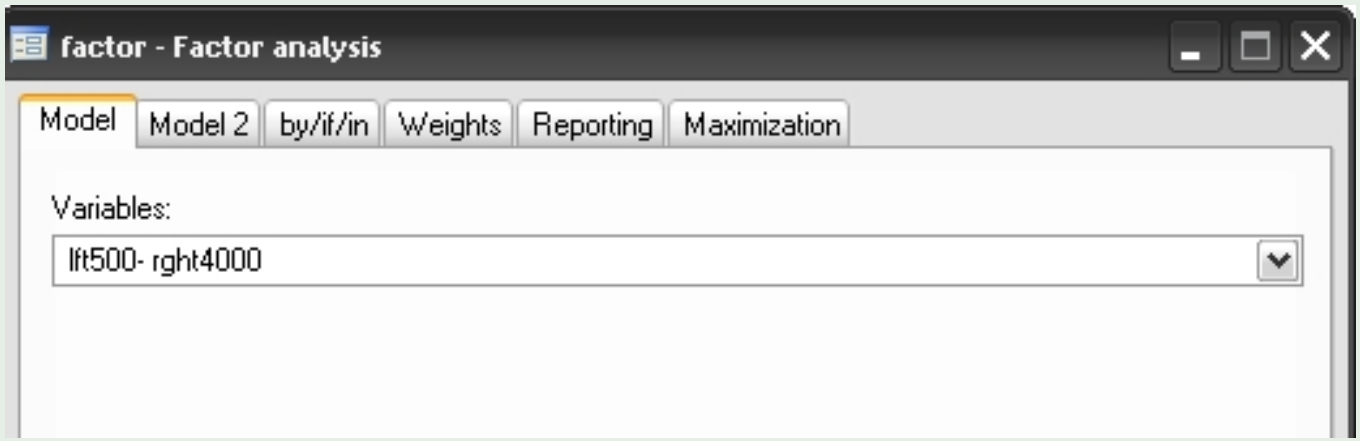
	lft500	lft1000	lft2000	lft4000	rght500	rght1000	rght2000
lft500	1.0000						
lft1000	0.7775	1.0000					
lft2000	0.4012	0.5366	1.0000				
lft4000	0.2554	0.2749	0.4250	1.0000			
rght500	0.6963	0.5515	0.2391	0.1790	1.0000		
rght1000	0.6416	0.7070	0.4460	0.2632	0.6634	1.0000	
rght2000	0.2372	0.3597	0.7011	0.3165	0.1589	0.4142	1.0000
rght4000	0.2041	0.2169	0.3262	0.7097	0.1321	0.2201	0.3746

	rght4000
rght4000	1.0000

La matriz de correlaciones de todas las variables nos proporciona una información interesante sobre la relación existente entre ellas.

Ejemplo de Análisis de Componentes Principales con Stata:

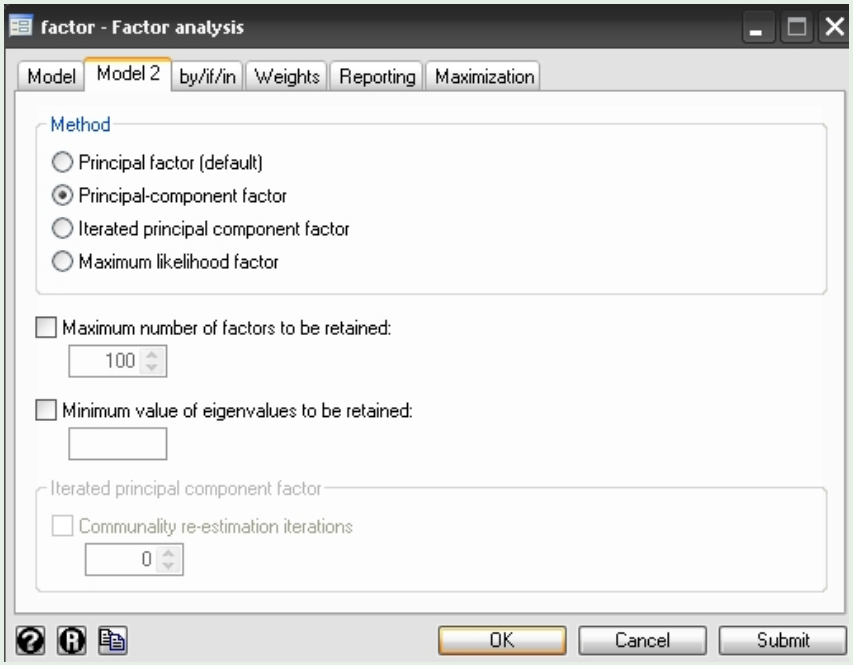
Paso 2. Análisis de Componentes Principales
Statistics / Multivariate Analysis / Factor and Principal Component Analysis / PCA



Seleccionamos las variables sobre las que realizaremos el análisis factorial.

Ejemplo de Análisis de Componentes Principales con Stata:

Paso 2. Análisis de Componentes Principales
Statistics / Multivariate Analysis / Factor and Principal Component Analysis / PCA



Elegimos el método de análisis factorial.
En este caso PCA (Principal Component Factor).

Ejemplo de Análisis de Componentes Principales con Stata:

Paso 2. Análisis de Componentes Principales
Statistics / Multivariate Analysis / Factor and Principal Component Analysis / PCA

Summary statistics of the variables

Variable	Mean	Std. Dev.	Min	Max
lft500	-2.8	6.408643	-10	15
lft1000	-.5	7.571211	-10	20
lft2000	2	10.94061	-10	45
lft4000	21.35	19.61569	-10	70
rght500	-2.6	7.123726	-10	25
rght1000	-.7	6.396811	-10	20
rght2000	1.6	9.289942	-10	35
rght4000	21.35	19.33039	-10	75

Descripción de las variables.

Ejemplo de Análisis de Componentes Principales con Stata:

Paso 2. Análisis de Componentes Principales

Statistics / Multivariate Analysis / Factor and Principal Component Analysis / PCA

```
. pca lft500- rght4000, means

Principal components/correlation                  Number of obs    =        100
                                                  Number of comp.  =          8
                                                  Trace            =          8
    Rotation: (unrotated = principal)             Rho              =     1.0000
```

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	3.92901	2.31068	0.4911	0.4911
Comp2	1.61832	.642997	0.2023	0.6934
Comp3	.975325	.508543	0.1219	0.8153
Comp4	.466782	.126692	0.0583	0.8737
Comp5	.34009	.0241988	0.0425	0.9162
Comp6	.315891	.11578	0.0395	0.9557
Comp7	.200111	.0456375	0.0250	0.9807
Comp8	.154474	.	0.0193	1.0000

Lista de **valores propios** y la **proporción de variabilidad explicada**.

Ejemplo de Análisis de Componentes Principales con Stata:

Paso 2. Análisis de Componentes Principales
Statistics / Multivariate Analysis / Factor and Principal Component Analysis / PCA

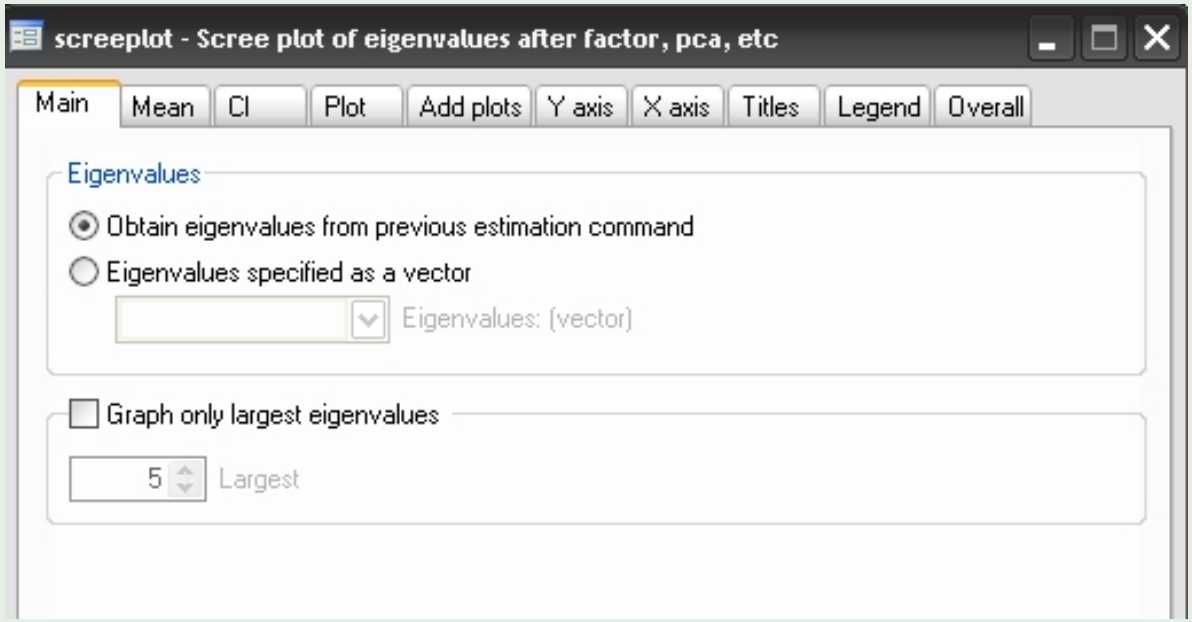
Principal components (eigenvectors)

Variable	Comp1	Comp2	Comp3	Comp4	Comp5	Comp6	Comp7	Comp8	Unexplained
lft500	0.4011	-0.3170	0.1582	-0.3278	0.0231	0.4459	0.3293	-0.5463	0
lft1000	0.4210	-0.2255	-0.0520	-0.4816	-0.3792	-0.0675	-0.0331	0.6227	0
lft2000	0.3664	0.2386	-0.4703	-0.2824	0.4392	-0.0638	-0.5255	-0.1863	0
lft4000	0.2809	0.4742	0.4295	-0.1611	0.3503	-0.4169	0.4269	0.0839	0
rght500	0.3433	-0.3860	0.2593	0.4876	0.4975	0.1948	-0.1594	0.3425	0
rght1000	0.4114	-0.2318	-0.0289	0.3723	-0.3513	-0.6136	-0.0837	-0.3614	0
rght2000	0.3115	0.3171	-0.5629	0.3914	-0.1108	0.2650	0.4778	0.1466	0
rght4000	0.2542	0.5135	0.4262	0.1591	-0.3960	0.3660	-0.4139	-0.0508	0

Componentes principales para cada variable.

Ejemplo de Análisis de Componentes Principales con Stata:

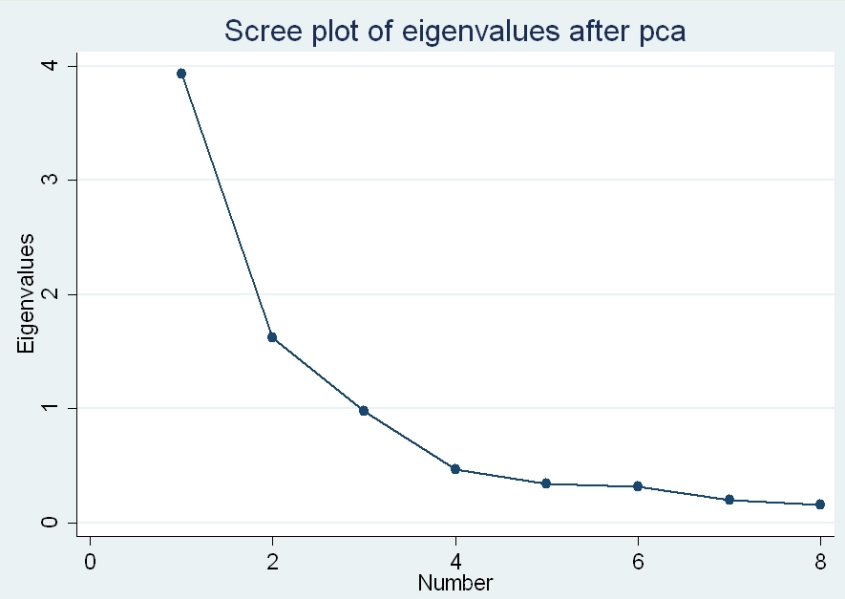
Post-estimación: evaluación de resultados.
Statistics / Multivariate Analysis / Factor and Principal Component Analysis / Postestimation



Screeplot. Gráfico de los valores propios.

Ejemplo de Análisis de Componentes Principales con Stata:

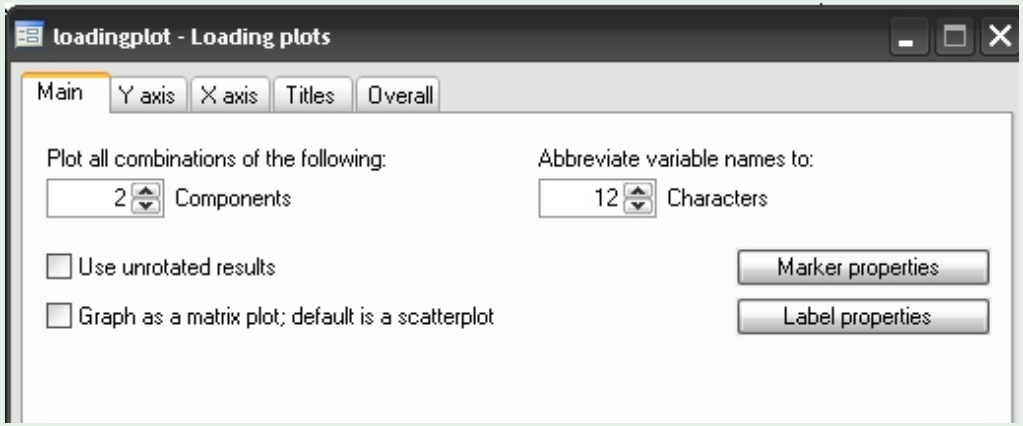
Post-estimación: evaluación de resultados.
Statistics / Multivariate Analysis / Factor and Principal Component Analysis / Postestimation



Screeplot. Gráfico de los valores propios.

Ejemplo de Análisis de Componentes Principales con Stata:

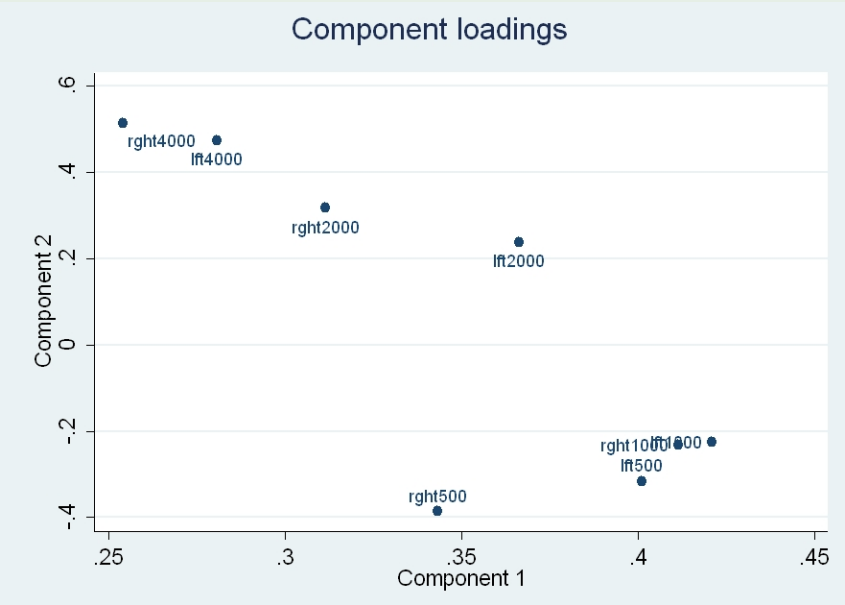
Post-estimación: evaluación de resultados.
Statistics / Multivariate Analysis / Factor and Principal Component Analysis / Postestimation



Representación gráfica de las variables mediante sus dos primeras componentes principales.

Ejemplo de Análisis de Componentes Principales con Stata:

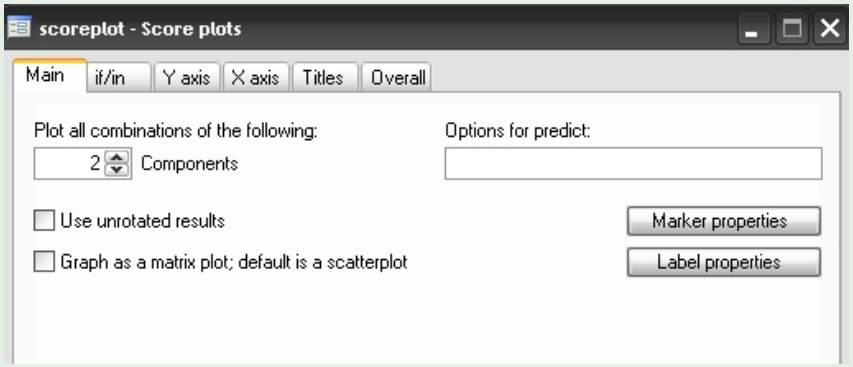
Post-estimación: evaluación de resultados.
Statistics / Multivariate Analysis / Factor and Principal Component Analysis / Postestimation



Representación gráfica de las variables mediante sus dos primeras componentes principales.

Ejemplo de Análisis de Componentes Principales con Stata:

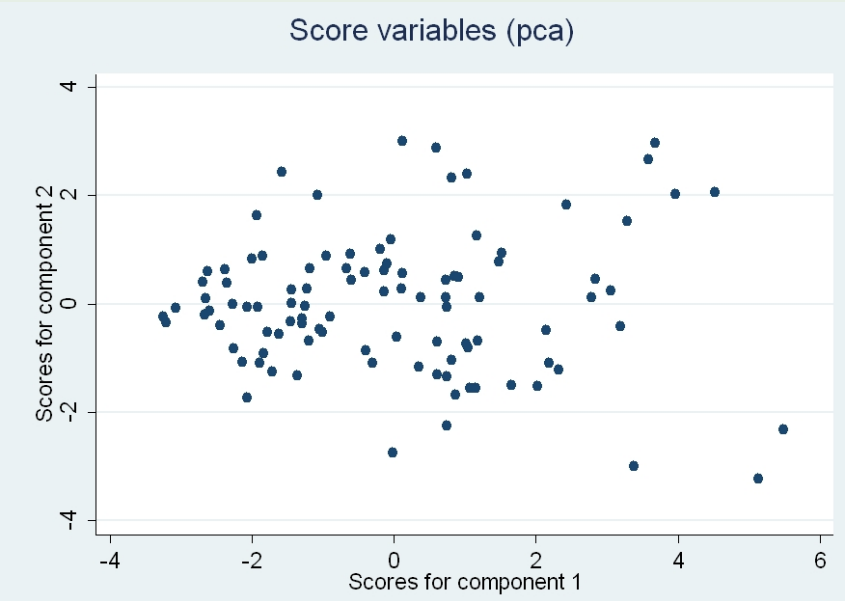
Post-estimación: evaluación de resultados.
Statistics / Multivariate Analysis / Factor and Principal Component Analysis / Postestimation



Representación de las observaciones mediante las puntuaciones en sus dos primeras componentes principales.

Ejemplo de Análisis de Componentes Principales con Stata:

Post-estimación: evaluación de resultados.
Statistics / Multivariate Analysis / Factor and Principal Component Analysis / Postestimation



Representación de las observaciones mediante las puntuaciones en sus dos primeras componentes principales.

Referencias bibliográficas:

- **Análisis Multivariante. Aplicación al ámbito sanitario**. Beatriz González López-Valcárcel. Editores SG.
- **Manual de Stata, versión 12**.