CAPÍTULO X.

ANÁLISIS DE DATOS EN LA DECISIÓN GERENCIAL (II)

CONTENIDO.

- **Introducción**
- **Introducción a las técnicas de reducción de dimensionalidad.**
- **Análisis de Componentes Principales (ACP). Ejemplo con SPSS y SAS**
- **Correspondencia (AC).- Proyección de las filas.-Proyección por columnas. Ejemplo con SAS.**
- **Análisis de conglomerado. Ejemplos con SAS.**
- **Análisis Discriminante.-Análisis lineal de Fisher-Ejemplo con SAS. Logit.-Ejemplo con SAS.**
- **Bibliografía.**

Prof: Andrés E Reyes P

INTRODUCCIÓN.

En este capítulo, estudiaremos algunas técnicas estadísticas que permiten analizar una gran masa de datos, que suelen estar en las base de datos de diferentes corporaciones financieras, públicas y de otra naturaleza, ellas son parte de las herramientas que se emplean en lo que se conoce como minería de datos y, que son fundamentales a la hora de orientar a los gerentes en la toma de decisiones. La minería de datos se propone encontrar patrones de comportamiento por grupos, tendencias en el tiempo, relaciones entre variables, escondidas en una base de datos que en principio pareciera no guardar ninguna información sino unos datos sin forma apreciable. Las herramientas que veremos junto con las que se verán en el capítulo de Inteligencia Artificial, permiten comprender mejor la estructura de la información, las relaciones existentes entre grupos de diferentes variables o categorías. La forma que pueden agruparse los individuos formando target naturales que pueden ser de interés para una campaña publicitaria, igualmente permiten la clasificación de observaciones en grupos previamente definidos.

El análisis multivariante de datos tiene su antecedentes en el siglo XX en los trabajos de Fisher entre los cuales se encuentran los problemas de clasificación y discriminación en 1937, Pearson con su trabajo del método de los ejes principales en1901, Spearman con las primeras ideas del análisis factorial en1904, Hotelling que propuso los componentes principales y el análisis de correlación entre grupos de variables entre 1933 y 1936 y más recientemente, los trabajos de Lawley y Maxwell que tratan formalmente desde el punto de vista de la inferencia estadística el análisis factorial en 1971, P. Benzecri que, junto con su colaboradores, estudiaron los datos cualitativos mediante el análisis de correspondencia y publicaron sus resultados en dos tomos (1973). A pesar que muchas ideas fueron desarrolladas en la primera mitad del siglo XX, solo a partir de los años ochenta es cuando

se desarrollan sotfware de estadística que facilitan la aplicación de estos métodos tales como SAS, STATGRAPHICS, SPSS, BMD entre otros. Estos programas son fácilmente instalables en CP y este hecho ha aumentado la popularidad de los métodos multivariantes aplicados al mercadeo, administración de personal, finanzas, medicina etc, se puede decir que en todas las disciplinas en donde se recolecta un número apreciable de variables y observaciones.

Por razones de espacio sólo veremos dos técnicas de reducción de dimensionalidad: componentes principales y correspondencia. La primera técnica se emplea en observaciones en escala métrica y la segunda a datos categóricos. También veremos algunas técnicas de análisis de conglomerados basados en algoritmos jerárquicos y, finalmente el análisis discriminante lineal y de regresión logística, éstas últimas técnicas permiten predecir a que grupo pertenece una observación dado un número de grupos previamente definidos de observaciones, tales como clientes solventes y no solventes. Todas las herramientas se presentan con una breve exposición teórica acompañada de ejemplos con datos de problemas reales que se analizan con salidas del software SPSS y SAS.

I.-INTRODUCCIÓN A LAS TÉCNICAS DE REDUCCIÓN DE LA DIMENSIONALIDAD.

Hay situaciones en donde un archivo de datos está formado por un gran número de variables y de observaciones y se quiere conocer la estructura subyacente de esa masa de datos. Una forma de analizar este gran volumen de datos y transformarlo en información útil para la toma de decisiones, es empleando técnicas de reducción de dimensionalidad, que recojan la máxima información contenida en los datos originales en un número mucho menor de nuevas variables conocidas genéricamente como factores. Esto es, encontrar el conjunto de variables que son relevantes en el conjunto de datos. Las técnicas de reducción de dimensionalidad se clasifican de acuerdo a la escala de medición empleada. Cuando la escala es métrica, una técnica de reducción de dimensionalidad es el análisis de componentes principales (ACP) donde se parte de nuevas variables no observables llamadas componentes principales, que son combinación lineal de las originales y, es posible expresar además las variables originales en combinación de las componentes, por tanto se podrá reproducir los datos originales lo más exactamente posible en la medida que la pérdida de información sea mínima. En la aplicación de esta técnica hay experiencias con escala ordinal que resultan satisfactorias cuando se emplea la escala con más de cinco opciones, es decir, cuando se pueden asignar rangos desde uno hasta siete o más. Otra técnica es el análisis factorial (AF) que consiste en expresar cada variable en función de un conjunto de factores comunes no observables y un factor específico adicional, asociado a una y solamente a una variable, esta técnica no la estudiaremos, pero se puede encontrar ampliamente desarrollada en textos como Hair et al(1999) y Johnson (2000). Cuando la escala es nominal, se emplea el análisis de correspondencia binaria (ACB) o múltiple (ACM). La correspondencia binaria, se refiere a datos que se codifican en dos categorías cada una con varias modalidades y múltiple cuando son más de dos categorías con diferentes modalidades.

II.-ANÁLISIS DE COMPONENTES PRINCIPALES.

El ACP es una técnica que permite descubrir la verdadera dimensión de los datos, cuando están medidos en escala métrica. La idea se basa en considerar la proyección de los datos originales lo más fiel posible a un nuevo espacio con una dimensión menor a la original. Para ello se definen unas nuevas variables que son combinaciones lineales exactas de las variables originales y, además son ortogonales entre si (no correlacionadas), el primer problema es determinar los valores de los pesos tales que recojan la máxima información. En principio se pueden definir tantas nuevas variables como variables originales se posean, pero de ellas tomaremos un número preferiblemente mucho menor al número de variables originales, con la condición de que estas expliquen una buena proporción de la variabilidad contenida en los datos originales y, se pueda por tanto reproducir con estas nuevas variables, la matriz de datos originales lo más exactamente posible. El análisis se puede hacer en dos vías: en el espacio de los individuos u observaciones o análisis en R^p y en el espacio de las variables, o análisis en R^n en otras palabras, si recordamos la matriz de datos explicada en el capítulo anterior, el análisis puede hacerse por filas que representan a las observaciones y cuyas dimensiones están dadas por el número de variables p o, por columnas que representan la variables y cuyas dimensiones están definida por el número de observaciones n.
Luego se pueden proyectar las variables y las observaciones en la nueva dimensión de los componentes principales.
Entre los usos mas importantes del ACP están:
1.-Reducción de la dimensionalidad: encontrar la verdadera dimensión de la población.
2.-Servir de apoyo para otros métodos:
a) Regresión múltiple: cuando hay problema de cuasicolinealidad.
b) Análisis factorial: como análisis preliminar para determinar el número de factores.
c) Análisis de conglomerado: para establecer un primer número de conglomerados.
d) Gráficas de control multivariantes: permite saber si un sistema está bajo control estadístico considerando varias variables simultáneamente.
3.-Detección de valores atípicos en la masa de datos.
El desarrollo de la técnica parte de considerar el conjunto de nuevas variables definidas como:

$$Y_1 = a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + ... + a_{1p}X_p$$
$$Y_2 = a_{21}X_1 + a_{22}X_2 + a_{23}X_3 + ... + a_{2p}X_p$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots$$
$$Y_p = a_{p1}X_1 + a_{p2}X_2 + a_{p3}X_3 + ... + a_{pp}X_p$$

Las variables Y_i son no observables y son ortogonales entre sí, esto es, no están correlacionadas, las variables X_j son las variables originales y se conocen sus valores, los parámetros a_{ij} son los pesos que se obtienen resolviendo el siguiente problema:
Dada la primera variable Y_1 determinamos la varianza de la misma:

$Var(Y_1) = Var(a_{11}X_1 + a_{12}X_2 + a_{13}X_3 + ... + a_{1p}X_p) = Var(a_1X) = a_1^T Sa_1$

S es la matriz de varianzas covarianzas en la muestra, entre las variables X_j y, a_1 es el vector de los pesos. El problema es recoger la máxima variabilidad asociada a las variables X_j en la varianza de la primera variable Y_1 luego, hay que maximizar la varianza de Y_1 . El vector a_1 es norma la unidad, esto es: $a_1^T a_1 = 1$. Por tanto el problema es:

Maximizar

$Var(Y_1) = a_1^T Sa_1$

Sujeto a

$a_1^T a_1 = 1$

Esto equivale como vimos en el segundo capítulo a:

Maximizar

$Var(Y_1) = a_1^T Sa_1 + \lambda(a_1^T a_1 - 1)$

Se demuestra que la condición necesaria de existencia de un punto extremo nos lleva a resolver el siguiente sistema de ecuaciones homogéneas:

$Sa_1 - \lambda a_1 = 0$

Para que el sistema de ecuaciones homogéneas tenga al menos una solución distinta a la trivial debe verificarse que**:** $|S - \lambda I| = 0$. El determinante $|S - \lambda I|$ da origen al polinomio en λ : $f(\lambda)$ **.** Por tanto, si $|S - \lambda I| = 0$ entonces debe verificarse que: $f(\lambda) = 0$**.**

La mayor raíz λ_1 se asocia al primer componente principal y el autovector asociada a la misma son los valores buscado de los pesos a_1 y este vector conforma el factor principal. Además, λ_1 es la varianza del primer componente principal. La covarianza entre Y_1 y una variable X_j es $a_{1j}\lambda_1$ y la correlación entre este componente y la variable X_j es: $a_{1j}\sqrt{\lambda_1 / Var(X_j)}$ esta mide la contribución de las variables en la formación del componente. En vez de emplear la matriz de varianza covarianza S podemos partir de datos estandarizados y utilizar la matriz de correlación R. Por tanto el problema se transforma en:

Maximizar

$Var(Y_1) = a_1^T Ra_1$

Sujeto a

$a_1^T a_1 = 1$

En este caso, la correlación entre la variable X_j y el primer componente Y_1 es $a_{1j}\sqrt{\lambda_1}$, puesto que X_j se ha estandarizado para obtener la matriz de correlación, entonces se pueden emplear estos valores como pesos estandarizados de cada variable en el primer componente principal.

Si queremos usar dos componentes principales para proyectar la masa de datos en dos ejes, debemos plantear un nuevo problema para poder obtener el segundo componente principal, ortogonal al primero. La condición de ortogonalidad se garantiza considerando como restricción $a_2^T a_1 = 0$. Entonces, debemos maximizar la varianza de Y_2 con las condiciones:

$a_2^T a_2 = 1;\ a_2^T a_1 = 0$

Maximizar

$Var(Y_2) = a_2^T S a_2$

Sujeto a:

$a_2^T a_2 = 1$

$a_2^T a_1 = 0$

Esto equivale a:

Maximizar

$Var(Y_2) = a_2^T S a_2 + \lambda(a_2^T a_2 - 1) + \mu a_2^T a_1$

Al diferenciar para aplicar la condición necesaria de la existencia de un óptimo y hacer algunas operaciones algebraicas, obtenemos las siguientes ecuaciones homogéneas:

$Sa_1 - \lambda a_1 = 0$

$Sa_2 - \lambda a_2 = 0$

En general, en el caso de p componentes, partimos de: $SA = \Lambda A$ donde A es la matriz de los pesos y Λ es la matriz diagonal: $\Lambda = Diag(\lambda_1, \lambda_2 ... \lambda_p)$ los pesos se determinan con base a los autovectores asociados a estos autovalores. Los pesos estandarizados se obtienen multiplicando los componentes de cada autovector por la raíz del autovalor: $a_{ij}\sqrt{\lambda_i}$. Estos coeficientes forman lo que se denomina matriz de coeficientes de puntuaciones factoriales o de saturaciones, con tantas filas como variables originales se están considerando y columnas como componentes se han seleccionados.

$F = A\Lambda^{1/2}$

Esta matriz presenta varias propiedades:

1.-Cada elemento $a_{ij}\sqrt{\lambda_i}$ de la matriz es la correlación entre variable y componente y, mide la contribución de cada variable en la formación del componente respectivo. Se conoce como ***contribución absoluta.***

2.- La suma de los cuadrados de los elementos de cada fila $\sum_{j=1}^{k}(a_{ij}\sqrt{\lambda_i})^2 \quad i = 1,2...p$ se denomina comunalidades y, mide la proporción de varianza explicada de cada variable i-ésima por los componentes, luego las comunalidades tienen el mismo significado del coeficiente de determinación visto en regresión múltiple, miden por tanto la buena representación de la variable en el espacio de los componentes principales. Cada correlación al elevarse al cuadrado $(a_{ij}\sqrt{\lambda_i})^2$ mide la proporción de varianza de la variable X_j explicada por el componente Y_i generalmente se denomina ***contribución relativa.***

3.-La suma de los cuadrados de los elementos de cada columna es igual al correspondiente autovalor de cada componente. $\sum_{i=1}^{p}(a_{ij}\sqrt{\lambda_i})^2 = \lambda_i$

4.-Los elementos de las columnas representan las coordenadas en el nuevo espacio de las componentes. Dado que estamos hablando de correlaciones, todas las variables se encontrarán en el espacio de un círculo de radio unitario con centro en el origen.

5.-Los datos originales pueden expresarse en función de los componentes principales:

$X = YA + E$

Donde X es la matriz de datos originales Y es el vector de componentes principales, A es la matriz de autovectores y E es el error debido a que solo se selecciona r<p componentes principales. Si se tomasen todos los componentes E es la matriz nula. Si los datos están bien representados por un número pequeño de componentes los elementos de E deberían ser muy pequeños. Consideremos ahora la matriz diagonal $\Lambda = Diag(\lambda_1, \lambda_2 ... \lambda_p)$, la contribución en la explicación de la variabilidad presente en la matriz X, está dada para cada componente como: $\lambda_i / \sum_{i=1}^{p}\lambda_i$. En efecto, se demuestra que la variación total de X es la $traza(S) = Var(X_1)...Var(X_p) = \sum_{i=1}^{p}\lambda_i$. Esto nos permite tener un criterio para seleccionar el número de componentes a considerar en el análisis. Cuando las correlaciones entre las variables es muy alto se necesitan pocos componentes para representar los datos, generalmente el número no excede de cuatro componentes. Algunos software dan como opción un gráfico que nos ayuda a seleccionar el número de componentes. Este gráfico, muestra una especie de codo tal que a partir de él los componentes son pocos relevantes, por tanto, se selecciona aquel número que se indica antes de la presencia de este codo. El gráfico se construye con los autovalores ordenados de mayor a menor.

La aplicación del método de componentes principales tiene sentido solo si las variables originales están correlacionadas, mientras mayor sean estas correlaciones más útil será la técnica pues se necesitará una dimensión menor para entender la estructura interna de los datos.

Si algunos autovalores nulos entonces existen variables con correlación exactas. Si las correlaciones entre las variables son bajas entonces se necesitará un mayor número de componentes a tal punto que puede perder sentido su aplicación. Esta situación puede darse cuando los datos están medidos en escala ordinal con muy pocas opciones, es el caso de ciertas encuestas en donde las preguntas dan sólo cinco opciones, tal como: muy bien al cual se le asigna el rango de cinco, bien: el rango cuatro, regular: el rango tres, mal: el rango dos y pésimo: el rango de uno y además, el número de pregunta es muy alto y con pocas observaciones. Hay casos en donde el número de observaciones es mucho menor que el de variables y se puede observar cierto grado de correlación pero generalmente bajo; en estos casos se tendrá cuidado de aplicar el método para descubrir existencia de conglomerados porque los dos primeros componentes principales tendrán poca capacidad para representar la masa de datos.

Por otra parte es importante tomar en cuente las unidades que están medidas las variables originales y las varianzas de las mismas a la hora de interpretar los resultados. Si están medidas en las mismas unidades y tienen más o menos la misma variabilidad no hay problema de aplicar los componentes principales partiendo de la matriz de varianza covarianza S o la matriz de correlación R**.** Pero cuando los datos no están medidos en las mismas unidades y las varianzas no son iguales es preferible trabajar con datos estandarizados y por tanto partir de la matriz de correlación. Los autovalores obtenidos a partir de R son diferentes a los obtenidos a partir de S. Las razones de emplear la matriz de correlación R en vez de la matriz de varianza covarianza S son:1) no está afectada por un cambio de escala porque la correlación es invariante a una transformación lineal 2) En el caso de existir fuerte diferencia de variabilidad entre las variables, al emplear la matriz S para obtener los componentes principales, las variables con mayor varianza serán las que tendrán mayor valor en el componente del vector de autovalores y no así en el caso que se emplee la matriz de correlación R. Este último punto está bien desarrollado en Johnson (1998).

Los libros especializados en análisis multivariante, suelen incluir en el tema de componentes principales, la hipótesis de que la población tiene una distribución normal multivariante con vector de media μ y matriz de varianza covarianza Σ y, partiendo de una muestra aleatoria estiman ambas características paramétricas y luego, proponen contraste de hipótesis sobre la significación de los autovalores y de homogeneidad de los datos. El enfoque en este capítulo no incluye este problema de inferencia, de tal suerte, que el método puede ser empleado para observaciones provenientes de una muestra aleatoria, opinática o de un censo. Por tanto, el método de recolección de datos queda al criterio del investigador.

Ejemplo 1 con SPSS.

Se ha efectuado un estudio sobre 45 naciones, considerando catorce variables de las características demográficas, económicas y geográficas. Las variables en estudio son: superficie en Km^2 (X_1), población en millones de habitantes (X_2), población económicamente activa ocupada en porcentaje, discriminada por el nivel educativo en: profesionales (X_3), técnicos (X_4), obreros especializados (X_5), obreros no especializados (X_6), tasa de mortalidad infantil en porcentaje (X_7), tasa de nacimiento en porcentaje (X_8), porcentaje de población alfabeta (X_9), ingreso per capital IPC en dólares(X_{10}), producto interno bruto del sector primario en porcentaje (X_{11}), producto interno bruto del sector secundario en porcentaje (X_{12}), producto interno bruto del sector servicio en porcentaje (X_{13}), pobreza en porcentaje (X_{14}).

El primer análisis que debemos hacer es el análisis de la matriz de correlación de las catorce variables:

CUADRO 1

Matriz de correlaciones

		Obreros-Esp	ingreso-PC	Alfabetismo	Obreros-NE	PIB-Primario	PIB-Secundario	PIB-Servicios	población	POBREZA	Profesionales	Superficie	tasa-de-mort	tasa-de-nac	TECNICOS
Correlación	Obreros- Esp	1,000	,680	,786	-,959	-,924	,913	,748	-,023	-,885	,831	,045	-,715	-,691	,908
	ingreso-PC	,680	1,000	,792	-,800	-,753	,728	,626	-,082	-,792	,876	,059	-,640	-,856	,787
	Alfabetismo	,786	,792	1,000	-,865	-,817	,765	,702	-,334	-,858	,869	-,261	-,827	-,729	,865
	Obreros- NE	-,959	-,800	-,865	1,000	,958	-,924	-,797	,063	,939	-,950	-,003	,754	,794	-,981
	PIB-Primario	-,924	-,753	-,817	,958	1,000	-,900	-,898	-,045	,908	-,898	,009	,695	,830	-,946
	PIB-Secundario	,913	,728	,765	-,924	-,900	1,000	,616	,106	-,878	,848	,132	-,678	-,708	,901
	PIB-Servicios	,748	,626	,702	-,797	-,898	,616	1,000	-,027	-,753	,767	-,149	-,570	-,785	,799
	población	-,023	-,082	-,334	,063	-,045	,106	-,027	1,000	,070	-,117	,541	,075	-,110	-,054
	POBREZA	-,885	-,792	-,858	,939	,908	-,878	-,753	,070	1,000	-,918	,002	,730	,720	-,917
	Profesionales	,831	,876	,869	-,950	-,898	,848	,767	-,117	-,918	1,000	-,007	-,731	-,828	,949
	Superficie	,045	,059	-,261	-,003	,009	,132	-,149	,541	,002	-,007	1,000	,027	,131	-,056
	tasa-de- mort	-,715	-,640	-,827	,754	,695	-,678	-,570	,075	,730	-,731	,027	1,000	,637	-,738
	tasa-de-nac	-,691	-,856	-,729	,794	,830	-,708	-,785	-,110	,720	-,828	,131	,637	1,000	-,810
	TECNICOS	,908	,787	,865	-,981	-,946	,901	,799	-,054	-,917	,949	-,056	-,738	-,810	1,000

Se puede observar que las correlaciones son en su mayoría mayores 0,60 salvo con las variables superficie y población. Esto nos conduce a pensar que la dimensión debe ser menor.

Ahora, veremos cual es grado de variabilidad de cada variable mediante la desviación estándar, recordemos que las varianzas de los componentes principales están en función de la matriz de varianza covarianza de las variables.

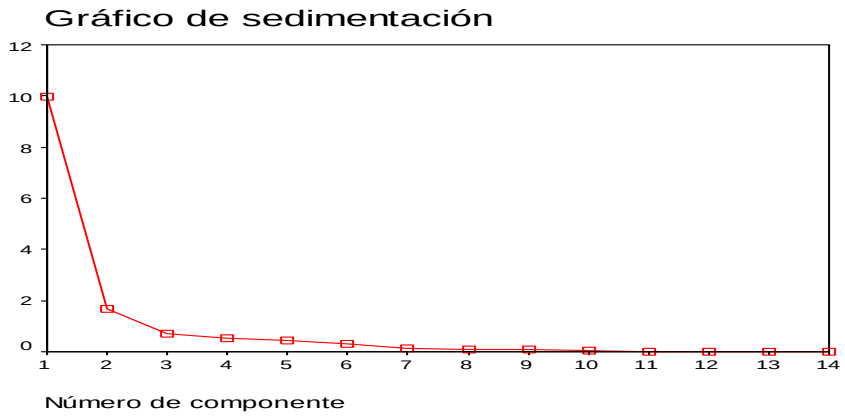
CUADRO 2

Estadísticos descriptivos

	Media	Desviación típica	N del análisis
Obreros- Esp	,3202	,1018	45
ingreso-PC	9789,6806	7929,0018	45
Alfabetismo	73,5889	19,2126	45
Obreros- NE	,3253	,2277	45
PIB-Primario	49,1444	21,8107	45
PIB-Secundario	27,9222	12,1876	45
PIB-Servicios	22,9333	12,0754	45
población	57,4181	49,0654	45
POBREZA	,4158	,2164	45
Profesionales	,1561	7,529E-02	45
Superficie	1419,6583	1855,9903	45
tasa-de- mort	3,715E-03	3,030E-03	45
tasa-de-nac	2,568E-02	1,294E-02	45
TECNICOS	,1984	5,965E-02	45

Como puede verse en el cuadro anterior, la variabilidad de las variables es diferente, además, están medidas en unidades diferentes, por tanto, es recomendable emplear la matriz de correlación y no la de varianza covarianza.

Para determinar el número de componentes, empleamos el gráfico de codo o de sedimentación:



De acuerdo a este gráfico, esperamos que con dos componentes principales se pueda explicar en un buen porcentaje la variabilidad de la masa de datos.

CUADRO 3

Varianza total explicada

Componente	Sumas de las saturaciones al cuadrado de la extracción		
	Total	% de la varianza	% acumulado
1	9,954	71,096	71,096
2	1,685	12,036	83,133

Método de extracción: Análisis de Componentes principales.

Puede verse que los dos componentes principales explican el 83,133%. Ahora, analizaremos los elementos de la matriz de saturaciones.

CUADRO 4

Matriz de componentes[a]

	Componente	
	1	2
Obreros- Esp	,923	8,956E-02
ingreso-PC	,854	2,516E-02
Alfabetismo	,907	-,306
Obreros- NE	-,985	-3,54E-02
PIB-Primario	-,966	-8,91E-02
PIB-Secundario	,905	,216
PIB-Servicios	,830	-5,70E-02
población	-6,24E-02	,877
POBREZA	-,946	-1,84E-02
Profesionales	,961	-1,68E-02
Superficie	-4,31E-02	,866
tasa-de- mort	-,795	4,848E-02
tasa-de-nac	-,857	-3,58E-02
TECNICOS	,974	3,539E-03

Método de extracción: Análisis de componentes principales.

a. 2 componentes extraídos

Los elementos de esta matriz son las correlaciones entre las variables y los componentes principales, por ejemplo: la correlación entre alfabetismo y el primer componente

$a_{13}\sqrt{\lambda_1}$ =0,907 , el cuadrado de esta correlación $(a_{13}\sqrt{\lambda_1})^2$ =0,8227 que es la proporción de la varianza de la variable alfabetismo explicada por el primer componente o contribución relativa. La suma de los cuadrados de los elementos de las columnas es igual al autovalor correspondiente. Para la primera columna tenemos:

$\sum_{i=1}^{14}(a_{i1}\lambda_i)^2 = (0{,}923)^2 + (0{,}854)^2 + (0{,}907)^2 + \ldots + (0{,}974)^2 = 9{,}954$

De igual forma podemos obtener el autovalor del segundo componente.

La suma de los cuadrados $\sum_{j=1}^{2}(a_{ij}\sqrt{\lambda_i})^2$ son las comunalidades.

CUADRO 4

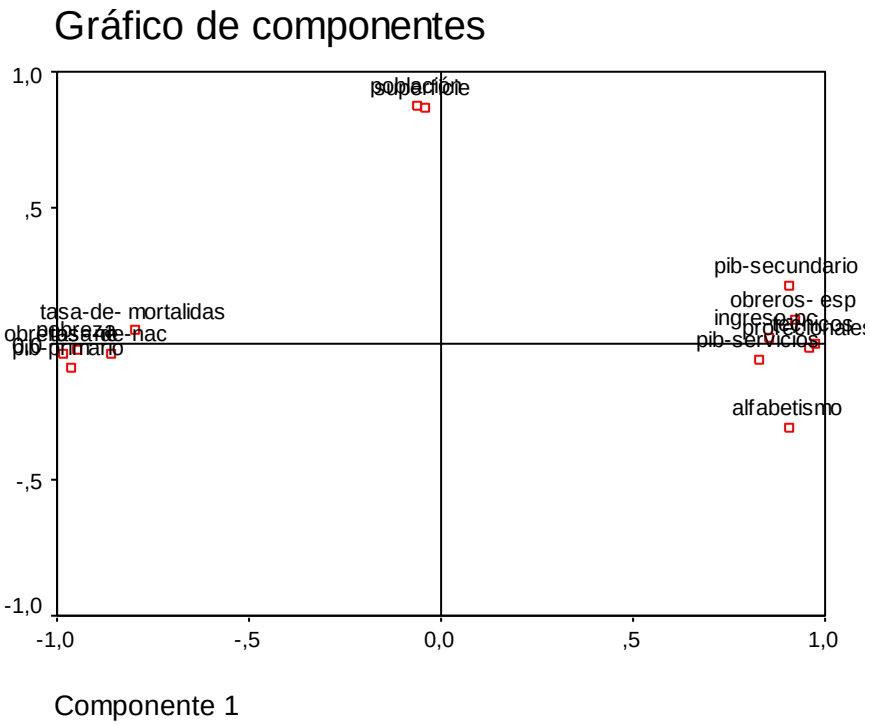
Comunalidades

	Inicial	Extracción
Obreros- Esp	1,000	,859
ingreso-PC	1,000	,729
Alfabetismo	1,000	,917
Obreros- NE	1,000	,971
PIB-Primario	1,000	,940
PIB-Secundario	1,000	,866
PIB-Servicios	1,000	,693
población	1,000	,774
POBREZA	1,000	,895
Profesionales	1,000	,923
Superficie	1,000	,751
tasa-de- mort	1,000	,635
tasa-de-nac	1,000	,736
TECNICOS	1,000	,948

Método de extracción: Análisis de Componentes principales.

Estas se interpretan como la proporción de variabilidad explicada por las componentes seleccionadas de cada una de las variables. Las mejores representadas son:obreros no especializados con el 97,1% de variabilidad explicada por los dos componentes, le sigue el PIB primario con 94%, profesionales con 92,3% y alfabetismos con 91,7%. En general, todas las variables están bien representadas, por tanto, podemos hacer el siguiente gráfico que representa la proyección de las catorce variables en dimensión R^{45} (cada variable está formada por un vector de 45 observaciones) en los dos componentes. Esto es, pasamos de un hiperplano de dimensión cuarenta y cinco a uno nuevo de dimensión dos, que contiene el 83,133% de la información y todas las variables están bien representadas en ese plano.

Este gráfico permite hacer la siguiente interpretación. El primer eje factorial contrapone el PIB primario versus el secundario y el de servicio; obreros no especializados versus obreros especializados, profesionales y técnicos, tasa de mortalidad y tasa de nacimiento versus alfabetismo, pobreza versus IPC: por tanto, este eje se puede asociar al nivel de desarrollo de las naciones consideradas. El segundo están las variables población y superficie, este se puede asociar al tamaño de las naciones.



De igual forma podemos proyectar las cuarenta y cinco observaciones en estos dos ejes factoriales, es decir, reducimos el espacio de R^{14} a R^{2}. Para ello se procede como sigue: Los elementos de la matriz de componentes se dividen entre sus respectivos autovalores obteniéndose la matriz de coeficientes para el cálculo de las puntuaciones en los componentes.

CUADRO 5

Matriz de coeficientes para el cálculo de las puntuaciones en las componentes

	Componente	
	1	2
Obreros- Esp	,093	,053
ingreso-PC	,086	,015
Alfabetismo	,091	-,182
Obreros- NE	-,099	-,021
PIB-Primario	-,097	-,053
PIB-Secundario	,091	,128
PIB-Servicios	,083	-,034
población	-,006	,521
POBREZA	-,095	-,011
Profesionales	,097	-,010
Superficie	-,004	,514
tasa-de- mort	-,080	,029
tasa-de-nac	-,086	-,021
TECNICOS	,098	,002

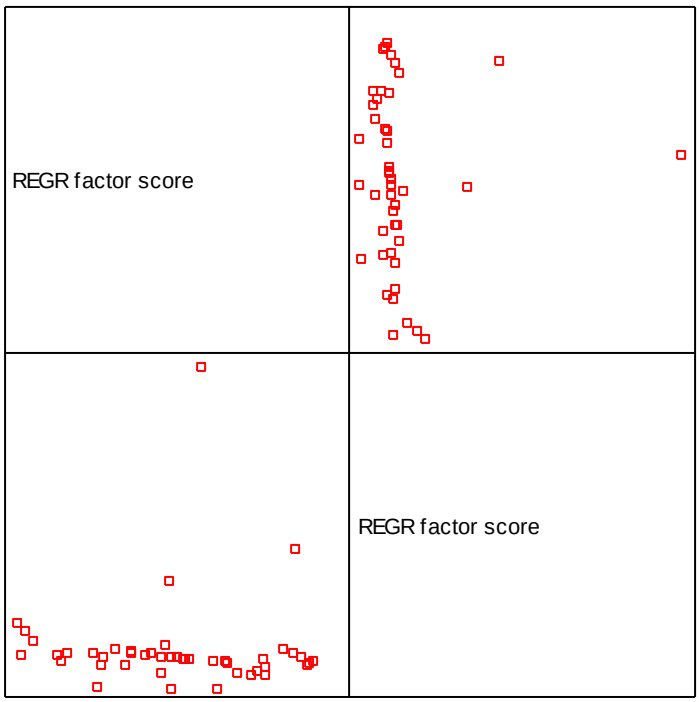
Método de extracción: Análisis de componentes principales.

Y_1 =0,093 obrero especializado + 0,086 ingreso-PC+…..+0.098TËCNICOS

Y_2 =-0,053 obrero especializado + 0,015 ingreso-PC+…..+0.002TËCNICOS

Al sustituir los valores de las variables originales de cada individuo se obtienen las coordenadas (Y_1, Y_2) de los mismos, por tanto tendremos 45 puntos.

Se puede observar que los 45 puntos se sitúan en dos cuadrantes: el positivo y el negativo, en ambos hay tres valores que se alejan de los demás. Una primera conclusión es que hay dos grupos bien diferenciados.



Ejemplo 2 con SAS

Ahora veremos un ejemplo con datos del Banco Central donde se consideran 9 variables y 276 observaciones. En primer lugar obtendremos la matriz de correlación entre todas las variables para ver si vale la pena aplicar la técnica de componentes principales:

CUADRO 6

IPMG **IPMG**	1.0000	0.99936 <.0001	0.99672 <.0001	0.99309 <.0001	0.97274 <.0001	0.97551 <.0001	0.98150 <.0001	0.98241 <.0001	0.98017 <.0001
IPMnac **IPMnac**	0.9993 <.0001	1.00000	0.99370 <.0001	0.98890 <.0001	0.97921 <.0001	0.98106 <.0001	0.98627 <.0001	0.98760 <.0001	0.97919 <.0001
IPMimp **IPMimp**	0.99672 <.0001	0.99370 <.0001	1.00000	0.99827 <.0001	0.95557 <.0001	0.96376 <.0001	0.96699 <.0001	0.96925 <.0001	0.98128 <.0001
TCN **TCN**	0.9930 <.0001	0.98890 <.0001	0.99827 <.0001	1.00000	0.94469 <.0001	0.95365 <.0001	0.95787 <.0001	0.95959 <.0001	0.97809 <.0001
BMN **BMN**	0.9727 <.0001	0.97921 <.0001	0.95557 <.0001	0.94469 <.0001	1.00000	0.98814 <.0001	0.99619 <.0001	0.99612 <.0001	0.95857 <.0001
CDN **CDN**	0.9755 <.0001	0.98106 <.0001	0.96376 <.0001	0.95365 <.0001	0.98814 <.0001	1.00000	0.98397 <.0001	0.99608 <.0001	0.96419 <.0001
M1N **M1N**	0.9815 <.0001	0.98627 <.0001	0.96699 <.0001	0.95787 <.0001	0.99619 <.0001	0.98397 <.0001	1.00000	0.99589 <.0001	0.96704 <.0001
M2N **M2N**	0.98241 <.0001	0.98760 <.0001	0.96925 <.0001	0.95959 <.0001	0.99612 <.0001	0.99608 <.0001	0.99589 <.0001	1.00000	0.96949 <.0001

RIN **RIN**	0.9801 <.0001	0.97919 <.0001	0.98128 <.0001	0.97809 <.0001	0.95857 <.0001	0.96419 <.0001	0.96704 <.0001	0.96949 <.0001	1.00000

Como puede observarse existe una altísima correlación entre todos los pares de variables. Por tanto la dimensión verdadera de los datos debe ser mucho menor. De haber sido lo contrario, esto es, bajas correlaciones, se pondría en duda el uso de los componentes principales.

The PRINCOMP Procedure

Este procedimiento corresponde al software SAS en el módulo de análisis multivariante. El primer resultado es la varianza generalizada o total de las observaciones.

Total Variance 6.3410714E13

En la siguiente tabla se muestran tres autovalores, en donde además se calcula la diferencia entre ellos y la proporción de cada uno de ellos.

CUADRO 7

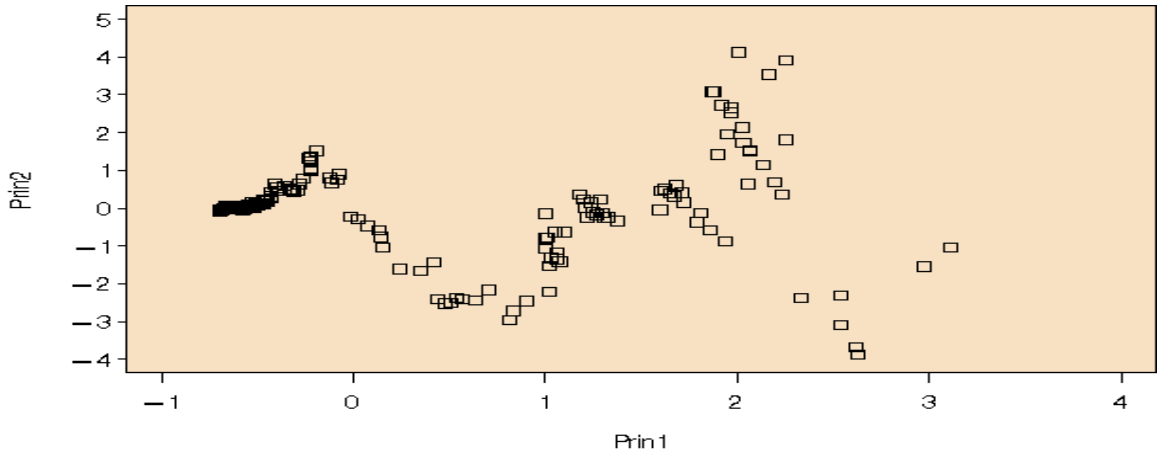
Eigenvalues of the Covariance Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	6.2473E13	6.16689E13	0.9852	0.9852
2	8.04181E11	6.8372E11	0.0127	0.9979
3	1.20461E11		0.0019	0.9998

El primer autovalor es: 6.2473E13 y la proporción de variabilidad de los datos originales explicado por el primer componente de acuerdo al auto valor es 98,52%. El segundo auto valor es 8.04181E11 y la variabilidad explicada por el segundo componente es de apenas 1,27 % que es casi despreciable. Por tanto podemos considerar que con el primer componente es suficiente para interpretar los datos. La cuarta columna de la tabla indica la proporción explicada acumulada, con los tres componentes se explica 99,98 %. Con estos tres componentes podríamos reconstruir casi toda la información contenida en los datos originales, formada por 276 observaciones sobre 9 variables.

CUADRO 8

Eigenvectors				
		Prin1	Prin2	Prin3
IPMG	IPMG	0.000009	-.000004	0.000004
IPMnac	IPMnac	0.000009	-.000002	0.000004
IPMimp	IPMimp	0.000008	-.000008	-.000000
TCN	TCN	0.000039	-.000046	0.000002
BMN	BMN	0.230947	0.203959	0.267781
CDN	CDN	0.337081	0.218510	-.707052
M1N	M1N	0.329300	0.182695	0.652189
M2N	M2N	0.666381	0.401206	-.054863
RIN	RIN	0.529658	-.846351	-.003239

En la tabla anterior tenemos los tres autovectores. El primero da las estimaciones de los coeficientes del primer componente principal esto es Y= 0.000009IPMG + 0.000009IPMnac + 0.000008 IPMimp + 0.000039 TCN+ 0.230947BMN+ 0.337081CDN+ + 0.329300 M1N + 0.666381M2N + 0.529658 RIN. De igual forma se escribe el segundo y tercer componente. Las correlaciones entre las variables y el componente principal es proporcional a estos coeficientes de tal forma que podemos deducir cuales son las variables que más contribuyen en la formación de cada componente, para el primero notamos que la variable que más contribuye es M2N, la siguiente RIN y así sucesivamente. El primer componente se interpreta como sigue: todos los valores tienen el mismo signo, pero no todas las correlaciones de las variables con el componente son altas, si lo fueran este componente se interpretaría como un promedio ponderado de todas las variables Peña (2002). Los siguientes componentes se interpretan como ¨forma¨. De acuerdo a los signos las variables se contraponen.



El gráfico muestra las 276 observaciones proyectadas en los dos primeros componentes. Es posible detectar en él agrupamiento de observaciones.

Ejemplo 3.

Ahora veremos una aplicación de ACP en el modelo de regresión. En el capítulo anterior estudiamos el modelo M2N= $\beta_0 + \beta_1$ BMN + β_2 CDN+ β_3 IPMG+ ε . En este modelo existe una alta correlación entre las variables explicativas, por tanto es plausible la sospecha de la presencia de cuasi colinealidad. Una solución propuesta en este caso es el uso de los componentes principales sobre las variables explicativas. El resultado es el siguiente:

CUADRO 9

Pearson Correlation Coefficients, N = 276 Prob > |r| under H0: Rho=0			
	IPMG	**BMN**	**CDN**
IPMG **IPMG**	1.00000	0.97274 <.0001	0.97551 <.0001
BMN **BMN**	0.97274 <.0001	1.00000	0.98814 <.0001
CDN **CDN**	0.97551 <.0001	0.98814 <.0001	1.00000

Puede observarse que en efecto hay una alta correlación entre las variables explicativas.

CUADRO 10

Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	2.95760166	2.92692785	0.9859	0.9859

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
2	0.03067381		0.0102	0.9961

De acuerdo a la tabla anterior basta usar solamente el primer componente principal para obtener la nueva ecuación de regresión. Se puede observar que el resultado ilustra lo dicho con respecto a las correlaciones entre las variables originales, en este caso son muy altas por tanto la verdadera dimensión debe ser menor.

CUADRO 11

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7.644578E15	7.644578E15	102785	<.0001
Error	274	2.037859E13	74374414381		
Corrected Total	275	7.664957E15			

CUADRO 12

Root MSE	272717	R-Square	0.9973
Dependent Mean	3793179	Adj R-Sq	0.9973
Coeff Var	7.18966		

CUADRO 13

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	3793179	16416	231.07	<.0001
Prin1		1	5272425	16445	320.60	<.0001

CUADRO 14

Durbin-Watson D	0.148
Number of Observations	276
1st Order Autocorrelation	0.906

El análisis de los resultados se deja al lector. Lo que si debemos tener presente es que podemos encontrar los estimadores de los parámetros del modelo de regresión lineal original, dado como:

$$\hat{\beta}_1 = \lambda_1^{-1} a_1^T X^T Y a_1$$

$$\hat{\beta}_2 = \lambda_1^{-1} a_1^T X^T Y a_2 + \lambda_2^{-1} a_2^T X^T Y a_2$$

El primer vector de estimadores sólo considera el autovector y el autovalor del primer componente, el segundo vector de estimadores, considera los autovectores y autovalores de dos primeros componentes principales.
Estos datos provienen de observaciones temporales por tanto hay que estudiar si hay presencia de autocorrelación entre los residuos.
En el análisis de conglomerados veremos otro ejemplo de aplicación del análisis de componentes principales.
El siguiente punto es una técnica que permite encontrar la verdadera dimensión de datos categóricos.

III.-CORRESPONDENCIA

Es una técnica de reducción de la dimensionalidad para variables categóricas (nominales) basada en tablas de contingencias. Si la tabla tiene sólo dos vías de clasificación (dos categorías), entonces se trata de correspondencia simple; si la tabla tiene más de dos vías de clasificación: correspondencia múltiple.
La técnica de reducción de la dimensionalidad está basada en la distancia chi cuadrado (χ^2) Empezaremos a exponer el análisis de correspondencia binaria, para esto partimos de la siguiente tabla de contingencia:

TABLA1

	B_1	B_2	B_3		B_s	
A_1	n_{11}	n_{12}	n_{13}		n_{1s}	$N_1.$
A_2	n_{21}	n_{22}	n_{23}		n_{2s}	$N_2.$
A_3	n_{31}	n_{32}	n_{33}		n_{3s}	$N_3.$
A_r	n_{r1}	n_{r2}	n_{r3}		n_{rs}	$N_r.$
Total	$N_{.1}$	$N_{.2}$	$N_{.3}$		$N_{.s}$	N

De esta tabla de contingencia en donde n_{ij} representa los valores observados en la celda i,j y $N._j$ son las sumas por filas y $N_i.$, la suma por columnas. Transformamos esta tabla de frecuencias absolutas n_{ij} a frecuencia relativas $f_{ij} = n_{ij} / N$ y la denotamos por F .

TABLA2

	B_1	B_2	B_3		B_s	Total
A_1	f_{11}	f_{12}	f_{13}		f_{1s}	$f_1.$
A_2	f_{21}	f_{22}	f_{23}		f_{2s}	$f_2.$
A_3	f_{31}	f_{32}	f_{33}		f_{3s}	$f_3.$
A_r	f_{r1}	f_{r2}	f_{r3}		f_{rs}	$f_r.$
Total	$f._1$	$f._2$	$f._3$		$f._s$	1

Definiremos dos nuevas matrices D_r y D_c que son matrices diagonales de los promedios de filas y columnas, dadas como: $D_r = diag(f_1., f_2.,... f_r.)$ y $D_c = diag(f._1 f._2, .,... f._s)$. Queremos proyectar los puntos en un espacio de menor dimensión con la condición de respetar la importancia de las celdas de acuerdo al número de observaciones que contienen y además, que los puntos con las mismas estructura estén los más cerca posible y los de estructura distintas lo más alejados. Para lograr la primera condición definimos dos nuevas matrices $R = D_r^{-1} F$ y $C = D_c^{-1} F$, la primera conforma lo que se denomina perfil fila y la segunda perfil columna, ahora se ha obtenido las frecuencias relativas condicionadas a los totales de filas y columnas en donde se conserva la importancia de cada celda por fila y columna. En el análisis de correspondencia binaria tratamos con dos espacios: filas y columna.

Perfil fila:

$$R = D_r^{-1} F$$

Perfil columna

$C = D_c^{-1} F$

Llamamos $r_i = (f_{i1} / f_i., f_{i2} / f_i .... f_{ir} / f_i.)$, al vector fila de la matriz **R**

y $c_j = (f_{1j} / f._j, f_{2j} / f._j .... f_{sj} / f._j)^T$ al vector columna de C.

Con esta información haremos las representaciones de las filas y las columnas en espacio de dimensiones menores.

1.-Proyección de las filas.

En una matriz rxs, las filas están en el espacio s y deseamos proyectar la información contenida en ellas en una dimensión menor. El problema se podría plantear partiendo de las

similitudes entre las mismas, tomando directamente la distancia euclidiana de las frecuencias relativas como una medida de variabilidad contenidas en las celdas, pero esto tiene el inconveniente que las frecuencias relativas no muestran adecuadamente la estructura subyacente de las filas, por tanto hay que recurrir a otro tipo de distancia que considere la importancia relativa de cada celda. Esto se logra definiendo las distancias entre dos filas cualesquiera r_a y r_b como:

$$D^2(r_a, r_b) = (r_a - r_b)D_c^{-1}(r_a - r_b) = \sum_{j=1}^{r}(f_{aj} / f_{a\cdot} - f_{bj} / f_{b\cdot})^2 / f_{\cdot j} =$$

$$\sum_{j=1}^{r}(f_{aj} / f_{a\cdot}\sqrt{f_{\cdot j}} - f_{bj} / f_{b\cdot}\sqrt{f_{\cdot j}})^2$$

Al hacer esta transformación sobre las frecuencias relativas, obteniendo: $f_{ij} / f_{i\cdot}\sqrt{f_{\cdot j}}$, la expresión final no es otra cosa que la distancia euclidiana. Esta distancia calculada con la transformación sí posee la propiedad de invarianza si se agrupan o desagregan unidades homogéneas. Esto es, definimos las distancias tomando las frecuencias relativas condicionadas por las filas. La matriz original de datos ha sido transformada en una matriz cuyos componentes son los obtenidos mediante la transformación. Denotaremos esta matriz por Y= $\left\{f_{ij} / f_{i\cdot}\sqrt{f_{\cdot j}}\right\}$, cuyos elementos son las frecuencias relativas condicionadas por las filas pero estandarizadas por la variabilidad asociada a las columnas, medida por la raíz del promedio de la columna correspondiente, haciendo posible que las celdas sean comparables entre sí. Es fácil comprobar que :

$$Y = RD_c^{-1/2} = D_r^{-1}FD_c^{-1/2}$$

Ahora, debemos encontrar la mejor proyección de los puntos filas de dimensión s en un espacio de dimensión menor, tal que las filas similares queden próximas y las filas diferentes queden alejadas y de esta forma reproducir las mismas características contenidas en la tabla original. Esta proyección se logra al ponderarlo por los promedios de las filas representados en la matriz diagonal, permitiendo representar mejor en la proyección aquellas filas que tienen mayor peso. El vector a es la dirección de norma la unidad ($a^T a = 1$) tal que el vector de puntos proyectados sobre esta dirección recoja la máxima variabilidad contenida en el espacio de las filas.

Por tanto el problema es:
Maximizar:

$$a^T Y^T D_r Y a$$

sujeto a

$$a^T a = 1$$

Se demuestra que el primer valor propio siempre es igual a la unidad, por tanto se toman en cuenta los siguientes valores propios o autovalores $\lambda_i \; i = 2,3...s$ y sus respectivos autovectores a_i. La primera proyección es:

$$P_r(a_1) = Ya_1$$

$P_r(a_1)$ se llama el primer factor y a_1 el eje factorial. Los componentes del vector resultante son las coordenadas de las filas en el primer eje factorial.

Si se desea proyectar en dos dimensiones, esto es, en R^2 entonces hay que tomar en cuenta el siguiente valor propio con su respectivo autovector y definir la proyección bidimensional como:

$$P_r(a_1, a_2) = YA$$

Donde A es una matriz de dos columnas formadas por los autovectores (a_1, a_2), obteniéndose la proyección en dos ejes factoriales. En general se puede proyectar hasta el espacio R^{s-1} definiendo la protección YA, donde $A = (a_1, a_2 .... a_{s-1})$.

2.-Proyección de las columnas.

Para la proyección de las columnas planteamos el problema de forma similar que la proyección de filas haciendo los cambios pertinentes. La distancia entre dos columnas cualesquiera se define como:

$$D^2(c_a, c_b) = (c_a - c_b) D_f^{-1} (c_a - c_b) = \sum_{i=1}^{s} (f_{ia} / f_{\cdot a} - f_{ib} / f_{\cdot b})^2 / f_i.$$

Ahora el problema es, recoger la máxima variabilidad en la dirección del vector **b** de norma la unidad $(b^T b = 1)$. Definimos la matriz $Z = D_c^{-1} F^T D_f^{-1/2}$,donde los elementos de la matriz Z son de la forma $\left\{ f_{ij} / f_{\cdot j} \sqrt{f_{i\cdot}} \right\}$ esto es las frecuencias relativas condicionadas por el promedio de las columnas, estandarizadas por las raíces del promedio de la fila correspondiente, logrando el mismo efecto que en el caso de las filas. Para recoger la máxima variabilidad en la proyección del nuevo espacio, se requiere resolver el problema:
Maximizar:

$$b^T Z^T D_c Z b$$

sujeto a:

$$b^T b = 1$$

La proyección en la dirección del primer vector propio o eje factorial b_1 es:

$$P(b_1) = Z b_1$$

Los componentes de este vector son las coordenadas en el primer factor asociado a las columnas. Para una segunda proyección se procede igual que en el caso de las filas.

3.-Análisis conjunto.

Se demuestra que existe una relación entre los valores propios de los dos espacios, de tal forma que obteniendo el valor propio del espacio de dimensión menor, se obtiene el correspondiente al otro espacio.

4.-Inercia.

La bondad del análisis se estudia mediante la inercia o variabilidad explicada de la misma forma que en el análisis de componentes principales. Esta se mide mediante el estadístico

Chi cuadrado $\chi^2 = n\sum_{i=2}^{k}\lambda_i^2$ en donde λ_i son los autovalores desde $i = 2,3...k$ i. La inercia total explicada es: χ^2 / n , la proporción que recoge cada factor es: λ_i^2 .

Ejemplo 4:
Supongamos que se realizó una investigación sobre la tenencia de tarjeta de crédito (si,no) de los clientes y la carga familiar (0,1..6) de un banco con la finalidad de conocer la relación entre estas dos categorías. Se propuso como método el análisis de correspondencia. Se empleó como herramienta computacional el software SAS.

CUADRO 15

The CORRESP Procedure

Contingency Table

	0	1	2	3	4	5	6	no	si	Sum
0	42	0	0	0	0	0	0	34	8	84
1	0	18	0	0	0	0	0	12	6	36
2	0	0	20	0	0	0	0	14	6	40
3	0	0	0	9	0	0	0	4	5	18
4	0	0	0	0	12	0	0	6	6	24
5	0	0	0	0	0	4	0	0	4	8
6	0	0	0	0	0	0	5	1	4	10
no	34	12	14	4	6	0	1	71	0	142
si	8	6	6	5	6	4	4	0	39	78
Sum	84	36	40	18	24	8	10	142	78	440

Esta tabla se conoce como tabla de Burt. Ella muestra varias tablas superpuestas de los dos atributos: carga familiar y posesión de tarjeta de crédito. Los atributos por separados forman dos matrices diagonales que indican el total de cada modalidad por atributo. Por ejemplo, los clientes que tienen tarjetas de créditos están totalizados en los dos últimos elementos de la diagonal principal: no tienen 71 casos y si tienen 39 cacos. Las filas con la notación: no y si interceptadas con las columnas desde cero a seis es la tabla cruzada de las dos categorías, por ejemplo, la intercepción entre la fila “si” y la columna 4 tiene seis observaciones.

A continuación se presenta el perfil fila, esto es: $R = D_r^{-1}F$ donde se verifica que la suma por fila es igual a la unidad. Se puede observar que la diagonal principal tiene el mismo valor y obviamente no pueden existir valores cruzados entre las modalidades.

CUADRO 16

Row Profiles									
	0	1	2	3	4	5	6	no	si
0	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.404762	0.095238
1	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.333333	0.166667
2	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.350000	0.150000
3	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000	0.222222	0.277778
4	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000	0.250000	0.250000
5	0.000000	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000	0.500000
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.500000	0.100000	0.400000
no	0.239437	0.084507	0.098592	0.028169	0.042254	0.000000	0.007042	0.500000	0.000000
si	0.102564	0.076923	0.076923	0.064103	0.076923	0.051282	0.051282	0.000000	0.500000

De forma similar, el cuadro muestra el perfil columna, esto es $C = D_c^{-1}F$ donde se verifica que la suma por columna es igual a la unidad.

CUADRO 17

Column Profiles									
	0	1	2	3	4	5	6	no	si
0	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.239437	0.102564
1	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.000000	0.084507	0.076923
2	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000	0.000000	0.098592	0.076923
3	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000	0.000000	0.028169	0.064103
4	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000	0.042254	0.076923
5	0.000000	0.000000	0.000000	0.000000	0.000000	0.500000	0.000000	0.000000	0.051282
6	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.500000	0.007042	0.051282
no	0.404762	0.333333	0.350000	0.222222	0.250000	0.000000	0.100000	0.500000	0.000000
si	0.095238	0.166667	0.150000	0.277778	0.250000	0.500000	0.400000	0.000000	0.500000

En el cuadro siguiente tenemos: la primera columna muestra los valores de los autovalores y la segunda los cuadrados de los mismos que es la parte de inercia explicada por cada factor. La siguiente columna es el valor de la chi cuadrado que se obtiene al multiplicar la

inercia por el total de observaciones de la tabla de Burt, esto es: 440x1,83887=809,104. Los grados de libertad se obtienen también de la misma tabla como: número de fila menos uno por el número de columna menos uno: $(f-1)(c-1)=gl$. Si se considera la inercia total y los grados de libertad, se concluye que para un nivel de significación del 5% se rechaza la hipótesis nula de independencia entre los atributos y esto es una condición para el buen uso del análisis de correspondencia. Luego tenemos dos columnas que indican respectivamente el porcentaje y el porcentaje acumulado de inercia explicado por los factores

CUADRO 18

The CORRESP Procedure

Inertia and Chi-Square Decomposition

Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	5 10 15 20 25 ----+----+----+----+----+---
0.71080	0.50523	222.303	27.48	27.48	**********
0.50000	0.25000	110.000	13.60	41.07	**********
0.50000	0.25000	110.000	13.60	54.67	**********
0.50000	0.25000	110.000	13.60	68.26	**********
0.50000	0.25000	110.000	13.60	81.86	**********
0.50000	0.25000	110.000	13.60	95.45	**********
0.28920	0.08364	36.800	4.55	100.00	*****
Total	1.83887	809.104	100.00		

Degrees of Freedom = 64

El cuadro siguiente muestra las coordenadas de los perfiles filas en los tres primeros ejes factoriales, en este caso la variabilidad total que se está tomando es 54,67%.

CUADRO 19

Row Coordinates

	Dim1	Dim2	Dim3
0	-0.5782	-0.0418	-0.3478
1	-0.0748	-0.1979	1.5846
2	-0.1922	-0.1615	-0.3447

Row Coordinates			
	Dim1	Dim2	Dim3
3	0.7084	2.2615	-0.0000
4	0.5126	-0.3800	-0.3391
5	2.2748	-0.9263	-0.3250
6	1.5699	-0.7078	-0.3306
no	-0.5268	0.0000	0.0000
si	0.9591	-0.0000	-0.0000

El cuadro siguiente se presenta en primer lugar la calidad de la representación de cada modalidad en los tres ejes factoriales (dimensión 1, 2 y 3) como: la suma de los cuadrados de los cosenos, se puede observar que la modalidad mejor representada en los tres ejes factoriales las cargas familiares uno (0.9996) y tres (0.9854). La masa representa la frecuencia relativa de cada modalidad en el perfil fila. Finalmente la inercia indica la proporción de la inercia explicada por cada modalidad a la inercia total.

CUADRO 20

Summary Statistics for the Row Points			
	Quality	Mass	Inertia
0	0.5264	0.1909	0.0901
1	0.9996	0.0818	0.1137
2	0.0806	0.0909	0.1116
3	0.9854	0.0409	0.1268
4	0.1264	0.0545	0.1225
5	0.4335	0.0182	0.1400
6	0.2812	0.0227	0.1351
no	0.8580	0.3227	0.0568
si	0.8580	0.1773	0.1033

Las contribuciones parciales a la inercia para los puntos filas en cada eje se muestra a continuación, éstas se interpretan como la contribución de cada modalidad en la formación de cada factor.

CUADRO 21

Partial Contributions to Inertia for the Row Points			
	Dim1	Dim2	Dim3
0	0.1263	0.0013	0.0924
1	0.0009	0.0128	0.8217
2	0.0066	0.0095	0.0432
3	0.0406	0.8369	0.0000
4	0.0284	0.0315	0.0251
5	0.1862	0.0624	0.0077
6	0.1109	0.0455	0.0099
no	0.1773	0.0000	0.0000
si	0.3227	0.0000	0.0000

Los cuadrados de los cosenos son las contribuciones relativas, esto es, la proporción de variabilidad explicada de cada modalidad por cada factor (Dim i-ésimo) y da la calidad de la representación en cada factor, la suma de los cuadrados de los cosenos muestra la calidad de representación de cada modalidad en el nuevo espacio de tres dimensiones.

CUADRO 22

Squared Cosines for the Row Points			
	Dim1	Dim2	Dim3
0	0.3851	0.0020	0.1393
1	0.0022	0.0153	0.9821
2	0.0164	0.0116	0.0527
3	0.0881	0.8974	0.0000
4	0.0636	0.0350	0.0278
5	0.3654	0.0606	0.0075
6	0.2254	0.0458	0.0100
no	0.8580	0.0000	0.0000
si	0.8580	0.0000	0.0000

Por ejemplo, la modalidad de una carga familiar de seis personas, el cuadrado del coseno en el primer factor es 0,2254 es decir, la proporción explicada de esta modalidad por este factor es 22,54%. La suma de los cuadrados de los cosenos es 0,2812 y esta cantidad indica la calidad de representación de esta modalidad en los tres factores. Una buena representación debe estar cercana a la unidad. Las modalidades mejor representadas en este nuevo espacio de tres dimensiones son: para la categoría carga familiar las modalidades de una carga familiar con un valor de la suma de los cuadrados de los cosenos de 0,9996 seguida con tres cargas familiares con 0,9855. Para la categoría posee o no tarjeta de crédito las dos modalidades están bien representadas, cada una con el valor de la suma de 0,850.

De la misma forma se interpreta el espacio de las columnas, lo que dejamos al lector.

CUADRO 23

Column Coordinates			
	Dim1	Dim2	Dim3
0	-0.5782	-0.0418	-0.3478
1	-0.0748	-0.1979	1.5846
2	-0.1922	-0.1615	-0.3447
3	0.7084	2.2615	0.0000
4	0.5126	-0.3800	-0.3391
5	2.2748	-0.9263	-0.3250
6	1.5699	-0.7078	-0.3306
no	-0.5268	0.0000	0.0000
si	0.9591	0.0000	0.0000

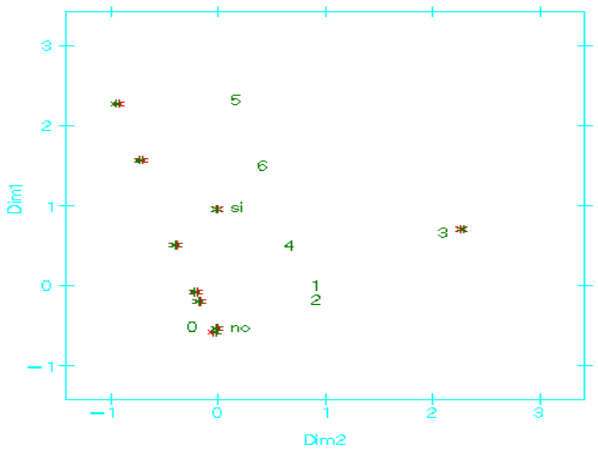
Summary Statistics for the Column Points			
	Quality	Mass	Inertia
0	0.5264	0.1909	0.0901
1	0.9996	0.0818	0.1137
2	0.0806	0.0909	0.1116
3	0.9854	0.0409	0.1268
4	0.1264	0.0545	0.1225
5	0.4335	0.0182	0.1400

Summary Statistics for the Column Points			
	Quality	Mass	Inertia
6	0.2812	0.0227	0.1351
no	0.8580	0.3227	0.0568
si	0.8580	0.1773	0.1033

Partial Contributions to Inertia for the Column Points			
	Dim1	Dim2	Dim3
0	0.1263	0.0013	0.0924
1	0.0009	0.0128	0.8217
2	0.0066	0.0095	0.0432
3	0.0406	0.8369	0.0000
4	0.0284	0.0315	0.0251
5	0.1862	0.0624	0.0077
6	0.1109	0.0455	0.0099
no	0.1773	0.0000	0.0000
si	0.3227	0.0000	0.0000

Squared Cosines for the Column Points			
	Dim1	Dim2	Dim3
0	0.3851	0.0020	0.1393
1	0.0022	0.0153	0.9821
2	0.0164	0.0116	0.0527
3	0.0881	0.8974	0.0000
4	0.0636	0.0350	0.0278
5	0.3654	0.0606	0.0075
6	0.2254	0.0458	0.0100
no	0.8580	0.0000	0.0000
si	0.8580	0.0000	0.0000

El grafico siguiente muestra la proyección de las modalidades de los dos atributos en el nuevo espacio.



IV.-ANÁLISIS DE CONGLOMERADO

Una de las preocupaciones de los investigadores, es el problema de clasificar bien objetos o personas. Es común en investigación de mercado hacer una clasificación de los clientes actuales y potenciales describiendo sus características y así definir estrategias que le permita mantener los clientes activos y en lo posible incorporar a su cartera nuevos clientes. Igualmente puede pensarse en clasificar las instituciones financieras con base a los índices que se obtienen de los estados financieros. Esta clasificación puede hacerse a priori bajo un criterio predefinido o empleando algoritmos que permitan hacer una clasificación a posteriori. Al clasificar los objetos o personas, se obtiene un número de grupos llamados conglomerados o cluster en inglés. Consideremos que poseemos una matriz de datos **X** con p variables y n observaciones, el problema es determinar si es posible agrupar las n observaciones de acuerdo a los valores de las p variables en varios grupos lo más homogéneos posible dentro de ellos y los mas separados posible entre los grupos. Esto se llama formar cluster o conglomerados de individuos. También se pueden agrupar las p variables de acuerdo a las n observaciones en grupos de variables, esto se llama cluster de variables.

No existe para este problema una única solución, esto es, existe un gran número de algoritmos que fijado el número de conglomerados que se desean, puede ocurrir que al utilizarlos todos den diferentes conglomerados en cuanto a los individuos que los integran. Tampoco existe un criterio que indique cuantos grupos se pueden formar. Por tanto, el número de grupos y el algoritmo a emplear para obtener esos grupos dependen en buena medida de la experiencia que tiene el investigador sobre el problema.

Sin embargo, es posible ayudarse en cuanto al número de conglomerados empleando la técnica de componentes principales o cualquier otra técnica de reducción de dimensionalidad y graficar las coordenadas de los individuos en los dos primeros ejes factoriales y, visualmente fijar un número tentativo de conglomerados.

Sobre los algoritmos tan solo indicaremos en que se basan, sin entrar en mayores detalles. Los algoritmos para definir conglomerados de individuos se dividen en jerárquicos y no jerárquicos.

1.-El no jerárquico consiste en definir previamente k puntos como centro de gravedad o semillas y luego se calculan las distancias de los individuos a estas semillas agrupándolos de acuerdo a su cercanía con los mismos, una vez obtenidos estos conglomerados se recalculan los centro de gravedad y las distancias y se obtienen nuevos conglomerados, el algoritmo se repite hasta tener grupos o conglomerados estables, esto es, ya no es posible redefinir nuevos conglomerados. Este procedimiento se llama de k medias. Un inconveniente de este algoritmo es que la escogencia de las semillas es en cierto sentido arbitrario, de tal forma que dos investigadores pueden llegar a conglomerados muy diferentes en cuanto a los elementos que los integran, una variante del algoritmo anterior es el desarrollado por E. Diday conocido como nube dinámica.

Estos tipos de algoritmos no siempre están disponibles en los software de estadística, entre los que los ofrecen está el software SAS que facilita el uso de varios procedimientos.

2.-El otro grupo de algoritmo se denomina jerárquico, y está basado en definir la distancia entre los puntos tomando en cuenta diferentes criterios, que no son otra cosa que la forma de definir distancias entre conjuntos (distancia de Hausdorff). Para cada criterio se obtiene un árbol de jerarquía o dendrograma y la pertenencia de los individuos a cada conglomerado. El dendrograma puede ayudar a definir el número de conglomerado y de esta forma hacer varias pruebas hasta obtener conglomerados que satisfagan al investigador. Supongamos que tenemos dos conjuntos de puntos o conglomerado A y B entonces los criterios más usuales o distancias son:

2ª-Criterio de vecinos más próximos

$$H(A;B) = \min\big(d(x,y); x \in A, y \in B\big)$$

Esto es, se van agrupando los elementos más cercanos

2b-Los vecinos más lejanos:

$$H(A;B) = \max\big(d(x,y); x \in A, y \in B\big)$$

Se obtienen los conglomerados descartando los más alejados.

2c-Centroide:

$$H(A;B) = d(\mu_A^*, \mu_B^*)$$

Donde μ_A^*, μ_B^* son las medias de los respectivos conglomerados. El centoide está dado por: $(n_A\mu_A^* + n_B\mu_B^*)/(n_A + n_B)$ y n_A, n_B son el número de elemento que contiene cada conglomerado.

2d.-Mediana

Se asemeja el anterior, pero en vez de usar media utiliza la mediana, se recomienda cuando los tamaños de los conglomerados son muy diferentes.

2e.-Distancia promedio

$$H(A,B) = \sum_i^{na}\sum_j^{nb} d(x,y)/n_A n_b$$

2f.-Mínima varianza de Ward.

Este se basa en descomponer la variabilidad total de la masa de datos en dos componentes: la variabilidad entre los grupos y la variabilidad dentro de los grupos.

Hay otros criterios que ofrecen los software como opciones. Además, el problema puede plantearse como un modelo de programación entera binaria pero no está disponible en los paquetes más usuales de estadística.

Dado los diferentes criterios que se pueden aplicar a un mismo conjunto de datos es importante tener también varias medidas que permitan determinar cuál es el número "satisfactorio" de conglomerados que se pueden formar con los mismos y cual de los criterios empleados luce mejor. Consideremos el vector de media μ^* formado por las medias de las diferentes variables en el conjunto de datos, igualmente a cada conglomerado le asociamos la media: μ_k^*. La variabilidad del conjunto de datos y entre los conglomerados la definimos como sigue; dados la n vector de observaciones X_i entonces:

$T = \sum_{i=1}^{n} \left\| X_i - \mu^* \right\|^2$: variabilidad total de la masa de puntos

$W_k = \sum_{i}^{nk} \left\| X_{ik} - \mu_k^* \right\|^2$: variabilidad dentro del conglomerado k de la masa de puntos

$P_G = \sum_{k=1}^{g} W_k$ suma de la variación dentro los conglomerado hasta el conglomerado g.

La primera medida de ajuste es el pseudo estadístico F dado como:

$$F = (T - P_G)(n - g)/(g - 1)P_G$$

Si los datos fuesen asignados aleatoriamente dentro de los conglomerados este estadístico tendría una distribución F con g-1 y n-g grados de libertad, pero como los datos son asignados a los conglomerados por un mecanismo determinístico esta F no es propiamente un estadístico. Sabemos por otra parte que el número máximo de conglomerados que se puede formar es igual al número de objetos y el número mínimo es un conglomerado con todos los objetos. Por tanto, F toma el valor cero para el conglomerado formado con todos los objetos, puesto que $T = P_G$. En la medida que aumenta el número de conglomerados aumentará el valor de F dependiendo de la variabilidad existente dentro de los conglomerados.

El otro seudo estadístico es el T^2 que no debe confundirse con la T de Hotelling por que parten de matrices diferentes. Para su construcción consideremos en primer lugar la unión de dos conglomerados de tamaño n_k y n_l dada como : $C_m = C_k \cup C_l$ entonces, $W_m \geq W_k + W_l$ de donde $B_{kl} = W_m - W_k - W_l$. El pseudo estadístico es:

$$T^2 = B_{kl}/(W_k + W_l)/(n_k + n_l)$$

Otra medida es el coeficiente de determinación dado por:

$$R^2 = 1 - P_G/T$$

El cual aumenta en la medida que aumenta el número de conglomerados, puesto que la suma de la variación dentro los conglomerado $P_G = \sum_{k=1}^{g} W_k$ disminuye hasta un valor nulo

en el caso extremo que el número de conglomerados es igual al número de objetos. El software SAS tiene además los siguientes indicadores de ajuste:

$$RMSST = \sqrt{\sum_{j=1}^{p} s_{qj}^2 / p}$$

Esta medida se obtiene para cada conglomerado y s_{qj}^2 es la varianza de la variable j en el conglomerado q. En la medida que la suma sea menor, entonces, las agrupaciones son mas homogéneas.

Finalmente, tenemos el criterio (criterio cúbico cluster) *CCC* Este criterio está dado por el producto de:

$$\ln\left((1-E(R^2))/(1-R^2)\right)$$

y

$$\sqrt{(np)/2}/(0{,}001+E(R^2))^{1,2}$$

R^2 es el coeficiente de determinación explicado por el conglomerado. El criterio asume que los conglomerados son obtenidos de una distribución uniforme sobre hipercubos (Khattree R; Naik D (2000)). Bajo este supuesto $E(R^2)$ es la esperanza del coeficiente de determinación R^2.

Para aplicar el criterio se grafica los pares número de conglomerados versus valor del CCC; si $CCC > 3$ nos indica que el número puede ser adecuado

3.- Así como se definen conglomerados de individuos podemos definir conglomerado de variables. La distancia se define de acuerdo a que los datos estén medido es escala ordinal o en escala por lo menos ordinal. Para este último caso se emplea como distancia la correlación de Pearson si los datos están medidos en escala de intervalo o razón o de Sperman para datos medido en escala ordinal.

$$d^2(X;Y) = 2n(1-r_{xy})$$

Donde n es el número de observaciones, r_{xy} es la correlación de Pearson o de Sperman.

En el caso de variables medidas en escala nominal o categórica la distancia es:

$$d(X;Y) = 1-\sqrt{\chi^2/n}$$

Donde χ^2 es la distancia Chi cuadrado obtenida a partir de una tabla de contingencia y n el número de observaciones.

Ejemplo 5.

Supongamos que en una empresa quiere obtener una clasificación de los veintinueve empleados seleccionados al azar de su base de datos, y combinando la información que posee con un cuestionario adicional se recabó información de las siguientes variables: bonos, ingreso, egreso, tarjeta, ahorro y finalmente tiempo de trabajo en la empresa. En total se tiene 7 variables y 29 observaciones. Lo primero que se analizó fue la matriz de correlación encontrando lo siguiente:

CUADRO 24

Correlation Matrix							
		bonos	**ingreso**	**egreso**	**tarjeta**	**ahorro**	**ttrabajo**
bonos	bonos	1.0000	0.7419	0.3679	0.7415	0.8140	0.0372
ingreso	ingreso	0.7419	1.0000	0.5385	1.0000	0.7561	0.3027
egreso	egreso	0.3679	0.5385	1.0000	0.5397	0.3648	0.0659
tarjeta	tarjeta	0.7415	1.0000	0.5397	1.0000	0.7570	0.3048
ahorro	ahorro	0.8140	0.7561	0.3648	0.7570	1.0000	0.0652
ttrabajo	ttrabajo	0.0372	0.3027	0.0659	0.3048	0.0652	1.0000
estudio	estudio	0.7420	1.0000	0.5385	1.0000	0.7562	0.3025

Se puede observar que existe una altísima correlación entre casi todas las variables, encontrándose casos de correlación perfecta. Esto hace pensar que los datos tienen una dimensión menor a nueve. Como los datos están medidos en escala métrica podemos encontrar la verdadera dimensión empleando los componentes principales.

CUADRO 25

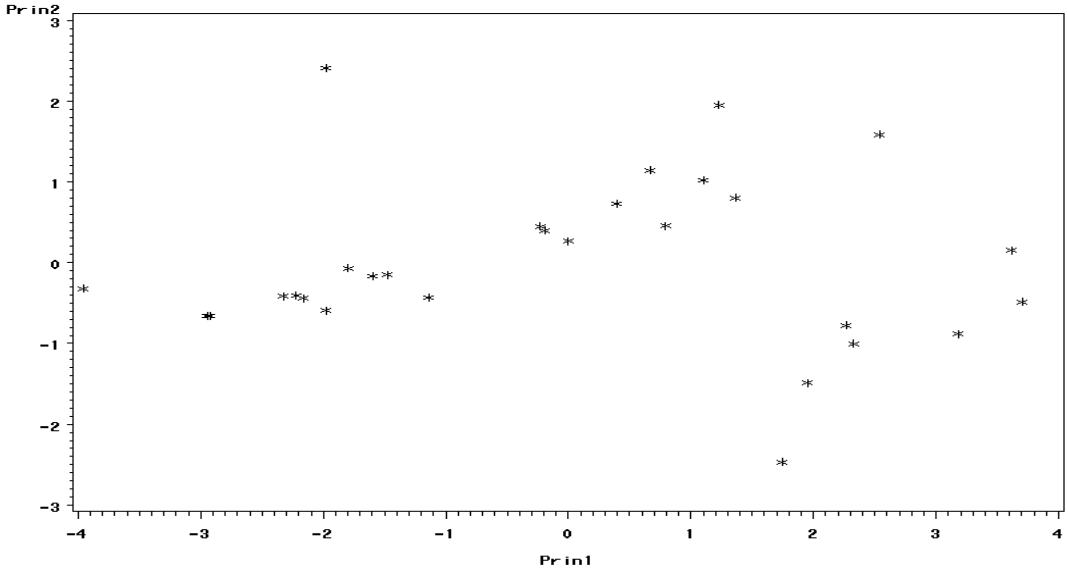
Eigenvalues of the Correlation Matrix				
	Eigenvalue	**Difference**	**Proportion**	**Cumulative**
1	4.70352541	3.65673998	0.6719	0.6719
2	1.04678543	0.30163730	0.1495	0.8215
3	0.74514814		0.1064	0.9279

Al mirar los autovalores de la matriz de correlación observamos que con los tres primeros componentes se explica casi toda la variabilidad de la masa de datos. De esto se concluye que la verdadera dimensión es tres. Por tanto podemos definir los conglomerados con estas tres nuevas variables.

CUADRO 26

Eigenvectors				
		Prin1	Prin2	Prin3
bonos	bonos	0.384838	-.283021	-.294361
ingreso	ingreso	0.450188	0.080880	-.002090
egreso	egreso	0.275409	-.069612	0.899667
tarjeta	tarjeta	0.450348	0.082534	-.001183
ahorro	ahorro	0.389809	-.248988	-.304716
ttrabajo	ttrabajo	0.126242	0.912793	-.105300
estudio	estudio	0.450199	0.080705	-.002104

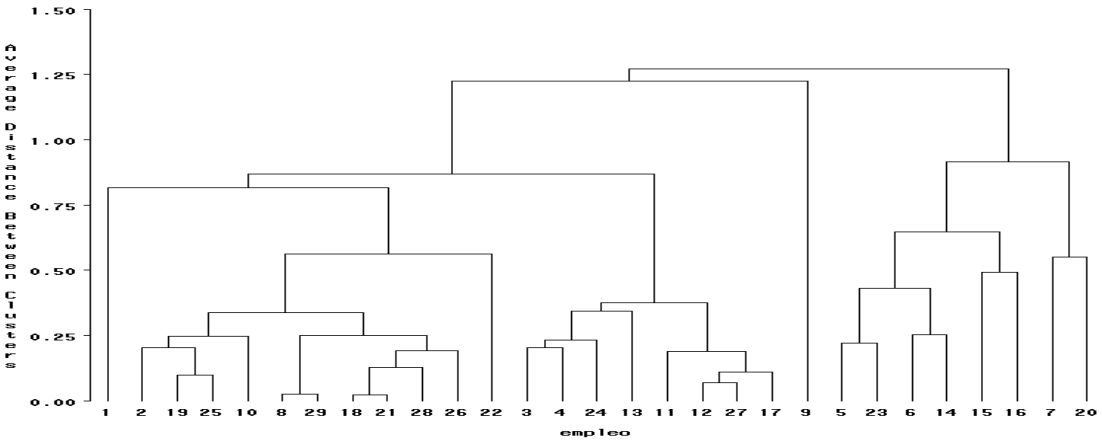
En el cuadro anterior son los tres primeros autovectores con los que se reconstruirían el 92,79% de los datos.
A continuación presentamos el gráfico de las proyecciones de las observaciones en los dos primeros ejes factoriales.



El gráfico sugiere que existen de dos a tres conglomerados. Para definirlos usaremos varios métodos para formar conglomerados y ver hasta que punto coinciden los resultados. La selección del método y del número de conglomerados depende del criterio del investigados.
Ejemplo 6.

El primer criterio que seleccionamos es la distancia promedio dada por: $H(A,B)=\sum_{i}^{na}\sum_{j}^{nb}d(x,y)/n_An_b$ que en software SAS se conoce como average. Este método tiende a agrupar conglomerados que tienen varianzas pequeñas.

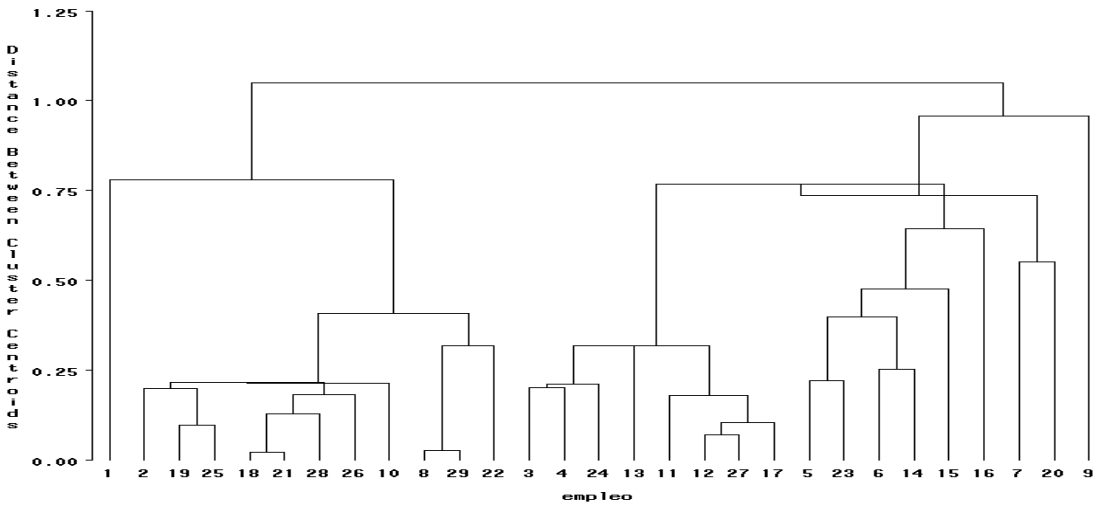
El dendrograma es un gráfico cuya ordenada representa las distancias y la abscisa las observaciones, si se realiza un corte imaginario en la distancia con valor de 1,00 se obtienen tres conglomerados. El primero formado por los empleados enumerados como 1,2,19,25,10,8,29,18,21,28,26,22,3,4,24,13,11,12,27 y 17 es decir, está formado veinte empleados; el segundo conglomerado está solamente el 9 y el tercero por: 5,23,6,14,15,16,7 y 20, es decir, ocho empleados. Si hacemos un corte en 0,5 tendremos más conglomerados y así sucesivamente.



Este gráfico (dendrograma) es útil si se tiene pocas observaciones tal que se puedan ver con claridad las mismas.

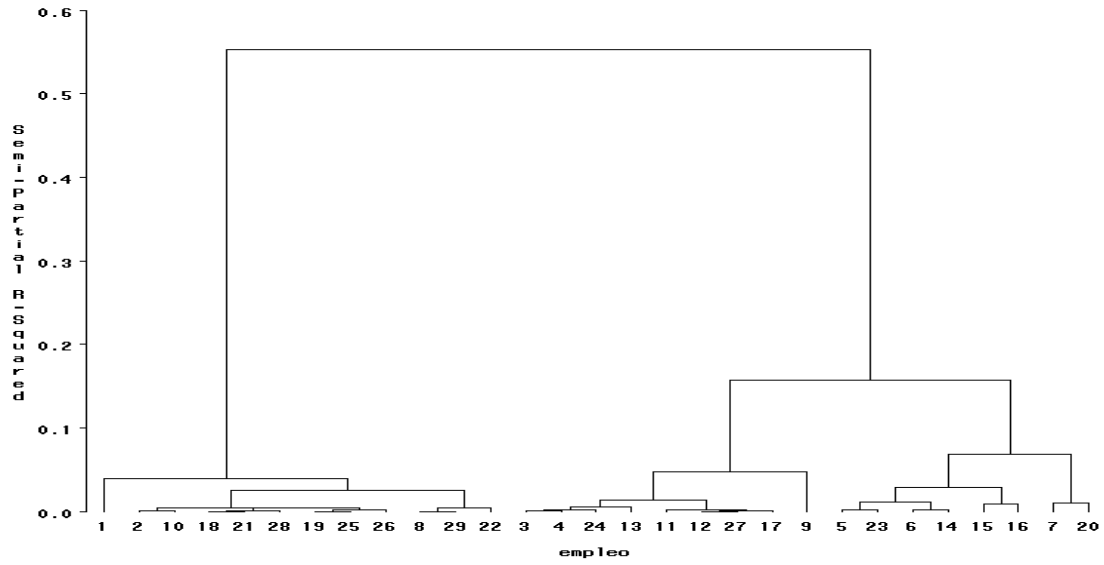
A continuación veremos el desarrollo de otros algoritmos en la formación de conglomerado.

Empleamos como segundo método: el centroide: $H(A;B)=d(\mu_A^*,\mu_b^*)$. En general es conveniente utilizar el mayor número de criterios. Este método se considera uno de los más robustos entre la clase de los métodos jerárquico frente a la presencia de valores atípicos.



El dendrograma obviamente es diferente al anterior por la diferencia de criterios. Si cortamos imaginariamente en el punto 1 como anteriormente tendríamos sólo dos conglomerados. Si el corte imaginario lo hacemos a la altura de 0,75 tendremos seis conglomerados. Como indicamos al inicio de este punto, el investigador decide cuantos conglomerados desea y cual método selecciona. Puede ocurrir que empleando varios métodos casi todos den la misma conformación de los conglomerados, sin embargo puede preferir uno de ellos que le dio una conformación distinta a los otros. Igualmente puede tomar como información las variables originales o usar las coordenadas asociadas a los individuos después de aplicar algún método de reducción de la dimensionalidad tal como los componentes principales obtenidos a partir de la matriz de correlación. Si toma los datos de las variables originales debe tener cuidado de la escala y unidades de medición. En todo caso es preferible trabajar con los datos estandarizados.

Ejemplo 7.
Ahora veremos por último el método de Ward cuyo dendrograma se muestra a continuación:



Se puede observar que el árbol o dendrograma es completamente diferente a los obtenidos anteriormente. Las distancias son menores. Si imaginariamente cortamos en 0,10 tendremos tres conglomerados.
El primero está formado por los clientes 1,2,10,18,21,28,19,25,26,8,29 y 22; el segundo por los clientes 4,24,13,11,12,27,17 y 9; el tercero por los clientes 5,23,6,14,15,16,7 y 20. Este método tiende a reunir conglomerados con un pequeño número de elementos, por otra parte, es muy sensible a valores atípicos.
Ahora analizamos para el caso en donde hemos empleado el método de average las diferentes medidas de bondad de ajuste.

CUADRO 27

Cluster History											
NCL	Clusters Joined		FREQ	SPRSQ	RSQ	ERSQ	CCC	PSF	PST2	Norm RMS Dist	T i e
8	7	20	2	0.0109	.913	.	.	31.5	.	0.5515	
7	CL13	22	11	0.0182	.895	.	.	31.2	6.3	0.5621	
6	CL10	CL9	6	0.0288	.866	.	.	29.8	4.8	0.6458	
5	1	CL7	12	0.0399	.826	.821	0.22	28.5	9.0	0.8162	
4	CL5	CL11	20	0.2098	.616	.774	-4.7	13.4	34.3	0.8691	
3	CL6	CL8	8	0.0684	.548	.703	-3.4	15.8	6.4	0.9151	
2	CL4	9	21	0.0867	.461	.569	-1.7	23.1	5.2	1.2243	
1	CL2	CL3	29	0.4613	.000	.000	0.00	.	23.1	1.2717	

El análisis lo empezaremos de la última fila a la primera, el primer agrupamiento contiene todas las observaciones, eso es como haber hecho el corte del árbol en la máxima distancia, ya de por sí no tiene sentido tomarlo en cuenta, por otra parte, los valores de los diferentes criterios indican que hay que seleccionar otra partición.
Cuando existen demasiados datos se hace poco comprensible el gráfico dendrograma entonces, es preferible fijar el número de conglomerados con la ayuda de la proyección de los datos sobre los dos primeros ejes factoriales si los componentes principales explican satisfactoriamente la variabilidad total de los mismos.
Ahora, en el mismo programa de SAS se le ha pedido que forme cuatro conglomerados indicando los miembros que lo forman. Los resultados son los siguientes:

Una vez decidido el número de conglomerados con sus respectivos miembros es importante caracterizar cada uno de ellos para que esta información permita tomar una decisión lo más cercana a la óptima. La forma de caracterizar a un conglomerado es mediante gráficos, medidas de tendencia y dispersión de cada variable, definir relaciones entre las variables si los conglomerados tienen el suficiente número de elementos y cualquier método descriptivo que se crea apropiado. Lo que no podrá hacerse es aplicar ninguna técnica de inferencia estadística por la falta de aleatoriedad en la asignación de los elementos en los conglomerados.

CUADRO 28

Obs	CLUSTER	empleo
1	1	18
2	1	21
3	1	8
4	1	29
5	1	12
6	1	27
7	1	19
8	1	25
9	1	17
10	1	28
11	1	11
12	1	26
13	1	3
14	1	4
15	1	2
16	1	24
17	1	10
18	1	13
19	1	22
20	1	1
21	2	5

Obs	CLUSTER	empleo
22	2	23
23	2	6
24	2	14
25	2	15
26	2	16
27	3	7
28	3	20
29	4	9

En este caso se le ha fijado un número de conglomerados igual a cuatro.

V.-ANÁLISIS DISCRIMINANTE.

En diferentes situaciones se tiene un conjunto de datos conformado por varias observaciones o individuos, cada una de ella con p mediciones clasificadas de acuerdo a un determinado criterio. Esta clasificación puede deberse a que las observaciones responden a un muestreo estratificado o, son el resultado de aplicar algún algoritmo de conglomerado o simplemente, la clasificación se hizo siguiendo algún criterio especial. Por ejemplo, una institución financiera puede tener sus clientes divididos en tres categorías: clientes con muy bajo riesgo, es decir, alta solvencia, con riesgo moderado o solvencia aceptable y con alto riesgo o insolvencia. En el criterio para hacer esta clasificación posiblemente emplearon varios índices financieros y otras variables. Entonces el problema que se presenta es dado que se tienen varios grupos previamente definidos ¿A cuál grupo debe clasificarse los nuevos clientes que solicitan un crédito?

A este tipo de pregunta responde el análisis discriminante. Hay varios métodos de análisis discriminantes que parten de diferentes supuestos y criterios, pero todos buscan minimizar el error de una mala clasificación. En la práctica se pueden emplear diferentes métodos tal como ocurre en el problema de pronóstico y, seleccionar aquel que tenga menor probabilidad de cometer el error de mala clasificación. Los métodos se dividen en los basados en probabilidades, entre ellos están el de máxima verosimilitud, bayesiano y, los que parten del concepto de distancia, entre ellos están el análisis discriminante lineal de Fisher y el discriminante cuadrático. Además se pueden clasificar de acuerdo a los supuestos sobre las distribuciones de las poblaciones como paramétricos y no paramétricos, y también, de acuerdo a la naturaleza de las variables en: análisis discriminante para variables continuas y análisis discriminante para variables discretas. En este punto solo veremos dos métodos que están al nivel de este texto introductorio y, que se pueden asociar al modelo lineal estándar visto en el capítulo VI. En algunos casos se hacen supuestos sobre la distribución de las poblaciones de donde provienen los grupos, generalmente estos

supuestos se refieren a poblaciones distribuidas normalmente, en otros no y, en este caso los métodos que se emplean son los métodos no paramétricos. El primer método que veremos es el de Fischer y se asume que los datos provienen de poblaciones con distribuciones normales multivariantes con iguales matrices de varianzas covarianzas. El segundo método es un modelo probabilístico y se conoce como modelo logit o logístico y se asume como distribución la binomial para el caso de dos grupos o, polinomial si hay más de dos grupos.

Partimos de la idea que tenemos dos grupos o poblaciones G_1 y G_2 y consideremos una observación x_0, entonces, $P(G_1)$ y $P(G_1)$ son las probabilidades a priori de que la observación x_0 pertenezca al grupo 1 o al grupo 2. Denotemos por $P(G_i / x_0)$ $i = 1,2$ las probabilidades a posteriori de que dada la observación x_0 sea de la población G_i; $i = 1,2$. Para determinar estas probabilidades empleamos el teorema de Bayes:

$$P(G_i / x_0) = P(x_0 / G_i)P(G_i)/(P(x_0 / G_i)(P(G_i) + P(x_0 / G_j)P(G_j)) \quad i \neq j \; i = 1,2 \; j = 1,2$$

Ahora, $P(G_i / x_0) \approx f_i(x_0;\Theta)\Delta x_0$ donde $\approx f_i(x_0;\Theta)$ es la función de densidad asociada a la población i ésima, Θ es el conjunto de parámetros. Luego la expresión anterior es:

$$P(G_i / x_0) = f_i(x_0;\Theta)\Delta x_0 P(G_i)/(f_i(x_0;\Theta)\Delta x_0 P(G_i) + f_j(x_0;\Theta)\Delta x_0 P(G_j))$$

$$P(G_i / x_0) = f_i(x_0;\Theta)P(G_i)(f_i(x_0;\Theta)P(G_i) + f_j(x_0;\Theta)P(G_j))$$

La observación x_0 se asigna al grupo G_i si $P(G_i / x_0) > P(G_j / x_0)$ P; para $i \neq j$; esto equivale a:

$f_i(x_0;\Theta)P(G_i) > f_j(x_0;\Theta)P(G_j)$ para $i \neq j$.

Una mala clasificación generalmente acarrea un costo, por tanto, es importante considerar el costo del error de clasificación. Llamaremos $C(i/j)$ los costos de clasificar la observación **x** en el grupo i cuando pertenece al grupo j. El costo promedio del error de clasificación en el caso de dos grupos es:

$$P(G_1)C(2/1)P(G_2 / G_1) + P(G_2)C(1/2)P(G_1 / G_2)$$

$P(G_i / G_j)$ es la probabilidad de clasificar la observación **x** en el grupo i cuando pertenece al grupo j. Este costo promedio permite seleccionar aquel método que dé el mínimo costo promedio.

Va.-Análisis lineal de Fisher.

Partiremos del caso más sencillo en donde sólo tenemos dos grupos: G_1 y G_2 con n_1 y n_1 observaciones p dimensionales. Asumimos que ambos grupos provienen de dos poblaciones que tienen la misma variabilidad medida por la matriz de varianza covarianza $(\Sigma_1 = \Sigma_2)$ y medias μ_1 y μ_2. Los estimaremos son la matriz de varianza covarianza de las observaciones dada por: S y, las medias μ_1^* y μ_2^* . La idea es encontrar una proyección

lineal que separe lo máximo posible los dos grupos. Entonces, tenemos el vector de direcciones α y el escalar z dado por:

$z = \alpha x$

Donde **x** es el vector de las p variables. En la proyección las medias de los dos grupos son: $\alpha\mu_1^*$ y $\alpha\mu_2^*$ y la varianza: $\alpha^T S\alpha$. El problema se traduce en encontrar el vector de direcciones **α** que maximice la distancia entre los dos grupos dada por:

$d(G_1, G_2) = (\alpha\mu_1^* - \alpha\mu_2^*)^2 / \alpha^T S\alpha$

Al resolver el problema de optimización encontramos el valor del vector que maximiza la distancia está dado por:

$\alpha^* = S^{-1}(\mu_1^* - \mu_2^*)$

La regla de decisión es: dado una nueva observación x_0, entonces se le asigna al grupo G_1 si $\alpha^* x_0 \geq h$ y al grupo G_2 en caso contrario. El valor de h es la distancia promedio de las dos medias en la proyección óptima, esto es: $h = (\alpha^*\mu_1^* - \alpha^*\mu_2^*)/2$. El análisis discriminante lineal se puede abordar como un caso particular del modelo de regresión estándar en donde la variable a explicar es una variable categórica, digamos que $y = 1$ cuando pertenece al primer grupo; $y = 0$ cuando pertenece al segundo grupo. Entonces, hay que encontrar el ajuste mínimo cuadrático del modelo lineal dado por:

$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_k x_k + \varepsilon = X\beta + \varepsilon$

El modelo ajustado es:

$Y^* = X\beta^*$

Para un valor $\mathbf{x_0}$ particular la regla de decisión es:

Si $Y^* = x_0\beta^* > 0$; x_0 entonces se asigna al grupo uno, en caso contrario al grupo dos. Hay varios problemas que debemos tener presente. El primero es el problema de la mala clasificación, esto es, debemos estimar la probabilidad $P(G_i / G_j)$, que representa la probabilidad de clasificar la observación x_0 como perteneciente a G_i dado que pertenece realmente a G_j. El otro problema es contrastar la hipótesis de igualdad de las matrices de varianza covarianza. $H_0 : \Sigma_1 = \Sigma_2 = \Sigma$ y finalmente seleccionar el mejor número de variables explicativas en el modelo. Hay varios métodos para estimar la probabilidad de mala clasificación, una de ellas es emplear las observaciones que se posee en ambos grupos y aplicar el método sobre estas observaciones para determinar la proporción de observaciones mal clasificadas tal como se presenta en el siguiente cuadro:

TABLA1

TABLA 3

	Perteneciente a G_1	Perteneciente G_2
Clasificada en G_1	n_{11} (frecuencia de bien clasificadas)	n_{12}(frecuencia de mal clasificadas)
Clasificada en G_2	n_{21}(frecuencia de mal clasificadas)	n_{22} (frecuencia de bien clasificadas)

Por tanto, el estimador de $P(G_i / G_j)$ es $P^*(G_i / G_j) = n_{ij} / n$. El segundo problema lo explicaremos con un ejemplo corrido en SAS.

Ejemplo 8.

Tenemos una muestra de 57 clientes de una entidad bancaria de los cuales se tiene el ingreso mensual y los pasivos, divididos en dos grupos. El primer grupo codificado como cero son clientes solventes y el segundo grupo codificado con uno conforman los clientes morosos, se seleccionó una muestra aleatoria de cada uno con tamaños $n_1 = 29$ y $n_2 = 28$. Se desea obtener un mecanismo que permita clasificar los futuros clientes en uno de los dos grupos con la finalidad de considerarlos como futuros clientes de los préstamos que facilita la entidad. Para procesar la información se empleó SAS.

Los resultados de la corrida del software SAS se muestran a continuación.

El cuadro siguiente muestra el número total de observaciones, el número de variables y el número de grupos o clases. Además, el número de elementos por grupos, sus pesos, la proporción y la probabilidad a priori de que un elemento sea miembro de uno de los grupos: $P(G_i)$. Finalmente se da el rango de la matriz de varianza covarianza, esta matriz es la estimación de la matriz poblacional: $\Sigma_1 = \Sigma_2 = \Sigma$ **(**recordemos que se asume que las dos poblaciones tienen igual matriz de varianza covarianza). Cuando el rango de la matriz es menor que el número de variables el software SAS emplea un algoritmo basado en la seudo inversa de Moore Penrose.

CUADRO 29

´Analisis discriminante´

Procedimiento DISCRIM

Observacione	57	**Total DF**	56
Variables	2	**Clases Within DF**	55
Clases	2	**Clases Between DF**	1

Información del nivel de la clase					
grupo	**Nombre de variable**	**Frecuencia**	**Peso**	**Proporción**	**Probabilidad anterior**
0	_0	29	29.0000	0.508772	0.500000
1	_1	28	28.0000	0.491228	0.500000

Matriz de Información de covarianza ponderada	
Rango de la matriz de covarianza	**Registro log natural de la Determinante de la matriz de covarianza**
2	19.48246

En la salida siguiente se muestra la matriz de las distancias entre las dos medias de los grupos. Ella es una matriz simétrica, cuya diagonal principal está formada por ceros. Además, se presenta la los vectores $\alpha_1^* = S^{-1}\mu_1^*$ y $\alpha_2^* = S^{-1}\mu_2^*$, asociados a las dos funciones discriminantes. La diferencia entre estos dos vectores da el valor de α^* y : $h = (\alpha^*\mu_1^* - \alpha^*\mu_2^*)/2$. Con esta información puede obtenerse la función lineal discriminante:

$$\alpha^* x = S^{-1}(\mu_1^* - \mu_2^*)x = (-2{,}5373 - 8{,}37818) + (0{,}05552 + 0{,}10212)x_1 + (-0{,}0553 - 0{,}102)x_2$$

´Analisis discriminante´ 20:34 Thursday, July 6, 2005 12
The DISCRIM Procedure

Pairwise Generalized Squared Distances Between Groups

$$D^2(i|j) = (\bar{X}_i - \bar{X}_j)' \, COV^{-1} \, (\bar{X}_i - \bar{X}_j)$$

Generalized Squared Distance to generalized Squared Distance to grupo

Procedimiento DISCRIM

Distancia cuadrada para grupo		
De grupo	**0**	**1**
0	0	3.39745
1	3.39745	0

Linear Discriminant Function

$$\text{Constant} = -.5\ \bar{X}_j' \text{COV}^{-1} \bar{X}_j \qquad \text{Coefficient Vector} = \text{COV}^{-1} \bar{X}_j$$

Procedimiento DISCRIM

Función discriminante lineal para grupo			
Variable	**Etiqueta**	**0**	**1**
Constante		-2.53730	-8.37818
ingreso	ingreso	0.05552	0.10212
pasivos	pasivos	-0.05530	-0.10201

La siguiente salida muestra los rangos de las matrices de covarianza de los dos grupos y los logaritmos de sus determinantes. La diferencia entre estos logaritmo se toma como un indicador de la diferencia entre las matrices de varianzas de las poblaciones de donde provienen los grupos.

Within Covariance Matrix Information

grupo	Covariance Matrix Rank	Natural Log of the Determinant of the Covariance Matrix
0	2	17.50155
1	2	20.44754
Pooled	2	19.48246

El siguiente paso es hacer el contraste de igualdad de las matrices de covarianzas:
$H_0 : \Sigma_1 = \Sigma_2 = \Sigma$ versus $H_1 : \Sigma_1 \neq \Sigma_2$

The DISCRIM Procedure
Test of Homogeneity of Within Covariance Matrices

Notation: K = Number of Groups

P = Number of Variables

N = Total Number of Observations - Number of Groups

N(i) = Number of Observations in the i'th Group - 1

$$V = \frac{\prod |\text{Within SS Matrix(i)}|^{N(i)/2}}{|\text{Pooled SS Matrix}|^{N/2}}$$

$$RHO = 1.0 - \left| \sum_{N(i)} \frac{1}{N(i)} - \frac{1}{N} \right| \frac{2P^2 + 3P - 1}{6(P+1)(K-1)}$$

$$DF = .5(K-1)P(P+1)$$

Under the null hypothesis:

$$-2\,RHO \ln \left| \frac{\prod_{N(i)} N(i)^{PN(i)/2}}{N^{PN/2}} \right|$$

is distributed approximately as Chi-Square(DF).

Chi-Square	DF	Pr > ChiSq
28.249627	3	<.0001

Since the Chi-Square value is significant at the 0.1 level, the within covariance matrices will be used in the discriminant function.
Reference: Morrison, D.F. (1976) Multivariate Statistical Methods p252.

De acuerdo al resultado del test la hipótesis nula se rechaza a un nivel de 0,05 puesto que el valor de la probabilidad que el estadístico tome un valor igual o mayor a 28,2496 es menor que 0,0001, Generalmente, para tamaños de muestras grandes con el estadístico propuesto tiende a rechazarse la hipótesis nula. En este caso se han propuesto varias soluciones. Una solución es definir: $d(x;G_i) = (x-\mu_i^*)^T S_i^{-1}(x-\mu_i^*)/2 + \ln|S_i| - \ln|P(G_i)C(j/i)|$; donde $P(G_i)$ es la probabilidad a priori de pertenecer al grupo i y $C(j/i)$ es el costo de clasificarlo en el grupo j cuando pertenece al grupo i. La regla de decisión es asignar **x** al G_1 si $d(x;G_1) < d(x;G_2)$.

El siguiente resultado se refiere a la distancia mencionada en el punto anterior y, las probabilidades a posteriori que se emplean en la estimación de la probabilidad en el error de clasificación.

´Analisis discriminante´ 20:34 Thursday, July 5, 2005 13

The DISCRIM Procedure
Classification Summary for Calibration Data: SASUSER.DISCRIMINANTE
Resubstitution Summary using Linear Discriminant Function

Generalized Squared Distance Function

$$D_j^2(X) = (X-\bar{X}_j)' \, COV^{-1} \, (X-\bar{X}_j)$$

Posterior Probability of Membership in Each grupo

$$Pr(j|X) = \exp(-.5\, D_j^2(X)) \,/\, SUM_k \exp(-.5\, D_k^2(X))$$

El cuadro siguiente da las frecuencias y los porcentajes como resultados de aplicar la regla de clasificación. Por ejemplo, de 29 observaciones que pertenecen al primer grupo 28 fueron clasificada correctamente y representa el 96,55% y una sola en el otro grupo, es decir erróneamente clasificada y representa el 3,46%.

Número de observaciones y porcentaje clasificado en grupo			
De grupo	**0**	**1**	**Total**
0	28 96.55	1 3.45	29 100.00
1	6 21.43	22 78.57	28 100.00
Total	34 59.65	23 40.35	57 100.00
Prior	0.5	0.5	

El último cuadro indica la estimación del error de clasificación, el total se obtiene multiplicando el error por grupo por la probabilidad a priori: 0,5x0,0345+0,5x0,2143=0,1244

Estimaciones de % de error para grupo			
	0	**1**	**Total**
Tasa	0.0345	0.2143	0.1244
Prior	0.5000	0.5000	

Vb.-Discriminante logit.

Este método está relacionado con el modelo de regresión lineal estándar y es aplicable para el caso de varios grupos sin embargo, sólo veremos el caso de dos grupos por tanto, asumimos que la pertenencia a uno de los dos grupos es una variable aleatoria con distribución de Bernoulli. Consideremos entonces la variable aleatoria Y que toma los valores 0 y 1, con probabilidad p y q respectivamente. Estas probabilidades responden a un

número de factores observables que denotamos por el vector **x.** De este supuesto, se deriva la idea que p se puede expresar como una función logística:

$p = \exp x\beta/(1+\exp x\beta)$, como $p+q=1$ entonces, $q=1/(1+\exp x\beta)$

Por tanto, podemos escribir la probabilidad que pertenezca al primer grupo como:

$P(y=1)=p=\exp x\beta/(1+\exp x\beta)$ y que pertenezca al segundo grupo es $P(y=0)=q=1/(1+\exp x\beta)$. Para simplificar podemos plantearnos el establecer la relación que existe entre la pertenencia a un grupo o a otro y, esto se logra tomando logaritmo de la división de las dos probabilidades:

$\ln\big(P(y=1)/P(y=0)\big)=x\beta$

De la misma forma que en el modelo de regresión estándar habrá que estimar: el valor ajustado, esto es; el valor Y^* que rara veces dará los valores 0 o 1, pero sí valores comprendidos entre esos dos valores; para ello habrá que estimar previamente el parámetro β y hacer los contrastes correspondientes**.** La estimación del parámetro β se puede realizar mediante el método de máxima verosimilitud. Los software de computación emplean diferentes algoritmos que no abordaremos. El contraste de hipótesis sobre los elementos del vector β se puede tratar de diferentes maneras. La primera hipótesis sobre el valor de un parámetro particular se puede tratar desde dos puntos de vista. La primera es emplear a igual que el modelo lineal estándar la prueba t de Student de la forma siguiente:

$H_0:\beta_j=0$ contra la alternativa que β_j tiene un valor diferente de cero. Esta hipótesis lo que establece es que el factor o variable no tiene ningún efecto en la probabilidad P. Por tanto, el estadístico t, tiene la forma:

$t^*=\beta_j^*/\sqrt{\text{var}^*\beta_j^*}$

Donde β_j^* es el estimador de β_j y $\sqrt{\text{var}^*\beta_j^*}$ es el estimador de la desviación estándar del estimador. Bajo la hipótesis nula, este estadístico tiene una ley de t de Stundent con n-k grados de libertad donde k es el número de parámetro. Rechazamos la hipótesis nula si $P\big(t>|t^*|\big)<\alpha/2$.

Otro test para realizar el mismo contraste de hipótesis es el estadístico W de Wald dado como:

$W^*=(\beta_j^*/\sqrt{Var^*(\beta_j^*)})^2$

Bajo la hipótesis nula este estadístico se distribuye aproximadamente como una Chi cuadrado con un grado de libertad. Rechazamos la hipótesis nula si para un valor estimado de $W:W^*$ se tiene $P\big(\chi^2>W^*\big)<\alpha/2$.

Para hacer el contraste $H_0:\beta_1=\beta_2=...\beta_p=0$, se emplea el test basado en la razón de máxima verosimilitud, en donde $L(x;\beta_0)$ es la función de máxima verosimilitud tomando solamente el parámetro de intesección β_0 y, $L(x;\beta)$ la función de máxima verosimilitud

construida tomando en cuenta todos los parámetros asociados a las variables explicativas. Esto es:

$$L = 2\ln\left(L(x;\beta_0)\right)/\left(L(x;\beta)\right) = 2(\ln L(x;\beta_0) - \ln L(x;\beta))$$

Este estadístico sigue una distribución χ^2 con p grados de libertad. Rechazamos la hipótesis nula si para un valor estimado L^* la probabilidad $P\left(\chi^2 > L^*\right) < \alpha$.

Ejemplo 9.

Apliquemos el modelo a los mismos datos del ejemplo anterior.

La primera información se refiere al número de grupos (2) y de observaciones (57). Cada grupo tiene 29 y 28 clientes respectivamente.

Procedimiento LOGISTIC

Información del modelo		
Conjunto de datos	SASUSER.DISCRIMINANTE	
Variable de respuesta	grupo	grupo
Número de niveles de respuesta	2	
Modelo	logit binario	
Técnica de optimización	Puntuación de Fisher	

Perfil de respuesta		
Valor ordenado	**grupo**	**Frecuencia total**
1	0	29
2	1	28

La probabilidad modelada es grupo=0.

La siguiente impresión corresponde a los valores del AIC, SC que son valores que se emplean cuando se están comparando varios modelos para el mismo conjunto de datos, se seleccionará aquél que presente los valores menores y, $2\ln L$, se refiere a los logaritmos de las funciones de verosimilitud $2\ln L(x;\beta_0)$=79,001 y $2\ln L(x;\beta)$=27,360. La diferencia da L=79,001-27,360=51,6411. La probabilidad de obtener este valor bajo la hipótesis nula:

$H_0 : \beta_1 = \beta_2 = 0$ es $P\left(\chi^2 > L^*\right) < 0{,}001$, por tanto para un nivel de α =0,05 se rechaza la hipótesis nula.

Estadístico de ajuste del modelo		
Criterio	**Sólo términos independientes**	**Términos independientes y Variables adicionales**
AIC	81.001	33.360
SC	83.044	39.489
-2 LOG L	79.001	27.360

Probar hipótesis nula global: BETA=0			
Test	**Chi-cuadrado**	**DF**	**Pr > ChiSq**
Ratio de verosim	51.6411	2	<.0001
Puntuación	26.6805	2	<.0001
Wald	12.4050	2	0.0020

Con el cuadro siguiente podemos obtener el modelo ajustado empleando el método de máxima verosimilitud. El modelo es: $\ln(P(y=1)/P(y=0)) = 15{,}0959 - 0{,}1255x_1 + 0{,}1239x_2$
Luego tomando antilogaritmo y recordando que $P(y=1) = \exp x\beta/(1+\exp x\beta)$ y $P(y=0) = 1/(1+\exp x\beta)$ obtenemos:
$P(y=1) = \exp(15{,}095 - 0{,}1255x_1 + 0{,}1239x_2)/(1+\exp(15{,}095 - 0{,}1255x_1 + 0{,}1239x_2))$ Si para un nuevo individuo x_0 $P(y=1) > 0{,}5$ entonces se clasifica en el grupo y= 1, en caso contrario; en el grupo y=0.

Análisis del estimador de máxima verosimilitud					
Parámetro	**DF**	**Estimador**	**Error estándar**	**Chi-cuadrado de Wald**	**Pr > ChiSq**
Intercept	1	15.0959	4.4236	11.6456	0.0006
ingreso	1	-0.1255	0.0359	12.2022	0.0005
pasivos	1	0.1239	0.0353	12.3160	0.0004

Las restantes columnas de este cuadro permiten realizar para cada parámetro el contraste: $H_0 : \beta_j = 0$, empleando el estadístico W de Wald. En efecto, para el caso del intercepto tenemos la siguiente relación: $W^* = (\beta_j^* / \sqrt{Var^*(\beta_j^*)})^2$, $W^* = (15{,}0959/4{,}4236)^2 = 11{,}6456$. La probabilidad bajo la hipótesis nula es $P(\chi^2 > W^*) = 0{,}0006 < \alpha = 0{,}05$. Por tanto, se rechaza la hipótesis nula. Igual ocurre con las demás hipótesis sobre los otros parámetros.

BIBLIOGRAFÍA.

Bastin, Ch; Benzécri et al (1980)
Pratique de L´Analyse des dones. Vol 2.
Dunod Paris. Francia

Benzécri et Collaborateurs (1976)
L´ Analyse des Donnees Vol 2 L´Analyse des Correspondance.
2da Êdition. Dunod Paris. Francia.

Cea D´Ancona M.A (2004)
Análisis Multivariante.-Teoría y Práctica en la Investigación Social.
Editoral Síntesis. Madrid. España

Hair J.F; Anderson R. E; Tatham R.L; Black W.C (2000)
Análisis Multivariante
Prentice Hall. 5ª Edición Madrid. España.

Hand, D.J (1981)
Discrimination and Classification.
John Wiley and Sons. Chichester. New York. USA

Johnson D.F (2000)
Métodos Multivariados Aplicados al Analisis de Datos
Internacional Thonson Editores. México-México.

Khattree R; Naik D (2000)
Multivariate Data Reduction and Discrimination
SAS institute. Inc. California. USA

Lebart; L. Morineau; A. Tabard: N. (1977)
Techniques de la Description Statistique.

Dunod-Paris. Francia.

Peña D (2002)
Análisis de Datos Multivariante
McGraw Hill-México DF. México.

Perez C (2004)
Técnicas de Análisis Multivariante de Datos.
Pearson-Prentice Hall.-Madrid-España

Stokes M.E; Davis, S.C; Koch, G.G (2000)
Categorical Data Analysis using the SAS System.
SAS institute. Inc. California. USA

Vega de la, Silvia (1995) @
Modelos Probabilística de Elección.
***Cuadernos Metodológicos Nº 12. CIS. Madrid. España*.**

SITIOS WEB
Introducción al Análisis Multivariante(2002)
jegn@estadistico.com
Creación de un Proyecto de Data Mining-Fases (2004)
jegn@estadistico.com
Salvador Figueras, M (2000): "Análisis Discriminante", [en línea] *5campus.com, Estadística* <http://www.5campus.com/leccion/discri>