8.- CORRELACIÓN MULTIPLE Y CORRELACIÓN CANÓNICA

8.1.- Correlación múltiple

Como se ha visto anteriormente, el coeficiente de correlación simple está basado en la suposición de la aproximación a la distribución *normal bivariante*. Si se tiene más de dos variables, el modelo básico para la correlación múltiple, sería una ampliación de esta distribución, denominada distribución *normal multivariante*.

Si hay tres variables, habrá tres correlaciones simples entre ellas, ρ_{12}, ρ_{13} y ρ_{23}. Estos coeficientes miden la relación lineal que existen entre estas variables, dos a dos, sin tener en cuenta la posible influencia de la tercera.

La correlación parcial se define como la correlación entre dos variables si las demás variable no varían, es decir, el valor de las demás variables son fijos. Por ejemplo, el coeficiente de correlación parcial $\rho_{12.3}$, es la correlación entre la variable 1 y 2 siendo constante el valor de la variable 3; o el coeficiente de correlación parcial $\rho_{23.1}$ es la correlación entre la variable 2 y 3 siendo constante el valor de la variable 1.

El mantener constante una variable puede hacerse experimentalmente o estadísticamente, debiendo dar en ambos casos resultados equivalentes. Para ver claro el porqué se necesita hallar una correlación haciendo constante el valor de otra u otras variables supóngase que se está interesado en conocer la correlación entre la longitud del brazo y de la pierna cuando el tamaño total del organismo permanece constante. Está claro que la longitud del brazo y de la pierna estarán altamente correlacionados debido al tamaño general; así, un individuo alto tendrá brazos y piernas largos, mientras que un individuo bajo tendrá extremidades cortas. Sin embargo, si este estudio se seleccionan individuos del mismo tamaño se puede esperar que exista alguna correlación residual entre la longitud del brazo y de la pierna. Esto es muy probable en vertebrados, debido a que ambas extremidades están determinadas embriológicamente con mecanismos homólogos responsables de la diferenciación y determinación. Por tanto existirá alguna correlación entre éstas dos longitudes, incluso en ausencia de una causa común como es el tamaño del individuo. Si una correlación significativa entre dos variables se convierte en correlación parcial no significativa cuando una tercera variable permanece constante, esto sugiere, aunque no prueba, que la variable que permanece constante es la causa común de la correlación de las otras dos.

Correlación parcial / correlación de lo residuos.-

La correlación parcial $r_{12.3}$, sería la correlación lineal entre la variable 1 y 2 dejando como constante la variable 3. Esto quiere decir que hay que medir la correlación entre la variable 1 y 2 que no sea un reflejo de sus relaciones con la variable 3. Por tanto, se puede obtener una estima muestral $r_{12.3}$ calculando la desviación o residuo e_{13}, de la regresión de la variable 1 sobre la variable 3, y la desviación o residuo e_{23}, de la regresión de la variable 2 sobre la variable 3. Y $r_{12.3}$ es
el coeficiente de correlación simple entre e_{13} y e_{23}.

Ejemplo.-

Se tiene el rendimiento (***prod***) de una línea de trigo observado en 11 años sucesivos y los datos meteorológicos correspondientes: precipitación en Noviembre y Diciembre (***prenov***), temperatura media en Julio (***tmjul***), precipitación en Julio (***prejul***) y radiación solar en Julio (***radjul***).

¿Cuál sería la correlación entre la producción y la precipitación en Noviembre si la temperatura media de Julio hubiera sido la misma todos los años?

Archivo del programa *SAS* (corpar1.sas).-

```
options ls=75 ps=60;
data corpar-1;
infile '\corpar.dat';
input prod prenov tmjul prejul radjul;
proc reg noprint;
model prod = tmjul;
output out=residuos R=rprod;
run;
proc reg noprint;
model prenov = tmjul;
output out=residuos R=rprenov;
run;
proc corr;
var rprod rprenov ;
run;
```

La correlación de los dos residuos es 0.3218, que como se verá en el siguiente ejemplo, es la correlación parcial de la producción con la precipitación en Noviembre si la temperatura media de Julio se hubiera mantenido como constante.

Cálculo de la correlación parcial.-

Puede demostrarse que $r_{12.3}$ satisface la siguiente fórmula

$$r_{12.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1 - r_{13}^2)(1 - r_{23}^2)}}$$

El error típico de esta estima es

$$S_{r_{12.3}} = \sqrt{\frac{1 - r_{12.3}^2}{n - 3}}$$

Por lo que podemos probar H_0: $r_{12.3} = 0$ por medio de la t

$$t = \frac{r_{12.3}\ \sqrt{n - 3}}{\sqrt{1 - r_{12.3}^2}}$$

Que se contrasta con la $t_{n-3;\ \alpha/2)}$.

De la misma manera se puede hallar la regresión entre la variable 1 y la variable 3 dejando constante la variable 2; o la correlación entre la variable 2 y la variable 3 dejando constante la variable 1.

$$r_{13.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1 - r_{12}^2)(1 - r_{23}^2)}};\quad S_{r_{13.2}} = \sqrt{\frac{1 - r_{13.2}^2}{n - 3}};\quad t = \frac{r_{13.2}\sqrt{n - 3}}{\sqrt{1 - r_{13.2}^2}}$$

$$r_{23.1} = \frac{r_{23} - r_{12}r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}; \quad S_{r_{23.1}} = \sqrt{\frac{1-r_{23.1}^2}{n-3}}; \quad t = \frac{r_{23.1}\sqrt{n-3}}{\sqrt{1-r_{23.1}^2}}$$

El coeficiente de correlación simple entre dos variables se le puede denominar *coeficiente de orden cero* y se simboliza por medio de una *r* con dos subíndices que hacen referencia a las variables de las que se está hallando la correlación. Los coeficientes de correlación parcial que se refieren a la correlación de dos variables dejando fija una tercera se denominan *coeficientes de primer orden* y se representan con la *r* con tres subíndices, los dos primeros separados del tercero por un punto, es decir, los dos primeros hacen referencia a las variables para las que se ha hallado la correlación y el tercero la variable que se ha hecho constante. De forma análoga se puede obtener coeficientes de *segundo*, *tercer*, *cuarto* o *n-ésimo* orden, dependiendo del número de variables que se mantienen constantes mientras se mide la correlación entre dos variables.

Los coeficientes de correlación parcial de un orden determinado pueden deducirse partiendo de los de orden inmediatamente inferior. Así, se puede obtener un coeficiente de primer orden aplicando la relación, ya conocida, de los de orden cero

$$r_{12.3} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

Para un coeficiente de segundo orden la relación es con coeficientes de primer orden, esta es

$$r_{12.34} = \frac{r_{12.3} - r_{14.3}\ r_{24.3}}{\sqrt{(1-r_{14.3}^2)(1-r_{24.3}^2)}}$$

Y una correlación parcial de orden *k-ésimo*

$$r_{12.34...k} = \frac{r_{12.34..(k-1)} - r_{1k.34...(k-1)}\ r_{2k.34...(k-1)}}{\sqrt{(1-r_{1k.34...(k-1)}^2)(1-r_{2k.34...(k-1)}^2)}}$$

La prueba t para contrastar si esta *r* es diferente de cero, es

$$t = \frac{r_{12.34...k}\sqrt{n-k}}{\sqrt{1-r_{12.34..k}^2}}$$

que se contrasta con la $t_{(n-k;\ \alpha/2)}$.

Por tanto, es posible, partiendo de los coeficientes de correlación de orden cero, calcular sucesivamente todos los coeficientes de orden más elevado.

Ejemplo.-

Siguiendo con el ejemplo del trigo calcúlese, a modo de ejemplo: (a) las correlaciones simples entre todas las variables; (b) correlación parcial entre la producción y la precipitación en Noviembre si la temperatura media de Julio fuera constante; (c) la correlación parcial entre la producción y la precipitación en Noviembre si la temperatura media de Julio y la precipitación de Julio fueran constantes; (d) la correlación parcial entre la producción y la precipitación en Noviembre si la temperatura media de Julio, la precipitación

de Julio y la radiación solar de Julio fueran constantes.

Archivo del programa *SAS* (corpar2.sas).-

```
options ls=75 ps=60;
data corpar-2;
infile '\corpar.dat';
input prod prenov tmjul prejul radjul;
proc corr;
run;
proc corr;
var prod prenov;
partial tmjul;
run;
proc corr;
var prod prenov;
partial tmjul preJul;
run;
proc corr;
var prod prenov;
partial tmjul preJul radJul;
run;
```

En el archivo de salida (**corpar-2.lst**) se observa que en la salida del primer procedimiento la correlación entre la producción y la precipitación en Noviembre es negativa, si bien es no significativa, mientras que en la salida del segundo procedimiento, esta correlación es positiva y bastante alejada del cero, si bien sigue siendo no significativa, tal vez por el pequeño tamaño de muestra. Se comprueba también que el valor de esta correlación parcial es el mismo que el de la correlación simple entre los residuos del ejemplo anterior.

CORRELACIÓN MÚLTIPLE.

Como ya se ha afirmado, la correlación parcial no involucra la noción de variables independientes y dependientes sino que es una medida de interdependencia. Por otro lado, el coeficiente de correlación múltiple se aplica a la situación en que una variable, a la que se puede seguir llamando *Y*, ha sido aislada para examinar su relación con el conjunto de las otras variables. Este coeficiente de correlación viene determinado por la expresión

$$R_{1.23} = \sqrt{\frac{r_{12}^2 + r_{13}^2 - 2\ r_{12}\ r_{13}\ r_{23}}{1 - r_{23}^2}}$$

Al igual que ocurría con el coeficiente de correlación simple, R^2 es el ***coeficiente de determinación múltiple***.

El valor de un coeficiente de correlación múltiple, *R*, se encuentra entre cero y uno. Cuanto más se acerque a uno mayor es el grado de asociación entre las variables. Y cuanto más se acerca a 0 la relación lineal es peor.

Existe una relación entre el coeficiente de correlación múltiple y los diferentes coeficientes de correlación parcial, que puede facilitar el cálculo de aquél, esta es

$$1-R^2_{1.23} = (1-r^2_{12})(1-r^2_{13.2})$$
$$1-R^2_{1.234} = (1-r^2_{12})(1-r^2_{13.2})(1-r^2_{14.23})$$

Una correlación múltiple de orden *k-ésimo*

$$1-R^2_{1.23\ldots k} = (1-r^2_{12})(1-r^2_{13.2})\ldots(1-r^2_{1k.23\ldots(k-1)})$$

Como se ve, la generalización de todas estas fórmulas para *k* variables es automática.

Así como en la prueba de ajuste de una regresión múltiple, R^2 es la fracción de la suma de cuadrados de las desviaciones de *Y* de su media, atribuible a la regresión, en tanto que ($1-R^2$) es la fracción no asociada a la regresión; ahora, la prueba de hipótesis nula, de que la correlación múltiple en la población es cero, es idéntica a la prueba *F* de la hipótesis nula que $\beta_1=\beta_2=\ldots=\beta_k=0$; y ésta es

$$F = \frac{(n-k)\ R^2}{(k-1)(1-R^2)}$$

Siendo *R* el coeficiente de determinación múltiple. Esta F_o se contrasta con la $F_{(k-1,n-k;\alpha)}$.

Ejemplo.-

Siguiendo con el ejemplo anterior, el coeficiente de correlación múltiple de la variable producción con las variables: precipitación en Noviembre, temperatura media en Julio, precipitación en Julio y radiación en Julio

Archivo del programa *SAS* (cormul.sas).-

```
title 'Correlación multiple';
dm 'log;clear;output;clear;';
options  ls = 65  ps = 60;
data cormul;
infile '\corpar.dat';
input prod prenov tmjul prejul radjul;
title 'Correlación múltiple por el procedimiento correlaciones canónicas';
proc cancorr;
var prod;
with prenov tmjul prejul radjul;
run;
title 'Correlación multiple, raíz cuadrada del coeficiente de derterminacion
de la regresión';
proc reg;
model prod = Prenov tmjul prejul radjul;
run;
```

Como se ve, la correlación múltiple es un caso particular de la correlación canónica en la que hay una sola variable (**VAR**).

En el archivo de resultado se observa que la correlación múltiple vale 0.8136, y la ***F*** de la prueba de significación vale 2.9383, que es no significativa al 0.05.

CORRELACIÓN CANÓNICA

La técnica del análisis de la correlación canónica se entiende mejor considerándola como una extensión de la regresión múltiple y de la correlación. El análisis de regresión múltiple consiste en encontrar la mejor combinación lineal de p variables independientes, X_1, X_2,..., X_p, para predecir la variable dependiente Y. La correlación múltiple es la correlación simple entre Y y sus valores estimados por la ecuación de regresión, $\hat{Y}$. Por tanto, el objetivo en los análisis de regresión y correlación múltiple está en examinar la relación entre varias variables, X, y una variable, Y.

El análisis de correlación canónica se aplica a situaciones donde es apropiada la técnica de la regresión pero para más de una variable dependiente. Aunque otra aplicación del análisis de correlación canónica es como un método para determinar la asociación entre dos grupos de variables. Es una generalización de la regresión múltiple al caso de más de una variable dependiente.

Este análisis esta íntimamente relacionado con el análisis canónico discriminante y tiene ciertas propiedades análogas al análisis de componentes principales y al análisis factorial, en el que en lugar de tratar de estudiar las dependencias internas entre las variables de un mismo grupo, en el caso de la correlación canónica lo que se estudia es la relación o dependencia entre dos grupos de variables.

Recuérdese que el análisis de regresión múltiple trataba de encontrar la combinación lineal de p variables, X_1, X_2,..., X_p, que mejor predigan la variable dependiente Y. El coeficiente de correlación múltiple es la correlación simple entre Y y su predicción por medio de la ecuación de regresión.

En el análisis de correlación canónica se examina la relación lineal entre un grupo de variables, X, y un grupo, o más de un grupo, de variables Y. Por lo que la diferencia es que ahora se tiene más de una variable Y. La técnica consiste en encontrar una combinación lineal de las variables X ($V_1=b_1X_1+b_2X_2+...+b_pX_p$) y otra combinación lineal de las variables Y ($U_1=a_1Y_1+a_2Y_2+...+a_qY_q$) de tal manera que la correlación entre U y V sea máxima. Después encontrar otras dos combinaciones lineales para cada grupo de variable que tenga correlación máxima y así sucesivamente se encuentran un conjunto de combinaciones lineales para cada grupo de variables que tienen correlación máxima. A estas combinaciones lineales se denominan *variables canónicas*, y las correlaciones entre los correspondientes pares de variables canónicas se denominan *correlaciones canónicas*.

En una aplicación común de esta técnica las Y se interpretan como variable *respuesta* o variables *dependiente*, mientras que las variables X representan variables *predictivas* o variables *independientes*. Las variables Y pueden ser más dificultoso de medir que X como ocurre con los problemas de calibración.

El análisis de correlación canónica se aplica a situaciones en las que es adecuada la técnica de la regresión pero existe más de una variable dependiente. Otra aplicación útil es para probar la independencia entre los dos grupos de variables, Y y X, como se verá dentro de un momento.

Ejemplos de aplicaciones de la correlación canónica pueden ser el estudio para relacionar las características de ciertas variedades de trigo y características de las harinas resultantes. En este estudio fue posible concluir que el trigo deseable es el que tiene valores altos de textura, densidad y contenido en proteínas y el que tiene valores bajos en granos deteriorados y en productos extraños. Similarmente, una harina buena debe tener un alto contenido en proteína y bajos valores de ceniza. La correlación canónica también puede usarse en psicología para calibrar dos grupos de pruebas de inteligencia hechas a los mismos individuos. También, ha sido usado para relacionar las combinaciones lineales de las escalas de personalidad con las combinaciones lineales de las pruebas psicológicas realizadas.

El análisis de correlación canónica es, de las técnicas multivariantes, uno de los menos utilizados. Esto es debido, en parte, a la dificultad que se puede encontrar a la hora de interpretar los resultados.

Ejemplo hipotético.-

Supóngase cierta especie, por ejemplo, caprino, en la que se toma una muestra de diez individuos en

los que se miden dos variables productivas de leche, como pueden ser *producción máxima diaria* (Y_1, *producción* en adelante) y *porcentaje de nitrógeno total* (Y_2, *porcentaje* en adelante), y dos variables de conformación, como pueden ser *longitud total del cuerpo* (X_1, *longitud* en adelante) y *anchura de las caderas* (X_2, *anchura* en adelante). Los datos son

Individuo	Y_1	Y_2	X_1	X_2
1	122	40	332	116
2	120	42	320	107
3	126	44	339	119
4	125	39	336	114
5	120	38	321	106
6	127	45	336	119
7	128	49	347	128
8	130	39	349	129
9	123	41	338	111
10	124	42	333	112

Las variables de conformación están medidas en la misma escala pero en diferentes unidades.

Los estadísticos básicos de estas variables son

Variable	$\overline{X}$	S
Y_1	124.50	3.3416
Y_2	41.90	3.3483
X_1	335.10	9.4334
X_2	116.10	7.8662

y la matriz de correlaciones es

	Y_1	Y_2	X_1	X_2
Y_1	1.00000	0.41212	0.92525	0.92782
Y_2		1.00000	0.38379	0.46868
X_1			1.00000	0.90874
X_2				1.00000

Como se observa en esta matriz, la variable *producción* está altamente correlacionada tanto con la *longitud* como con la *anchura*, mientras que la variable *proporción* no esta tan correlacionada con las variables de conformación. Las dos variables de conformación están, lógicamente, muy correlacionadas entre ellas, mientras que las dos variables de producción, siendo una de cantidad y la otra de proporción, están medianamente correlacionadas.

Conceptos básicos de la correlación canónica.-

Supóngase que se va ha estudiar la relación entre un grupo de variables, $x_1, x_2,..., x_p$ y otro grupo de variables, $y_1, y_2,..., y_q$. Las variables x pueden ser vistas como variables independientes o predictoras, mientras que las variables y se pueden considerar como variables dependientes o variables respuesta. Se asume que, en una muestra dada, se le resta a los datos originales la media de cada variable, por lo que la media de todas las x y todas las y valen cero.

Primera correlación canónica.-

La idea básica del análisis de correlación canónica comienza buscando una combinación lineal de las *y*, tal como

$$U_1 = a_1y_1+a_2y_2+...+a_qy_q$$

y una combinación lineal de las *x*, tal como

$$V_1 = b_1x_1+b_2x_2+...+b_px_p$$

Para cualquier elección de los coeficientes, *a* y *b*, se puede calcular los valores U_1 y V_1 de cada individuo de la muestra. Para los *N* individuos de la muestra se puede calcular la correlación simple de los *N* pares, U_1 y V_1, de la manera usual. La correlación resultante dependerá de la elección de los valores de *a* y *b*.

En el análisis de correlación canónica, se seleccionan los valores de los coeficientes *a* y *b* de manera que *maximice* la correlación entre U_1 y V_1. Como consecuencia de esta particular elección de los coeficientes, a la combinación lineal U_1 se le denomina *primera variable canónica* de las *y*, y a la combinación lineal V_1 se le denomina *primera variable canónica* de las *x*. Nótese que tanto U_1 como V_1 tienen media cero. La correlación entre U_1 y V_1 se le denomina *primera correlación canónica*.

La primera correlación canónica es, por tanto, la correlación mayor posible entre la combinación lineal de las *x* y la combinación lineal de las *y*. En este sentido, es la correlación lineal máxima entre el grupo de las *x* y el grupo de las *y*. La primera correlación canónica es análoga al coeficiente de correlación múltiple entre una variable *Y* y un grupo de variables *X*. La diferencia es que en la correlación canónica hay varias *y* por lo que también hay que encontrar una combinación lineal de ellas.

El *SAS* provee los coeficientes, *a* y *b*, estos son

Coeficientes		*Coeficientes tipificados*	
a_1=0.29107	b_1=0.04948	a_1=0.9727	b_1=0.4667
a_2=0.01864	b_2=0.07077	a_2=0.0624	b_2=0.5567

Como se ve, los coeficientes tipificados se calculan multiplicando los coeficientes por la desviación típica de la variable, así 0.9729=0.29107 x 3.34166.

Las variables canónicas se calculan con los coeficientes no tipificados y, tal como se dijo anteriormente, con la diferencia de cada dato original con su media, de manera que las primeras variables canónicas (U_1 y V_1) para, por ejemplo, el primer individuo de la tabla de datos es

$$U_1 = 0.29107(122-124.5) + 0.01864(40-41.9) = -0.76309$$
$$V_1 = 0.04948(332-335.1) + 0.07077(116-116.1) = -0.16045$$

Si se calcula U_1 y V_1 para todos los individuos y se estima la correlación lineal simple de estas dos variables se obtiene la primera correlación canónica, que en este caso vale 0.9499. Este valor representa la correlación mayor posible entre cualquier combinación lineal de las variables independientes y cualquier combinación lineal de las variables dependientes. Particularmente, es mayor que cualquier correlación entre una *X* y una *Y*, como se puede comprobar con la matriz de correlaciones de las variables expuesta anteriormente.

Un método para interpretar el valor relativo de cada variable en la combinación lineal canónica, es viendo el valor de los coeficientes tipificados. Así, para las *Y*, la primera variable canónica viene determinada

fundamentalmente por la variable *producción*, lo que quiere decir que un individuo que produzca relativamente mucho tendrán un alto valor de la primera variable canónica U_1. Mientras que en el valor de la primera variable canónica de las X tiene una influencia ligeramente superior la *anchura* que la *longitud*, se puede considerar que en ambas la influencia es la misma.

Otro método para interpretar el valor relativo de cada variable en la combinación lineal canónica, es viendo el valor de la correlación de cada variable original con su variable canónica (o con la variable canónica del otro grupo de variables). Estos valores los da el *SAS* por defecto y son, en el caso del ejemplo hipotético

	U_1	V_1
Y_1	0.9984	0.9484
Y_2	0.4633	0.4400
X_1	0.9239	0.9726
X_2	0.9317	0.9808

Como se ve, la primera variable canónica de las Y (U_1) está altamente correlacionada con la Y_1 y medianamente correlacionada con la Y_2 por lo que se puede determinar que la primera variable canónica viene determinada fundamentalmente por la variable *producción*, lo que quiere decir que un individuo que produzca relativamente más, tendrán un alto valor de la primera variable canónica U_1. Mientras que en el valor de la primera variable canónica de las X (V_1) tiene una correlación ligeramente superior con *anchura* que la *longitud*, pero en ambas es elevada.

De todo esto se deduce que los individuos muy *anchos* y *largos* tienen una elevada *producción* pero no indica nada de la *proporción*. Lógicamente, en un ejemplo real con más variables el análisis de estos coeficientes puede ser gran utilidad para conducir a múltiples y variadas conclusiones.

Segunda (y sucesivas) correlación canónica.-

Se puede realizar interpretaciones adicionales de la relación entre las X y las Y obteniendo otro conjunto de variables canónicas y su correspondiente correlación canónica. Concretamente, se puede hallar la segunda variable canónica V_2 (combinación lineal de las x) y la correspondiente variable canónica U_2 (combinación lineal de las y). Los coeficientes de estas combinaciones lineales se eligen teniendo en cuenta las siguientes condiciones:

1. V_2 esta incorrelacionada con V_1 y U_1.
2. U_2 esta incorrelacionada con V_1 y U_1.
3. Una vez cumplidas las condiciones anteriores, U_2 y V_2 tienen la máxima correlación posible.

La correlación entre U_2 y V_2 se denomina *segunda correlación canónica* y necesariamente es menor o igual que la primera correlación canónica.

Este paso se repite para calcular la tercera, cuarta, *etc*., variables canónicas. El número máximo de correlaciones canónicas y sus correspondientes variables canónicas es igual al número mínimo de variable en los grupos, esto es, si hay por ejemplo 10 variables X y 5 variables Y el número de correlaciones canónicas que se podrán calcular será de 5.

Para el ejemplo hipotético, la segunda correlación canónica vale 0.2147. Los coeficientes de las segundas variables canónicas son

Coeficientes		*Coeficientes tipificados*	
a_1=-0.15215	b_1=-0.24912	a_1=-0.5084	b_1=-2.3501
a_2= 0.32726	b_2= 0.29625	a_2= 1.0958	b_2= 2.3304

y las correlaciones con las variables originales son

	U_2	V_2
Y_1	-0.0569	-0.0122
Y_2	0.8862	0.1903
X_1	-0.0499	-0.2323
X_2	0.0418	0.1948

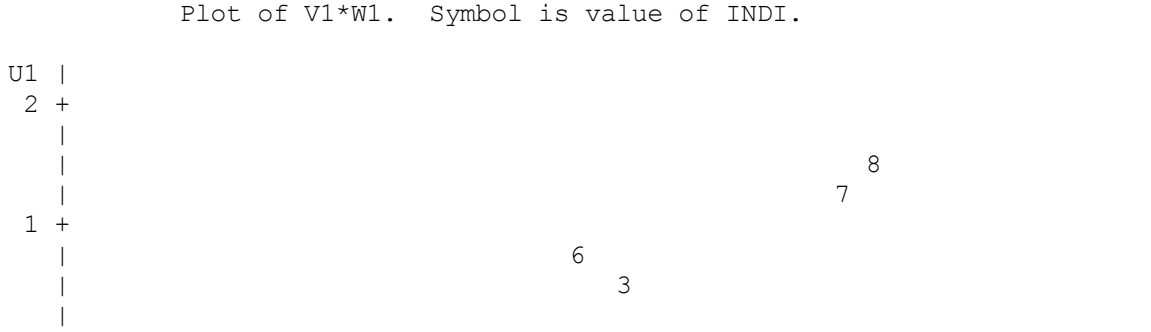
Tanto con los coeficientes tipificados como con las correlaciones se observa que en la segunda variable canónica de las Y (U_2) tiene una influencia negativa la variable Y_1, baja en valor absoluto comparada con la influencia positiva de la variables Y_2, esto significa que en la segunda variable canónica de las Y tiene una gran influencia positiva la variable *proporción* y una leve influencia negativa la variable *producción*. Mientras que en el valor de la segunda variable canónica de las X (V_2) tienen prácticamente la misma influencia las dos variables pero en sentido contrario, esto es, la *longitud* tiene una influencia negativa en el valor se su segunda variable canónica y la *anchura* tiene una influencia positiva, pero en ambas variables esta influencia es más bien baja.

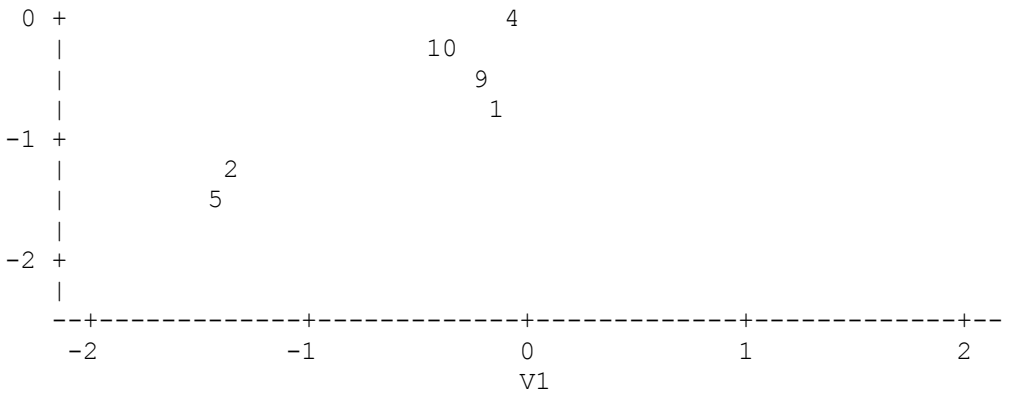
Pruebas de hipótesis.-

El paquete *SAS* realiza la prueba de razón de verosimilitud para probar las correlaciones canónicas. En el caso del ejemplo hipotético, la razón de verosimilitud para la primera correlación canónica es de 0.09318 que corresponde a una F aproximada de 6.8276 que para 4 y 12 grados de libertad es significativa al 0.01 por lo que se puede concluir que la primera correlación canónica es significativa, por lo que están correlacionadas ambos tipos de variables (las de producción con las de conformación) y son significativamente ciertas las conclusiones que obtuvimos al analizar los coeficientes tipificados y las correlaciones de las primeras variables canónicas. Mientras que la razón de verosimilitud de la segunda correlación canónica es 0.95389 (prácticamente 1), cuya F aproximada es 0.3383 (inferior a 1) por lo que la segunda correlación canónica es no significativa, y por tanto no son significativas ciertas las conclusiones que obtuvimos al analizar los coeficientes y las correlaciones de las segundas variables canónicas.

Representación de los valores de las variables canónicas.-

Puede ser útil la representación, en unos ejes cartesianos, de los individuos siendo el valor de sus coordenadas la de los valores de las variables canónicas, U_1 y V_1. Para el ejemplo hipotético esta representación sería





Como se observa, existe una pendiente positiva acentuada como era de esperar por la significación de la primera correlación canónica (0.9499). Para datos multivariante normal, esta gráfica será una elipse de dispersión que podrá ser útil para detectar posibles datos erróneos o individuos peculiares.

Programa *SAS*.-

Para obtener los resultados y análisis del ejemplo hipotético. el programa *SAS* es **corcan1.sas**

```
options ls=64 ps=30;
data corrcan;
infile 'corcan.dat';
input indi y1 y2 x1 x2;
proc cancorr corr out=canonica;
var y1 y2;
with x1 x2;
run;
proc print;
run;
proc plot;
plot V1*W1=indi;
run;
```

Correlación canónica usando los componentes principales.-

Otro método para examinar la relación entre un grupo de variables X y de variables Y es el siguiente:

1. Obtener los componentes principales de y_1, y_2,..., y_q y simbolizarlos por D_1, D_2,..., D_q.

2. Obtener los componentes principales de x_1, x_2,..., x_q y simbolizarlos por C_1, C_2,..., C_q.

3. Elegir las primeras m componentes principales de cada grupo.

4. Calcular la correlación entre C_1 y D_1; entre C_2 y D_2; ...; C_m y D_m.

Las correlaciones calculadas en el paso cuatro son, en general, menores que las correlaciones

canónicas, pero las componentes principales pueden ser interesantes por ellas mismas. Las componentes principales explican el máximo de varianza *dentro* del grupo de variables, mientras que las variables canónicas maximizan la correlación *entre* los dos grupos de variables.

Puesto que, por ejemplo, C_1 y D_2 pueden tener una correlación diferente de cero, puede ser útil calcular dichas correlaciones además de las descritas en el paso 4.

Programa *SAS*.- corcan2.sas

```
Options ls=64 ps=30;
Data corrcan;
infile 'corcan.dat';
input indi y1 y2 x1 x2;
proc princomp out=ccom;
var y1 y2;
run;
proc princomp
data=ccom(rename=(prin1=priny1 prin2=priny2))
out=ccom;
var x1 x2;
run;
proc corr data=ccom(keep=priny1 priny2 prin1 prin2);
run;
```

Aplicación al análisis discriminante.-

Cuando se estudió el Análisis Discriminante, se vio que consiste en la clasificación de un individuo en una de k>=2 poblaciones en base a las medidas X_1, X_2,..., X_p. Para clasificar un individuo se calcula cada una de las k funciones discriminantes y se asigna el individuo a la población para la cual ha dado un mayor valor la función discriminante. Este proceso es el aspecto predictivo de la clasificación. Para fines descriptivos está el Análisis Canónico Discriminante. Recuerde que en este análisis las variables canónicas se deducían de manera que fuera máxima la diferencia entre los grupos. Conocidas las correlaciones canónicas se puede estudiar el planteamiento de estas funciones discriminantes .

Se puede comenzar definiendo un conjunto de nuevas variables Y_1. Y_2,..., Y_{k-1}, que serán variables de diseño o de incidencia, indicando el valor de cada variable la pertenencia o no a uno de los grupos. Recuerdes cuando se estudio el análisis discriminante que se dijo que se necesitan k-1 variables de diseño para describir k grupos. Por ejemplo, supóngase que se han medido p variables en k=4 grupos, el valor de las variables Y_1. Y_2 e Y_3 serán

grupo	Y_1	Y_2	Y_3
1	1	0	0
2	0	1	0
3	0	0	1
4	0	0	0

Si un individuos pertenece al grupo 1 se le da el valor de 1 a la variable Y_1 y cero a las otras dos variables Y, *etc*, con lo que se tendrá q=k-1 variables Y y p variables X. Con estas variables se realiza el análisis de correlaciones canónicas en el que se obtendrán las variables canónicas U_1 y V_1, cuyo número es, tal como se dijo antes, igual al mínimo valor de p o de q.

La variable canónica V_1 es la función discriminante que se vio en el análisis discriminante. Como V_1

es la combinación lineal de las X que maximiza la correlación con U_1, en este sentido, V_1 maximiza la correlación con las variables de diseño que representan los grupos y además maximiza la diferencia entre los grupos. Similarmente, V_2 exibe la máxima diferencia entre grupos con la condición previa de que este incorrelacionada con V_2.

Ejemplo.-

Se ha medido en 429 cabras dos tipos de medidas. medidas productivas de la leche total y medidas productivas de una fracción de la leche. Las medidas productivas de la leche total son; producción total en *Kg* (**prod**), porcentaje de proteína total (**prot**), porcentaje de caseína total (**cas**), porcentaje de grasa (**gras**) y porcentaje de lactosa (**lac**). Las medidas productiva de la fracción de la caseína son: porcentaje de caseína α (**alfa**), porcentaje de la caseína β (**beta**) y porcentaje de la caseína κ (**kapa**).

Se quiere saber si las variables de la fracción de las caseínas influyen en la producción total

Archivo del programa *SAS* (coca.sas).-

```
options ls=80 ps=60 ;
data coca;
infile 'coca.dat';
input prod prot cas gras lac   alfa beta kapa;
proc cancorr corr out=canonica;
var prod prot cas gras lac;
with alfa beta kapa;
run;
proc plot;
plot V1*W1;
run;
```

Archivo de resultados (coca.lst).-

. Como se ha puesto la opción **CORR** la primera salida es la de las correlaciones entre las variables originales, que son muy elevadas como era de esperar por las características de los datos.

. Como el grupo más pequeño de variables es el de tres variables, solo se puede estimar tres correlaciones canónicas. Esta es la siguiente salida, la primera correlación canónica vale 0.9978, la segunda vale 0.3738 y la tercera 0.2539.

. Después de los valores propios vienen las tres pruebas de hipótesis para las tres correlaciones canónicas, indicando que las tres son altamente significativamente diferentes de cero.

. Después vienen los coeficientes de las variables canónicas, los coeficientes tipificados y las correlaciones de cada variable original con las variables canónicas. Estudiando los coeficientes tipificados y las correlaciones se observa que para la primera variable canónica de las variables productivas totales, la que más influye en ella es la caseína y la que menos la producción total, mientras que en la primera variable canónica de las variables de la fracción de caseína la que más influye es la α-caseína y la que menos la κ-caseína.

. En la salida del procedimiento **PLOT** se observa claramente que ambos grupos de variables están altamente correlacionadas, pues la nube de dispersión es prácticamente una recta