Gónzalez.N.; Abad, J. & Lèvy, J.P. (2006) " Normalidad y otros supuestos en análisis de covarianzas" en Lévy. J. (Ed.): " Modelización conestructuras de covarianzas" Netbiblo. La Coruña (31-57)

CAPÍTULO 3: NORMALIDAD Y OTROS SUPUESTOS EN ANÁLISIS DE COVARIANZAS

1. Introducción

Este capítulo aborda el estudio de los supuestos subyacentes al análisis de covarianzas. El primero de ellos y quizá, el más importante, es el requisito de que los datos disponibles sigan de forma conjunta una distribución normal. El cumplimiento de este supuesto resulta necesario con el fin de garantizar la validez de los resultados. Es por ello, en este capítulo, se dedica un primer apartado al estudio de dicho supuesto, tanto de forma univariante como multivariante, proponiéndose pruebas gráficas y estadísticas que permiten contrastar su cumplimiento.

El siguiente apartado se ocupa del análisis de la linealidad de los datos, otro de los supuestos requeridos para la utilización de los modelos de ecuaciones estructurales. La comprobación de este supuesto se realiza a través del estudio de los gráficos de dispersión.

Finalmente, se abordan algunos procedimientos que permiten detectar la existencia de valores atípicos o *outliers* entre los datos disponibles, tanto desde una perspectiva univariante como multivariante. La presencia de valores atípicos entre las observaciones puede influir en las relaciones entre variables y desvirtuar los resultados del análisis de ecuaciones estructurales, de ahí la importancia de detectar y, en su caso, eliminar dichas observaciones.

Todos los apartados se ilustran con un ejemplo práctico en el que se muestra cómo aplicar el programa estadístico SPSS y los programas de análisis de ecuaciones estructurales LISREL y AMOS en el estudio del cumplimiento de estas hipótesis.

2. Normalidad

Como acabamos de señalar, uno de los principales supuestos sobre el que se asienta el modelo de ecuaciones estructurales es que las variables observadas siguen de forma conjunta una distribución normal multivariante dado que, en caso contrario, ni los estimadores planteados serían óptimos, ni los contrastes individuales de los parámetros ni los de ajuste global resultarían adecuados.

En este sentido, el que cada una de estas variables verifique la normalidad univariante resulta ser una condición necesaria pero no suficiente para que conjuntamente sigan una normal multivariante, es decir, si la distribución conjunta es normal multivariante, cada una de las marginales es una normal univariante, pero no a la inversa. Por este motivo, se hace necesario comprobar en primer lugar que todas las variables consideradas individualmente se distribuyen normalmente para, a continuación, contrastar que todas ellas en conjunto cumplen la normalidad multivariante.

Por todo ello, vamos a referirnos en primer lugar a algunos procedimientos que permiten estudiar la normalidad univariante, posteriormente trataremos técnicas para contrastar la hipótesis de normalidad multivariante y, finalmente, ilustraremos todo ello con un ejemplo de aplicación en el que se mostrará cómo utilizar distintos programas estadísticos a la hora de evaluar la normalidad de los datos.

2.1. Normalidad univariante

Para estudiar la normalidad univariante de los datos, podemos comenzar realizando una inspección visual de los mismos utilizando para ello el histograma, que nos permitirá observar si la forma de la distribución es similar a la de la campana de Gauss (unimodal, campaniforme, simétrica,…). Otra opción es el gráfico de probabilidad normal, en el que se representan los datos frente a la teórica distribución normal de forma que los puntos deberían aproximarse a una línea recta para poder admitir que son normales, aunque conviene tener en cuenta que siempre tenderá a observarse una mayor desviación en los extremos. Además, los gráficos de probabilidad normal también permiten conocer la causa de esa desviación: si los puntos se disponen en forma de "U" o con alguna curvatura, ello se debe a que la distribución es asimétrica, mientras que si presentan forma de "S" significará que la distribución no es mesocúrtica.

El problema que plantean estos métodos gráficos es que siempre presentan un importante grado de subjetividad: la forma del histograma depende del número y amplitud de los intervalos que se consideren, y es el investigador el que ha de juzgar en qué medida los puntos se ajustan a una recta en el gráfico de probabilidad normal.

Por tanto, para valorar la normalidad de los datos desde una perspectiva más objetiva, resulta necesario emplear otro tipo de procedimientos: los denominados contrastes de normalidad, de entre los cuales destacaremos los siguientes:

Contraste de Kolmogorov-Smirnov-Lilliefors

Se trata de una modificación del contraste de bondad de ajuste de Kolmogorov-Smirnov para el caso en que la distribución de contraste es una normal de parámetros desconocidos (que es la situación más habitual). Este contraste compara la función de distribución empírica muestral con la teórica de una población normal, de manera que se rechazaría la hipóteisis nula de normalidad cuando el valor experimental del estadístico (que sería la mayor diferencia registrada entre ambas funciones) es significativamente grande. Este contraste no resulta muy apropiado cuando el tamaño de muestra es pequeño porque para ese tipo de muestras su potencia es baja.

Contraste de Shapiro-Wilks

Mide el grado de ajuste a una recta de las observaciones de la muestra representadas en un gráfico de probabilidad normal, de forma que se rechazará la hipótesis nula de normalidad cuando el ajuste sea malo, situación que se corresponde con valores pequeños del estadístico de contraste. Este contraste es el más adecuado cuando el tamaño de muestra es pequeño (no superior a 50) y tampoco requiere que los parámetros de la distribución estén especificados.

Contrastes de asimetría y curtosis

Permiten determinar si la forma de la distribución de las observaciones muestrales se aleja significativamente de la de un modelo normal en lo que a su simetría y curtosis se refiere. Antes de plantear este contraste, vamos a definir dichos conceptos.

Una distribución es simétrica cuando los valores que están a la misma distancia de la media tienen igual frecuencia, mientras que es asimétrica a la derecha (o con asimetría

positiva) cuando los valores bajos de la variable son los más frecuentes, y asimétrica a la izquierda (o con asimetría negativa) en caso contrario.

La curtosis, por su parte, se refiere al grado de apuntamiento que presenta una distribución al compararla con la distribución normal. Una distribución es leptocúrtica (o con curtosis positiva) cuando es más apuntada y con colas menos gruesas que la normal, platicúrtica (o con curtosis negativa) si es más aplastada y con colas más gruesas que la distribución normal, y mesocúrtica si es igual de apuntada que la normal.

Dada una variable aleatoria X con media o esperanza $E(X) = \mu$ y con varianza $\sigma^2 = E[(X-\mu)^2]$, se definen los coeficientes poblacionales de asimetría, γ_1, y de curtosis, γ_2, como:

$$\gamma_1 = \frac{\mu_3}{\sigma^3} \qquad \text{y} \qquad \gamma_2 = \frac{\mu_4}{\sigma^4} - 3$$

donde μ_k es el momento central de orden k definido como $\mu_k = E[(X-\mu)^k]$.

La interpretación de estos coeficientes, tal y como se muestra en las figuras 1 y 2, es la siguiente: si γ_1 es positivo, la distribución es asimétrica a la derecha, si es negativo, lo es a la izquierda, y si es nulo, es simétrica; por su parte, si γ_2 es positivo, la distribución es más apuntada que la normal, si es negativo, es más aplastada que la normal, y si es nulo, tiene una curtosis como la de la distribución normal. Por tanto, cuando una variable aleatoria sigue una distribución normal, ambos coeficientes son nulos.

Figura 1. Simetría

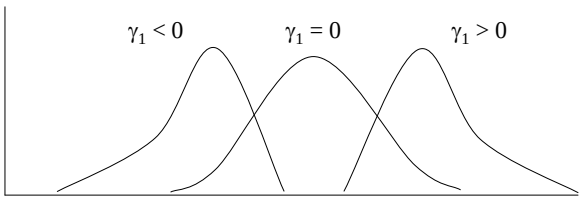
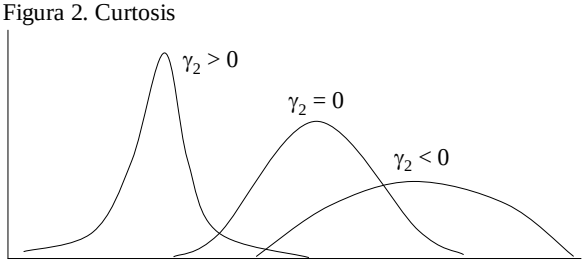


Figura 2. Curtosis



Dado que, como ya hemos señalado anteriormente, los valores de estos coeficientes se refieren a toda la población, un procedimiento para decidir a partir de una muestra si la distribución de la que procede es o no normal consiste en estimar el valor de γ_1 y γ_2 y valorar si difieren significativamente de cero, en cuyo caso, rechazaríamos la hipótesis de normalidad.

A tal fin, siendo $\{x_1, x_2, \ldots, x_n\}$ una muestra aleatoria de tamaño n de la variable aleatoria X se definen los coeficientes muestrales de asimetría, g_1, y de curtosis, g_2, como:

$$g_1 = \frac{m_3}{S^3} \qquad \text{y} \qquad g_2 = \frac{m_4}{S^4} - 3$$

donde m_k es el momento muestral central de k-ésimo orden: $m_k = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^k$, S^2 es la varianza muestral: $S^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$, y $\bar{x}$ es la media muestral: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$.

Estos coeficientes muestrales no son, en general, estimadores insesgados de γ_1 y de γ_2, por lo que se plantean sendos estimadores G_1 y G_2 que, bajo la hipótesis nula de normalidad que queremos contrastar, no tienen sesgo y sus varianzas son conocidas. Las respectivas expresiones de G_1 y G_2 y sus relaciones con g_1 y g_2 son las siguientes:

$$G_1 = \left(\frac{\sqrt{n(n-1)}}{n-2}\right)\frac{m_3}{S^3} = \frac{\sqrt{n(n-1)}}{n-2}g_1$$

$$G_2 = \frac{(n+1)(n-1)}{(n-2)(n-3)}\left[\frac{m_4}{S^4} - 3\left(\frac{n-1}{n+1}\right)\right] = \frac{n-1}{(n-2)(n-3)}[(n+1)g_2 + 6]$$

Una vez estimados a partir de la información muestral los coeficientes de asimetría y curtosis y sus respectivas varianzas, se pueden construir diversos estadísticos[1] que permitirán contrastar las hipótesis nulas de que los parámetros poblacionales γ_1 y γ_2 son nulos. Dado que la distribución de estos estadísticos de contraste, a los que denominaremos $z(G_1)$ y $z(G_2)$, se aproxima a una normal tipificada, para un nivel de significación del 5%, un valor experimental de $z(G_1)$ superior en valor absoluto a 1'96 permite rechazar la hipótesis nula $\gamma_1 = 0$ (la distribución es simétrica) y, de forma análoga, si

$|z(G_2)| > 1'96$, entonces se rechaza la hipótesis nula $\gamma_2 = 0$ (la distribución es mesocúrtica). Partiendo de los dos estadísticos de contraste individual $z(G_1)$ y $z(G_2)$, se puede efectuar un contraste conjunto de la simetría y curtosis de la muestra utilizando el siguiente estadístico:

$$k^2 = [z(G_1)]^2 + [z(G_2)]^2$$

que se distribuirá asintóticamente como una χ^2 con dos grados de libertad, de forma que un valor de k^2 superior a 5'99 permite rechazar la hipótesis nula $\gamma_1 = \gamma_2 = 0$ (simetría y curtosis igual a la normal) dado un nivel de significación del 5%.

De este modo, rechazar la hipótesis nula, bien en cualquiera de los dos contrastes individuales, bien en el contraste conjunto, supone rechazar también la hipótesis de que los datos proceden de una distribución normal.

2.2. Normalidad multivariante.

Hasta ahora nos hemos referido a métodos que permiten contrastar la hipótesis de normalidad para cada una de las variables observables consideradas por separado. Sin embargo, como ya se ha señalado, el modelo de ecuaciones estructurales se asienta en el supuesto de que las variables observadas siguen de forma conjunta una distribución normal multivariante. En este sentido, el que cada una de estas variables verifique normalidad univariante resulta ser una condición necesaria pero no suficiente para que conjuntamente sigan una normal multivariante (si la distribución conjunta es normal multivariante, cada una de las marginales es una normal univariante, pero no a la inversa).

Por este motivo, una vez comprobada la normalidad de cada una de las variables observadas consideradas individualmente, se hace necesario también contrastar la hipótesis de normalidad multivariante. A tal fin, MARDIA (1970) propuso algunos tests para contrastar si la asimetría y la curtosis multivariantes del conjunto de variables observables permite asumir o no la hipótesis de normalidad. Estos contrastes se construyen a partir de las siguientes medidas muestrales de asimetría y curtosis multivariantes:

---

[1] En BOLLEN (1989; 421) se relaciona un conjunto de estadísticos recomendados por D'AGOSTINO (1986) para el contraste de la simetría y la curtosis, indicando cuál resulta más adecuado según el tamaño de la muestra con que se trabaje.

– Asimetría: $G_{1,p} = \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left[(\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\mathbf{S}}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}})\right]^3$

– Curtosis: $G_{2,p} = \frac{1}{n}\sum_{i=1}^{n}\left[(\mathbf{x}_i - \bar{\mathbf{x}})'\hat{\mathbf{S}}^{-1}(\mathbf{x}_i - \bar{\mathbf{x}})\right]^2$

donde n representa el número total de observaciones, $\mathbf{x}_i$ y $\mathbf{x}_j$ son vectores columna con los valores de todas las variables para las observaciones i-ésima y j-ésima, respectivamente, $\bar{\mathbf{x}}$ es el correspondiente vector columna de medias muestrales y $\hat{\mathbf{S}}^{-1}$ es la inversa de la matriz de varianzas-covarianzas muestral.

Los estadísticos de contraste $z(G_{1,p})$ y $z(G_{2,p})$ obtenidos a partir de $G_{1,p}$ y $G_{2,p}$ se distribuyen asintóticamente según una normal estándar por los que su interpretación es semejante a la ya comentada para los estadísticos de asimetría y curtosis univariante $z(G_1)$ y $z(G_2)$: valores experimentales que en valor absoluto sean mayores que 1'96 permiten rechazar a un nivel de significación del 5% las respectivas hipótesis nulas de distribución multivariante simétrica y mesocúrtica. Asimismo, también se puede realizar un contraste conjunto de simetría y mesocurtosis multivariantes utilizando el estadístico:

$$k_p^2 = [z(G_{1,p})]^2 + [z(G_{2,p})]^2$$

que se aproxima a una distribución χ^2 con dos grados de libertad y que también se interpreta de forma análoga al estadístico conjunto k^2 de normalidad univariante, es decir, se rechaza la hipótesis nula para valores experimentales mayores que 5'99 dado un nivel de significación del 5%.

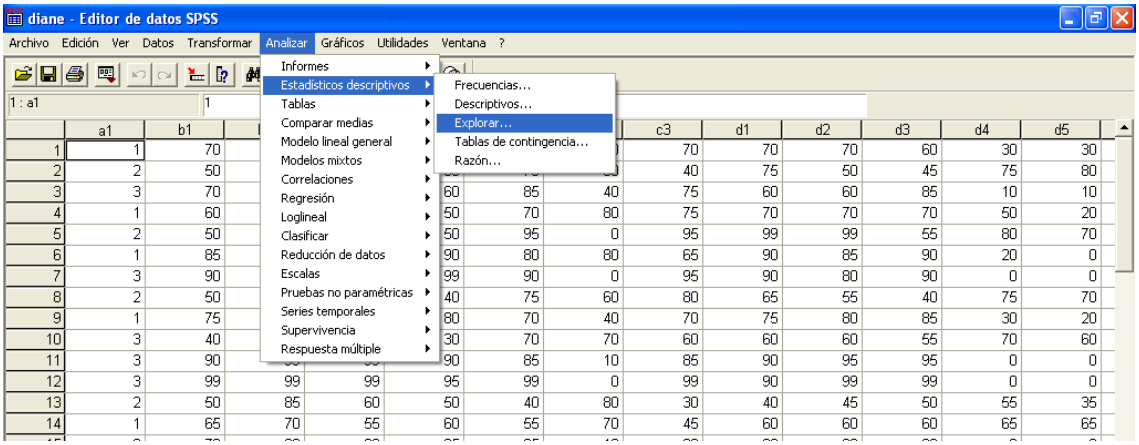
2.3. Aplicación

Con el objetivo de ilustrar la comprobación de los supuestos que se analizan en este capítulo, se examinará un conjunto de datos provenientes de un estudio de campo realizado en los canales de distribución de productos electrónicos domésticos canadienses en 1989 para el cual se envió un cuestionario a los comerciantes detallistas. De esta forma, la población inicial estaba constituida por todos los comercios de productos electrónicos domésticos de Canadá. En total rebasaban los 450 puntos de venta que se repartían en comercios independientes, afiliados o pertenecientes a una red corporativa y franquiciados. En total, se consiguieron 99 respuestas válidas (43 provenientes de comercios independientes, 35 de afiliados y 21 de franquiciados).

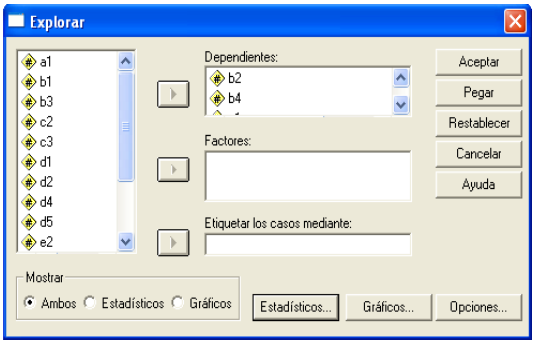
El objetivo perseguido con el cuestionario era la identificación de las variables que podrían regir las relaciones entre el mayorista y el detallista y que podrían conducir a la satisfacción de dichas relaciones y de los resultados financieros alcanzados por ambos. Para ello, los encuestados debían responder a un total de 22 cuestiones cada una de las cuales constituía una variable observable. Con el fin de simplificar el análisis en este apartado se escogen ocho de esas variables: B2, B4, C1, D3, D6, E1, F1 y F3. La estructura del cuestionario así como las preguntas concretas incluidas en el mismo aparecen recogidas en el anexo 1.

A continuación, se realiza el estudio del supuesto de normalidad para las variables implicadas en dicho estudio utilizando los programas *SPSS*, *LISREL* y *AMOS*.

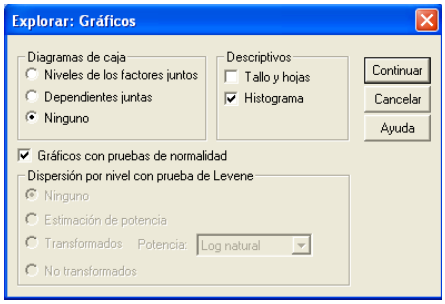
La versión 12 del paquete estadístico *SPSS* sólo permite estudiar la normalidad univariante, para lo cual hay que ir dentro del menú *Analizar* a *Estadísticos descriptivos* y después a *Explorar*:



Aparece así una ventana emergente en la que hemos de pasar al cuadro titulado *Dependientes* aquellas variables cuya normalidad queremos analizar. Asimismo, vamos a seleccionar en el cuadro *Mostrar* la opción *Ambos* y al pulsar en el botón *Gráficos…*



Aparece otra ventana emergente en la que señalaremos las opciones *Ninguno* dentro de *Diagramas de Caja*, *Histograma* dentro de *Descriptivos*, y marcaremos la opción *Gráficos con pruebas de normalidad*. Pulsamos *Continuar* y volvemos a la ventana anterior en la que haremos clic en el botón *Aceptar*.



Dentro de los resultados del análisis realizado, destacaremos, en primer lugar, una tabla con una serie de estadísticos descriptivos calculados para cada una de las variables seleccionadas, entre los que se encuentran los coeficientes de asimetría G_1 y curtosis G_2 y sus respectivos errores típicos:

Descriptivos

			b2	b4	c1	d3	d6	e1	f1	f3
Media		Estadístico	76,47	64,78	67,14	61,67	61,38	64,56	69,24	69,76
		Error típico	1,577	1,889	2,071	1,930	1,960	2,193	1,586	1,608
Intervalo de confianza para la media al 95%	LI	Estadístico	73,34	61,03	63,03	57,84	57,49	60,20	66,09	66,57
	LS	Estadístico	79,61	68,53	71,25	65,50	65,27	68,91	72,39	72,95
Media recortada al 5%		Estadístico	77,42	65,38	67,77	62,42	61,77	65,20	69,85	70,38
Mediana		Estadístico	80,00	70,00	70,00	60,00	65,00	65,00	70,00	70,00
Varianza		Estadístico	246,354	353,277	424,817	368,816	380,321	476,188	249,145	255,981
Desv. típ.		Estadístico	15,696	18,796	20,611	19,205	19,502	21,822	15,784	15,999
Mínimo		Estadístico	30	0	0	0	10	10	10	20
Máximo		Estadístico	100	100	100	100	100	100	100	99
Rango		Estadístico	70	100	100	100	90	90	90	79
Amplitud intercuartil		Estadístico	15	30	30	25	25	35	20	20
Asimetría		Estadístico	-,837	-,602	-,490	-,530	-,255	-,370	-,794	-,749
		Error típico	,243	,243	,243	,243	,243	,243	,243	,243
Curtosis		Estadístico	,790	,790	-,068	,949	-,248	-,501	1,416	,514
		Error típico	,481	,481	,481	,481	,481	,481	,481	,481

Dividiendo cada uno de los coeficientes entre su respectivo error típico hemos calculado los estadísticos $z(G_1)$ y $z(G_2)$, y sumando los cuadrados de estos últimos hallamos el valor experimental del estadístico de contraste conjunto k^2, siendo los resultados obtenidos los siguientes:

CONTRASTE	b2	b4	c1	d3	d6	e1	f1	f3
Asimetría: $z(G_1)$	-3,444	-2,477	-2,016	-2,181	-1,049	-1,523	-3,267	-3,082
Curtosis: $z(G_2)$	1,642	1,642	-0,141	1,973	-0,516	-1,042	2,944	1,069
Conjunto: k^2	14,562	8,835	4,086	8,650	1,367	3,403	19,343	10,643

Como se puede observar, de acuerdo con los criterios especificados con anterioridad para un nivel de significación del 5%, la hipótesis de simetría se rechaza para todas las

variables excepto para D6 y E1; en cambio, la hipótesis de distribución mesocúrtica se rechaza sólo para D3 y F1. Por su parte, el contraste conjunto de asimetría y curtosis, indica que sólo pueden considerarse como normales las variables C1, D6 y E1.

La siguiente tabla que se obtiene es la que contiene los resultados de los contrastes de normalidad de Kolmogorov-Smirnov-Lilliefors y de Shapiro-Wilks:

Pruebas de normalidad

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Estadístico	gl	Sig.	Estadístico	gl	Sig.
b2	,149	99	,000	,935	99	,000
b4	,135	99	,000	,964	99	,008
c1	,121	99	,001	,965	99	,009
d3	,100	99	,016	,967	99	,014
d6	,095	99	,028	,981	99	,164
e1	,104	99	,010	,966	99	,012
f1	,176	99	,000	,950	99	,001
f3	,193	99	,000	,942	99	,000

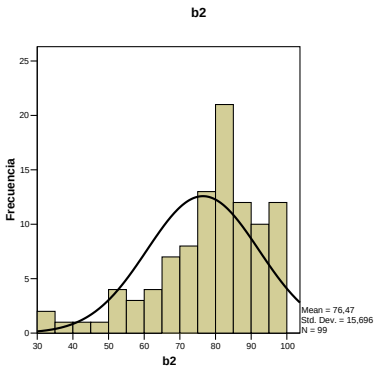
[a] Corrección de la significación de Lilliefors

Si nos fijamos en la significación de los estadísticos de Kolmogorov-Smirnov-Lilliefors, podemos observar que en todos los casos son inferiores a 0'05, por lo que se rechazará[2] la hipótesis nula de normalidad de todas las variables para dicho nivel de significación. De acuerdo con ese mismo criterio, el contraste de Shapiro-Wilks nos permite rechazar la hipótesis de normalidad para todas las variables excepto para D6, si bien es cierto que, como ya señalamos, este contraste no sería el más indicado en este caso dado que el tamaño de muestra es demasiado grande (99 individuos). No obstante, si consideramos una significación del 1%, ambos contrastes dan lugar a que rechacemos la normalidad de todas las variables excepto D3, D6 y E1.

Los últimos resultados que se obtienen del análisis son los histogramas y los gráficos de probabilidad normal[3] de cada una de las variables seleccionadas, los cuales se recogen a continuación:

[2] Recordar que la hipótesis nula de un contraste se rechazará siempre que la significación o p-valor asociado al estadístico sea inferior al nivel de significación α, habitualmente fijado en el 5% o en el 1%.

[3] Concretamente, los gráficos que genera *SPSS* con este comando son una tipo especial que se denominan "gráficos cuantil-cuantil" o simplemente "gráficos q-q" (en inglés, *quantile-quantile plots* o *q-q plots*) cuya interpretación es similar a la comentada para los gráficos de probabilidad normal.


b2
Frecuencia
Mean = 76,47
Std. Dev. = 15,696
N = 99
b2

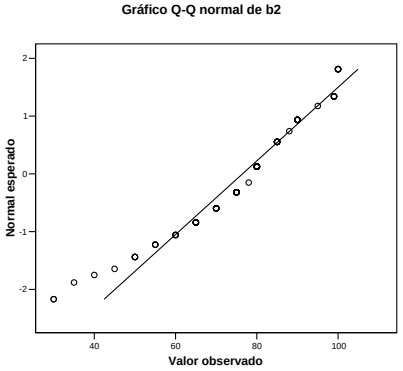
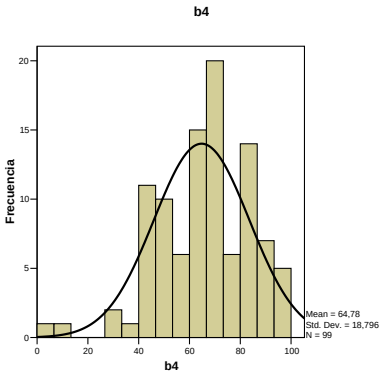


Gráfico Q-Q normal de b2
Normal esperado
Valor observado


b4
Frecuencia
Mean = 64,78
Std. Dev. = 18,796
N = 99
b4

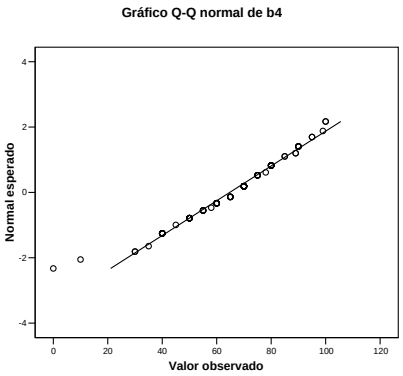
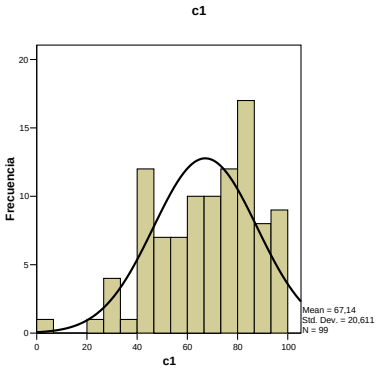


Gráfico Q-Q normal de b4
Normal esperado
Valor observado


c1
Frecuencia
Mean = 67,14
Std. Dev. = 20,611
N = 99
c1

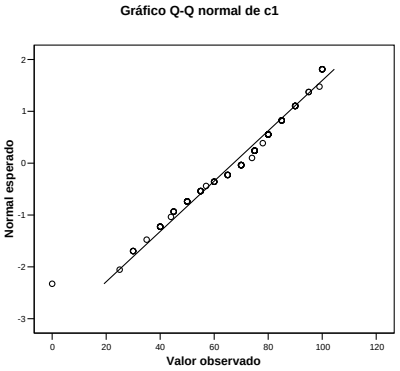
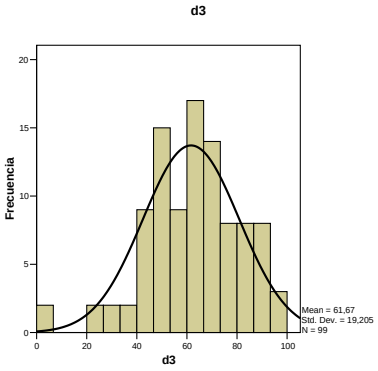


Gráfico Q-Q normal de c1
Normal esperado
Valor observado


d3
Frecuencia
Mean = 61,67
Std. Dev. = 19,205
N = 99
d3

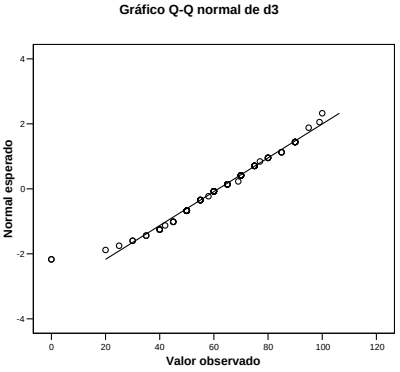
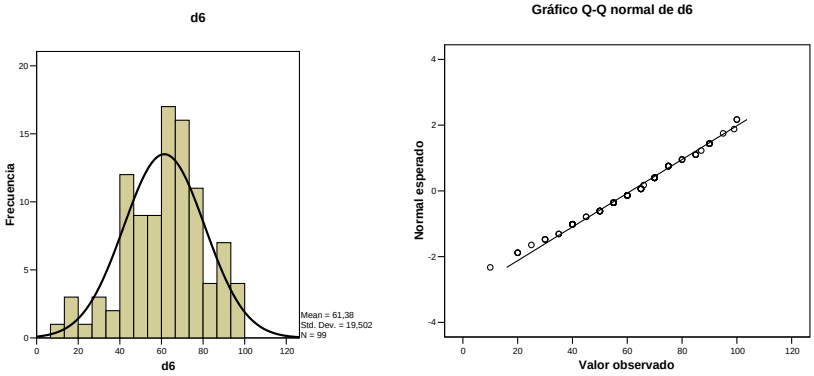
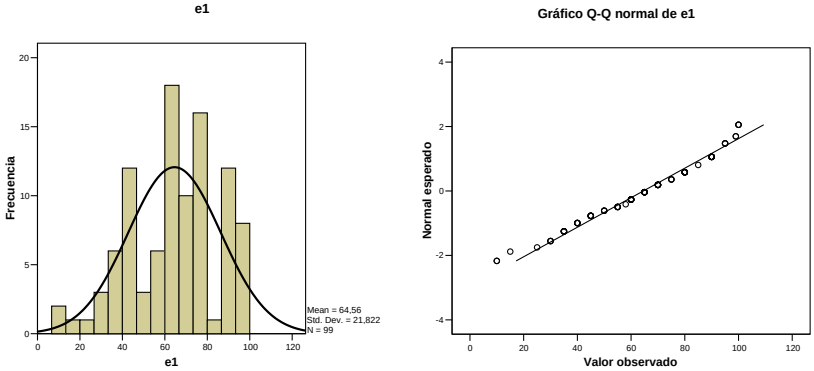
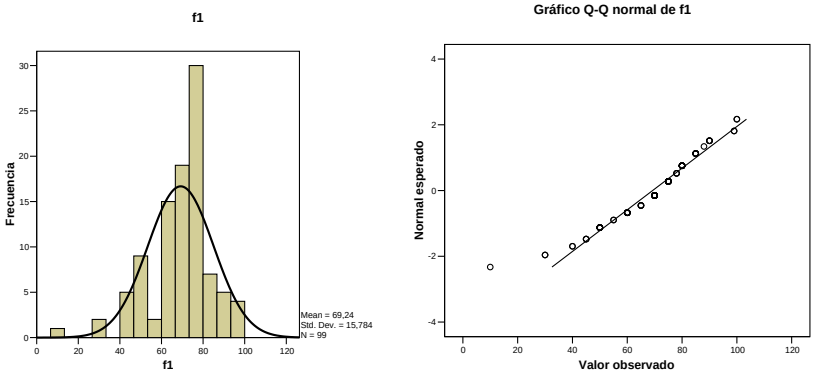
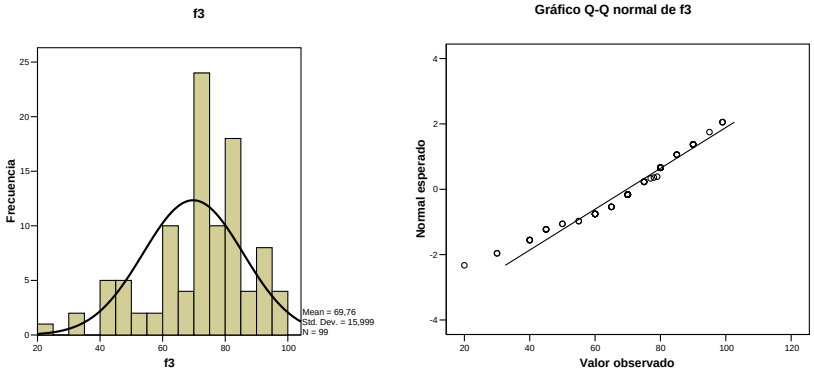


Gráfico Q-Q normal de d3
Normal esperado
Valor observado





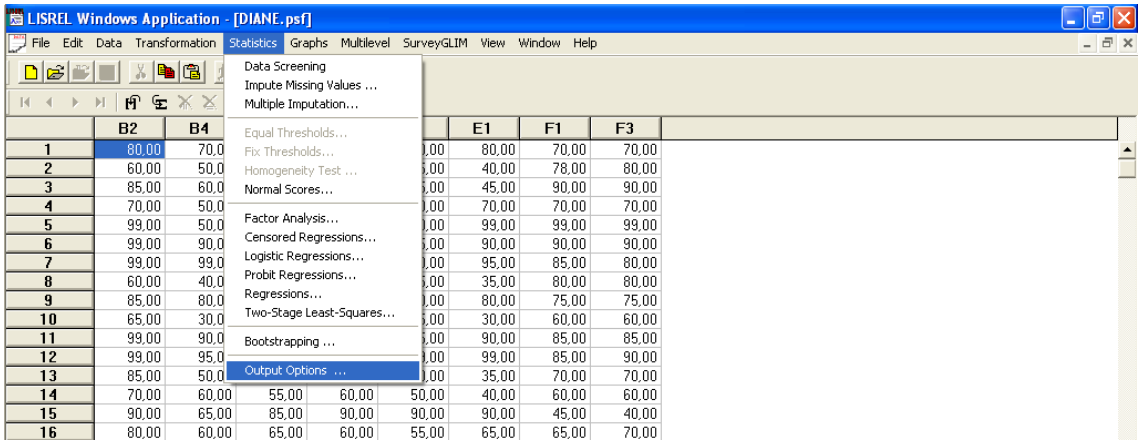




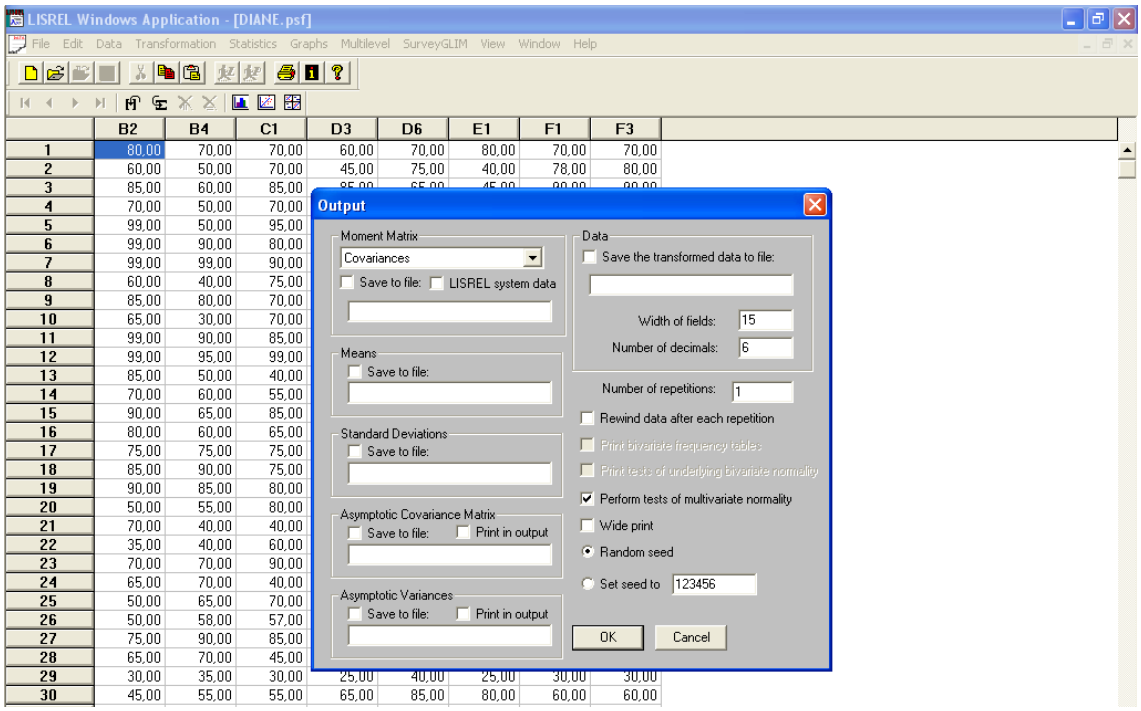
A la vista de los histogramas no parece claro que alguna de las variables tenga una distribución muy semejante a la normal. En cambio, en los gráficos q-q de las variables D3, D6 y E1 parece que el ajuste a la recta es bastante bueno por lo que podrían ser

normales. En cierto modo, esta interpretación resulta congruente con la obtenida de los contrastes Kolmogorov-Smirnov-Lilliefors y Shapiro-Wilks antes comentados si fijamos como nivel de significación el 1%.

Para contrastar la normalidad de los datos con el módulo *PRELIS* del programa *LISREL* (versión 8.7), hay que ir a *Output Options* dentro del menú *Statistics*.



y en la ventana emergente que aparece marcamos la opción *Perform tests of multivariate normality*.



Pulsamos *OK* y el programa genera una ventana con los resultados siguientes:

– Una primera tabla con medidas estadísticas de síntesis univariantes para cada una de las variables continuas entre las que se encuentran los coeficientes de asimetría (*skewness*) G_1 y curtosis (*kurtosis*) G_2:

Total Sample Size = 99

Univariate Summary Statistics for Continuous Variables

Variable	Mean	St. Dev.	T-Value	Skewness	Kurtosis	Minimum	Freq.	Maximum	Freq.
B2	76.475	15.696	48.479	-0.837	0.790	30.000	2	100.000	6
B4	64.778	18.796	34.291	-0.602	0.790	0.000	1	100.000	2
C1	67.141	20.611	32.412	-0.490	-0.068	0.000	1	100.000	6
D3	61.667	19.205	31.949	-0.530	0.949	0.000	2	100.000	1
D6	61.384	19.502	31.318	-0.255	-0.248	10.000	1	100.000	2
E1	64.556	21.822	29.435	-0.370	-0.501	10.000	2	100.000	3
F1	69.242	15.784	43.648	-0.794	1.416	10.000	1	100.000	2
F3	69.758	15.999	43.382	-0.749	0.514	20.000	1	99.000	3

– La segunda tabla muestra el contraste univariante de normalidad para cada una de las variables consideradas, que incluye los respectivos valores experimentales de los estadísticos de contraste de asimetría $z(G_1)$ y curtosis $z(G_2)$ univariantes (*z-score*) y sus p-valores asociados, así como el valor experimental del estadístico k^2 de contraste conjunto de simetría y curtosis univariantes (*chi-square*) y su correspondiente p-valor[4]:

Test of Univariate Normality for Continuous Variables

	Skewness		Kurtosis		Skewness and Kurtosis	
Variable	Z-Score	P-Value	Z-Score	P-Value	Chi-Square	P-Value
B2	-3.200	0.001	1.515	0.130	12.533	0.002
B4	-2.412	0.016	1.514	0.130	8.111	0.017
C1	-2.000	0.045	0.034	0.973	4.003	0.135
D3	-2.149	0.032	1.712	0.087	7.551	0.023
D6	-1.072	0.284	-0.418	0.676	1.324	0.516
E1	-1.537	0.124	-1.210	0.226	3.825	0.148
F1	-3.063	0.002	2.212	0.027	14.278	0.001
F3	-2.914	0.004	1.126	0.260	9.760	0.008

Como se puede observar, para un nivel de significación del 5%, la hipótesis de simetría se rechaza para todas las variables salvo para D6 y E1, la de mesocurtosis se rechaza sólo para F1, y según el contraste conjunto de asimetría y curtosis, se rechaza la normalidad de todas las variables salvo C1, D6 y E1.

– La tercera tabla recoge el contraste multivariante de normalidad, que consta de los coeficientes (*value*) de asimetría $G_{1,p}$ y curtosis $G_{2,p}$ multivariantes, los respectivos valores experimentales (*z-score*) de los estadísticos de contraste $z(G_{1,p})$ y $z(G_{2,p})$ y

[4] Si bien los coeficientes de asimetría G_1 y de curtosis G_2 calculados por *LISREL* coinciden con los obtenidos por *SPSS*, sin embargo los valores experimentales de los estadísticos de contraste $z(G_1)$, $z(G_2)$ y k^2 no coinciden con los que obtuvimos anteriormente porque para su cálculo *LISREL* sigue los procedimientos recomendados por D'AGOSTINO (1986) a las que ya hicimos referencia. No obstante, los resultados obtenidos así como las conclusiones a las que se llega son prácticamente iguales.

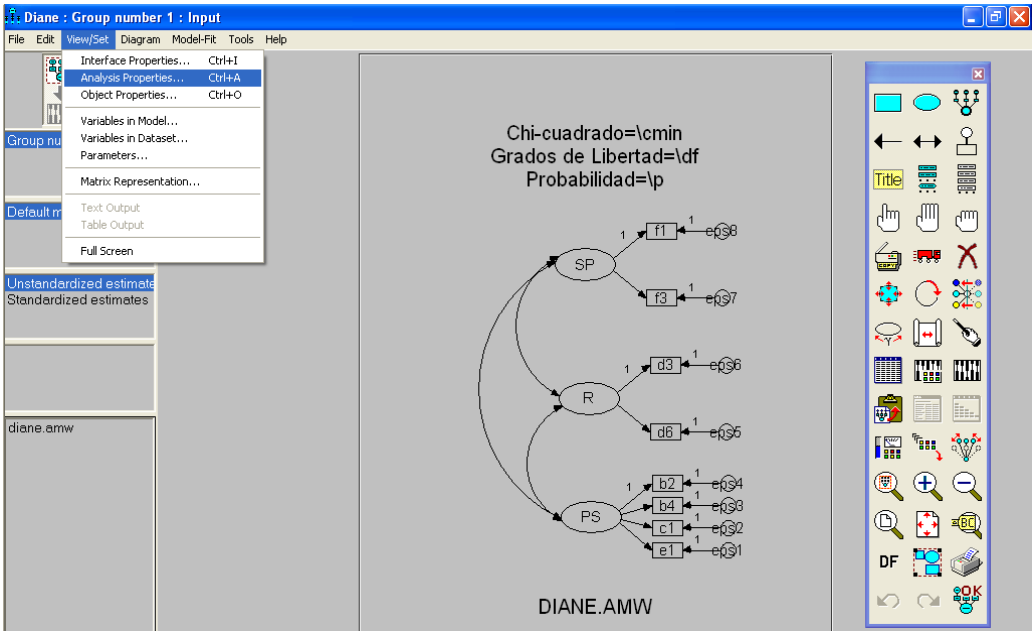
sus p-valores asociados, así como el estadístico k_p^2 de contraste conjunto de simetría y curtosis multivariante y su correspondiente p-valor:

```
Test of Multivariate Normality for Continuous Variables

            Skewness                  Kurtosis          Skewness and Kurtosis

     Value  Z-Score P-Value     Value  Z-Score P-Value      Chi-Square P-Value
    ------  ------- -------   -------  ------- -------      ---------- -------
    35.497   16.223   0.000   130.514    8.449   0.000         334.576   0.000
```

En este caso, los contrastes de asimetría y curtosis multivariantes considerados tanto por separado como conjuntamente permiten rechazar la hipótesis nula de distribución normal multivariante para cualquier nivel de significación puesto que todos los p-valores asociados a los estadísticos son nulos.

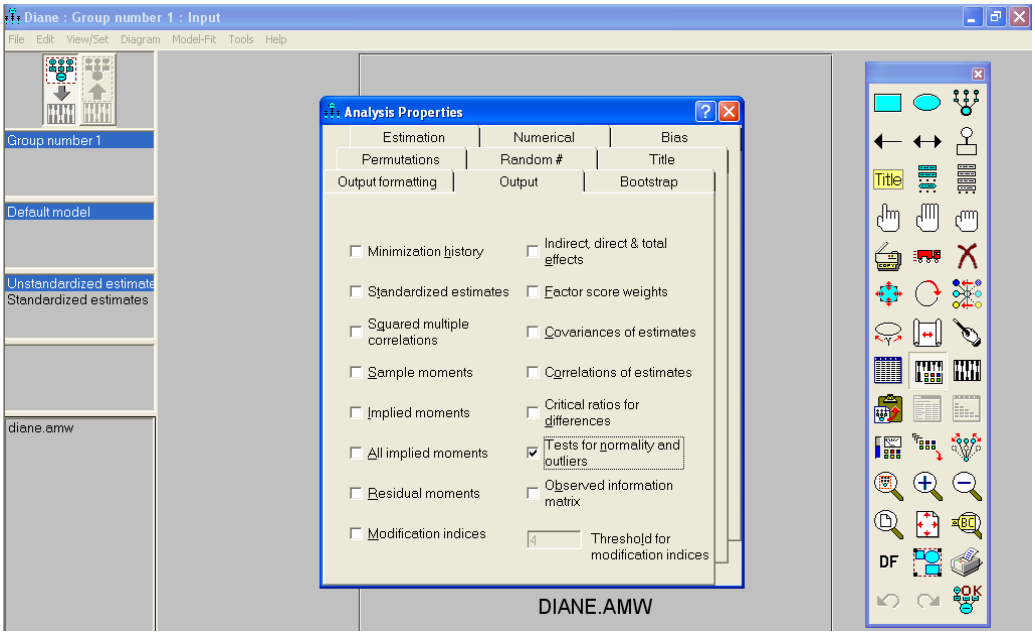
– Finalmente, se muestran los respectivos histogramas de frecuencias de cada una de las variables, la matriz de varianzas-covarianzas, y las matrices de medias y de desviaciones típicas[5].

Para contrastar la normalidad multivariante de los datos con programa AMOS (versión 4.0), hay que ir a *Analysis Properties* dentro del menú *View/Set*, o bien, hacemos clic en el icono (*Analysis Properties*) de la barra de herramientas:



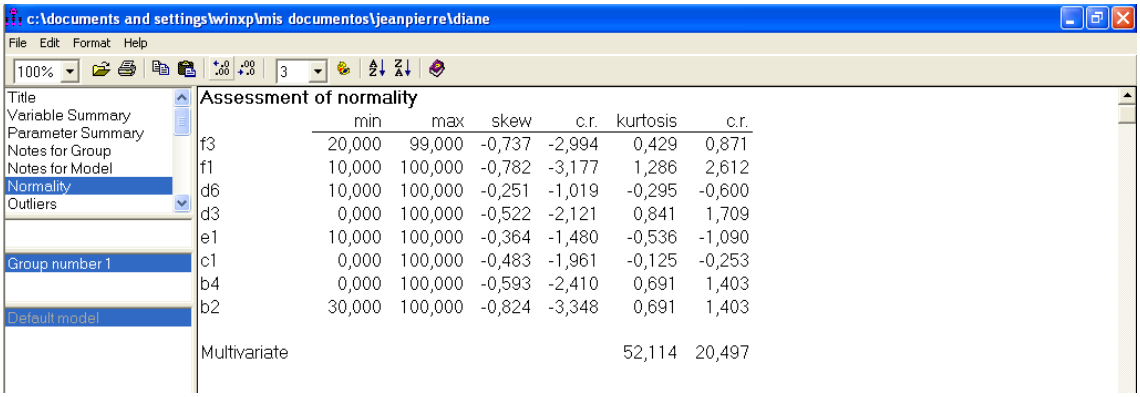
[5] Omitimos esta parte de los resultados porque esta información no resulta necesaria para el estudio de la normalidad de los datos.

Aparece una ventana emergente, y en la pestaña *Output* marcamos la opción *Tests for normality and outliers*.



Cerramos esa ventana y hacemos clic en el icono (*Calculate estimates*) de la barra de herramientas para ejecutar el comando.

Para ver los resultados, podemos ir a *Table Output* dentro del menú *View/Set*, o bien, pulsar en el icono (*View spreadsheets*) de la barra de herramientas. Aparece una ventana con los resultados, y seleccionamos el campo titulado *Normality* dentro de la lista que aparece en la parte superior izquierda de la ventana, mostrándose así en la parte derecha de la ventana una tabla con los contrastes de normalidad correspondientes.



	min	max	skew	c.r.	kurtosis	c.r.
f3	20,000	99,000	-0,737	-2,994	0,429	0,871
f1	10,000	100,000	-0,782	-3,177	1,286	2,612
d6	10,000	100,000	-0,251	-1,019	-0,295	-0,600
d3	0,000	100,000	-0,522	-2,121	0,841	1,709
e1	10,000	100,000	-0,364	-1,480	-0,536	-1,090
c1	0,000	100,000	-0,483	-1,961	-0,125	-0,253
b4	0,000	100,000	-0,593	-2,410	0,691	1,403
b2	30,000	100,000	-0,824	-3,348	0,691	1,403
Multivariate					52,114	20,497

Esta tabla presenta, para cada una de las variables consideradas los valores mínimo y máximo, los coeficientes de asimetría y curtosis –que, en este caso no son G_1 y G_2, sino g_1 y g_2– los respectivos valores experimentales (*critical ratio* o *c.r.*) de los estadísticos de

contraste $z(g_1)$ y $z(g_2)$ asociados que, admitiendo la normalidad, corresponderían a una normal estándar. Así, los resultados indican que, dado un nivel de significación del 5%, todas las variables excepto D6 y E1 tienen una asimetría significativa, y que sólo la variable F1 se aleja significativamente de la normal en lo que a la curtosis se refiere (conclusiones que coinciden con las obtenidas con *LISREL*).

En cuanto a la normalidad multivariante, *AMOS* proporciona sólo contraste para la curtosis multivariante[6], cuya estimación y valor experimental se muestran al final de la tabla anterior. La interpretación de este valor es la misma que en el caso univariante por lo que podemos concluir que conjuntamente las variables presentan una curtosis significativamente distinta de la de una normal multivariante.

Por último, hemos de señalar que la diferencia en el estadístico de curtosis multivariante y en su valor experimental según sea obtenido con *LISREL* o con *AMOS* se debe a que cada uno de estos programas utiliza un estadístico diferente para realizar el contraste de hipótesis[7]. No obstante, las conclusiones a que se llegan con uno y otro procedimiento suelen ser muy similares (como se puede comprobar, en nuestro ejemplo son iguales).

3. Linealidad

Este supuesto se refiere a que las relaciones entre distintas variables sean lineales. El método más comúnmente utilizado a la hora de examinar la estructura de las relaciones entre distintas variables es el gráfico de dispersión el cual representa los valores para cada dos variables. En dicho gráfico, cada variable se representa en un eje y el patrón seguido por los puntos representa la relación entre dichas variables, de tal forma que si los puntos siguen una línea recta, la combinación de las dos variables es lineal. Cuando los puntos siguen una línea curva, representan una relación no lineal y cuando no siguen

---

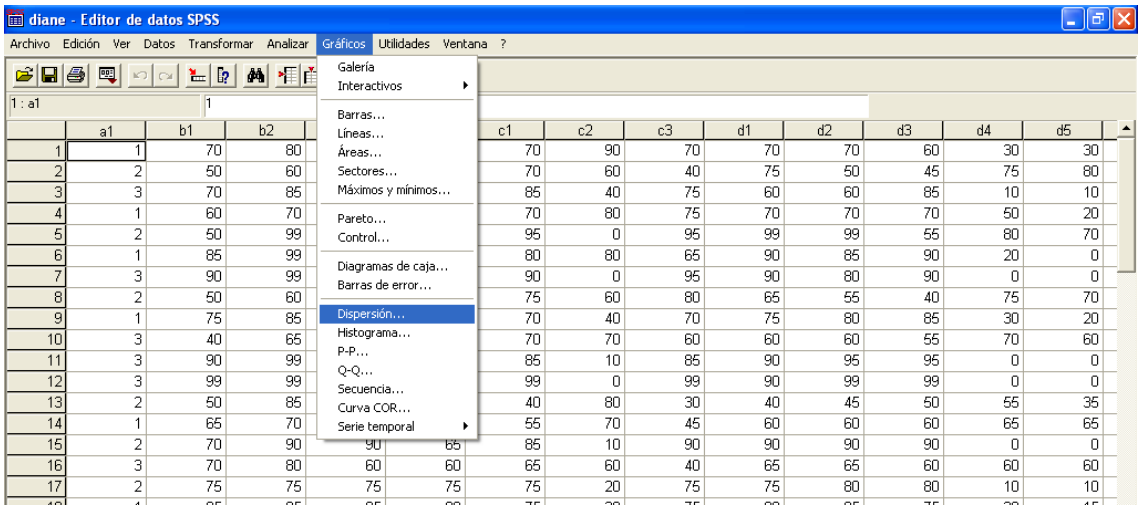
[6] El que *AMOS* sólo permita contrastar la curtosis multivariante y no la asimetría puede deberse al mayor efecto en la validez de los resultados que tiene un significativo exceso o defecto de curtosis de la distribución conjunta de las variables observadas.

En este sentido, BOLLEN (1989; 416) señala que si la distribución no es normal pero es mesocúrtica, las propiedades de los estimadores máximo-verosímiles y de mínimos cuadrados generalizados son las mismas que si se cumpliese la hipótesis de normalidad. Sin embargo, si la distribución presenta una curtosis significativamente distinta de la normal, queda garantizada la consistencia de los estimadores pero no su eficiencia asintótica, ni serían adecuadas las matrices de covarianzas para los test de significación individual de los parámetros, ni se podrían aplicar los tests χ^2 de ajuste global del modelo puesto que los estadísticos de contraste no seguirían asintóticamente esta distribución

[7] En BOLLEN (1989; 424) se pueden encontrar sendas listas con la expresión de los estadísticos de contraste.

ninguna estructura aparente, se pone de manifiesto la no existencia de relación alguna entre las dos variables.

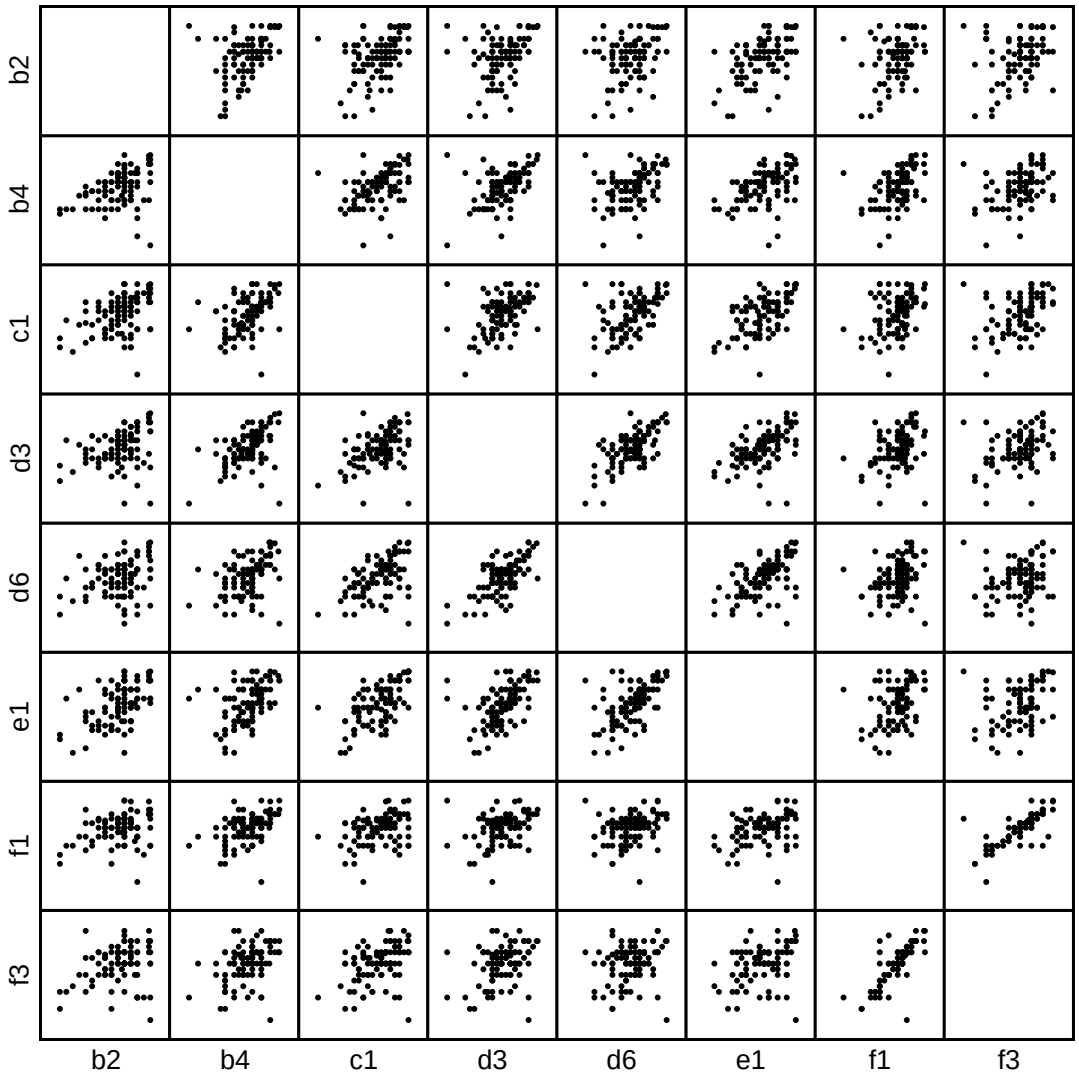
Cuando el investigador se enfrenta a muchas variables, la forma más eficiente de comprobar el supuesto de linealidad es utilizar la opción matricial del gráfico de dispersión en la cual se representan los gráficos de dispersión correspondientes a todas las posibles combinaciones de variables. Para ello, en *SPSS*, se debe ir a *Dispersión* dentro del menú de *Gráficos* y elegir la opción matricial:



Aparece una ventana emergente donde se debe elegir las variables sobre las que se quiere realizar el gráfico. En nuestro caso, se eligen todas aquellas implicadas en el estudio.

Se pulsa en *Aceptar* y el programa estadístico proporciona el gráfico de dispersión en forma matricial, tal y como se muestra a continuación. Como se puede observar, los gráficos de dispersión parecen indicar que existe cierto grado de linealidad en las relaciones existentes entre las variables, ya que los puntos, en cada uno de ellos, están

más o menos situados alrededor de una recta, si bien algunos casos, como F1-F3, parecen mucho más claros que otros, por ejemplo B2-F3.



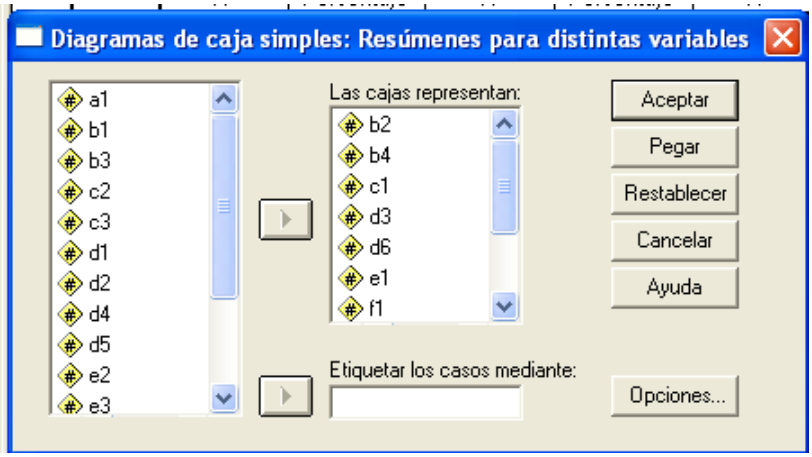
4. Valores atípicos

Los valores atípicos (en inglés "outliers") son individuos que presentan un valor o combinación de valores en la(s) variable(s) observada(s) que les diferencia claramente del grueso de las observaciones. Estos valores pueden aparecer por diversas razones, como errores de procesamiento y/o codificación de los datos, como consecuencia de una situación extraordinaria, o pueden deberse a causas desconocidas. Dado que ciertos valores atípicos –que se denominan observaciones influyentes– pueden provocar una importante distorsión en los resultados de los análisis, se hace necesario examinar los datos para detectar su presencia, estudiar la influencia que ejercen y, en caso de tratarse

de observaciones influyentes, estudiar cuáles son las causas que los originan y decidir en cada caso si se deben retener o excluir del análisis. En este sentido, la detección de los valores atípicos se puede realizar desde una perspectiva univariante o multivariante.

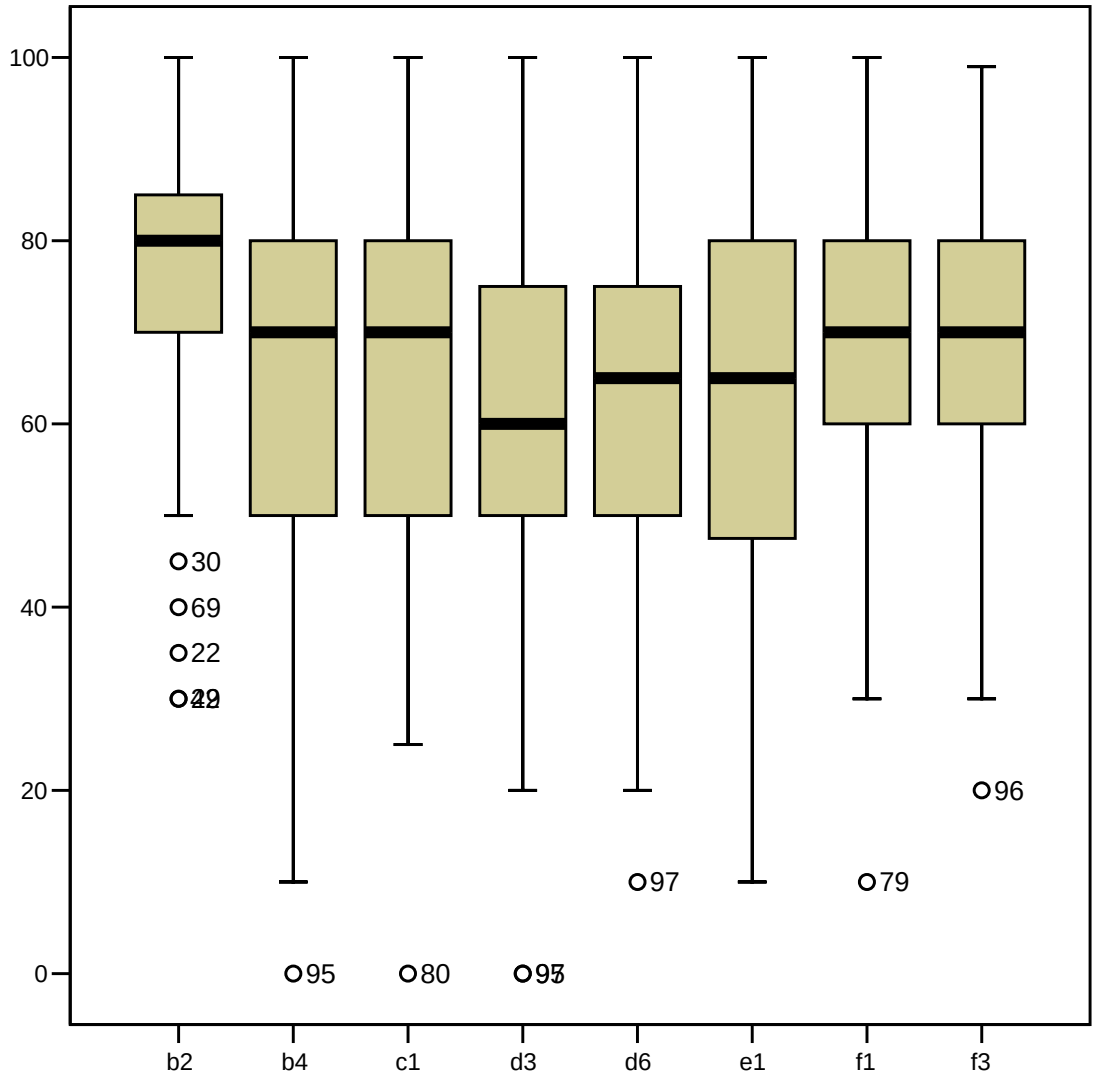
Desde un punto de vista univariante, una primera aproximación se puede realizar utilizando los denominados diagramas de caja. Para obtener estas gráficos utilizando el programa *SPSS*, hemos de ir a *Diagramas de caja* dentro del menú *Gráficos*; aparece una ventana emergente y seleccionamos la opción *simple* e indicamos que los datos del gráfico son *resúmenes para distintas variables:*

Vamos a *definir* y aparece una nueva ventana; seleccionamos como variables a representar aquéllas que son objeto de estudio, y si tuviésemos alguna variable de identificación de los individuos (en este ejemplo no existe) la incorporaríamos al campo *Etiquetar los casos mediante*; pulsamos aceptar y obtendremos los resultados.



En primer lugar, se obtiene una tabla resumen con el número (y porcentaje) de casos validos, perdidos y el total para cada una de las variables y, a continuación tendremos la siguiente representación gráfica con los distintos diagramas de caja de cada una de las

variables que aparecen identificadas en el eje de abscisas (en ordenadas se sitúan los valores de la variable).



Los valores atípicos de cada variable aparecen representados como pequeños círculos, si la distancia a la que están por debajo del borde inferior de la caja –que es el primer cuartil (Q_1)– o por encima borde inferior de la caja –que es el tercer cuartil (Q_3)– es entre 1'5 y 3 veces la longitud de la caja, es decir, el rango intercuartílico ($Q_3 - Q_1$); o como asteriscos, si esa distancia es más del triple de dicho rango. Tanto unos como otros aparecen etiquetados con el número de caso –o con el valor de variable de identificación que hayamos seleccionado–.

En el ejemplo con que estamos trabajando, observamos que son atípicas las observaciones 42, 29, 22, 69 y 30 en la variable B2, la observación 95 en B4, la 80 en la variable C1, las

observaciones 95 y 97 en D3, y las observaciones 97, 79 y 96 en las variables D6, F1 y F3, respectivamente. Asimismo, comprobamos que la variable E1 no presenta ningún valor atípico.

Un método complementario consiste en tipificar todas las variables de manera que todas tengan media 0 y desviación típica 1. Así, se considerarían valores atípicos aquellos que, una vez tipificados, son en valor absoluto superiores a 2'5 –si la muestra es menor de 80– o a un valor entre 3 y 4 –para muestras mayores.

Aplicando este método a nuestro ejemplo, encontraríamos los valores atípicos que aparecen en la tabla siguiente (que no aparece completa), señalados en rojo los que son superiores a 2'5, y resaltados en amarillo los que son mayores que 3).

individuo	B2	B4	C1	D3	D6	E1	F1	F3
1	0,2257	0,2793	0,1394	0,0872	0,4441	0,7114	0,0482	0,0152
2	1,0550	0,7902	0,1394	0,8723	0,7018	1,1310	0,5577	0,6434
…	…	…	…	…	…	…	…	…
21	0,4146	1,3250	1,3235	1,6573	1,1021	1,3613	2,4988	2,4976
22	**2,6559**	1,3250	0,3482	0,4361	0,0713	0,2508	1,2253	1,5553
23	0,4146	0,2793	1,1147	0,4361	0,0713	1,1719	1,0034	0,9575
24	0,7348	0,2793	1,3235	0,6106	0,4441	0,2098	0,3666	0,3293
25	1,6953	0,0119	0,1394	0,0872	0,4441	0,9007	0,5885	0,6130
26	1,6953	0,3624	0,4945	0,3489	1,3598	0,4401	0,3666	0,3293
27	0,0944	1,3487	0,8709	0,4361	1,3202	0,2098	0,3666	0,6434
28	0,7348	0,2793	1,0797	0,6978	0,2379	0,4811	0,0482	0,2989
29	**2,9761**	1,5924	1,8112	1,9190	1,1021	1,8219	2,4988	2,4976
30	2,0155	0,5229	0,5921	0,1745	1,2171	0,7114	0,5885	0,6130
…	…	…	…	…	…	…	…	…
41	1,3752	1,3250	1,3235	0,6106	0,3290	1,3613	0,9069	0,9271
42	**2,9761**	1,3250	1,3235	1,0293	2,1329	1,5916	1,8621	1,5553
43	0,0944	0,7071	0,3345	0,3838	0,7018	0,2508	0,6850	0,5806
…	…	…	…	…	…	…	…	…
68	0,4146	0,8140	0,6270	0,3489	0,4441	0,4811	0,0482	0,2989
69	2,3357	1,3250	2,0550	1,3956	1,3598	**2,5128**	1,2253	1,2412
70	1,5065	1,6161	1,3585	1,1339	0,4441	0,0205	1,0034	0,6434
…	…	…	…	…	…	…	…	…
78	0,4146	0,2555	0,8359	0,6106	1,1021	1,1310	0,0482	0,6130
79	0,8661	0,8140	0,8359	0,6106	0,4441	0,7114	**3,7723**	1,8694
80	0,8661	0,8140	**3,2741**	2,1807	2,1329	0,2098	0,5885	1,8694
81	0,8661	0,8140	0,6270	1,4828	0,4441	0,2508	0,0482	0,0152
…	…	…	…	…	…	…	…	…
91	0,2257	0,5466	1,6023	0,1745	1,1021	1,6325	0,5885	1,5857
92	0,8661	**2,9292**	0,6270	0,0872	0,4441	0,7114	0,5885	0,0152
93	1,5065	1,8835	0,8359	2,0062	0,0713	1,1719	1,3218	1,2716
94	0,2257	0,7902	1,8112	1,6573	2,1329	**2,5128**	0,5885	0,0152
95	1,5065	**3,4640**	0,8359	**3,2274**	1,6175	0,2508	1,2253	1,8694
96	1,5065	1,3487	1,6023	1,4828	1,9902	1,6325	0,6850	**3,1258**
97	0,2257	1,8835	1,6023	**3,2274**	**2,6482**	1,1719	1,9585	0,6434
98	1,0550	0,2555	0,6270	0,4361	0,4441	1,1719	0,0482	0,0152
99	0,2257	0,7902	0,1394	0,4361	0,4441	1,1719	0,6850	0,6130

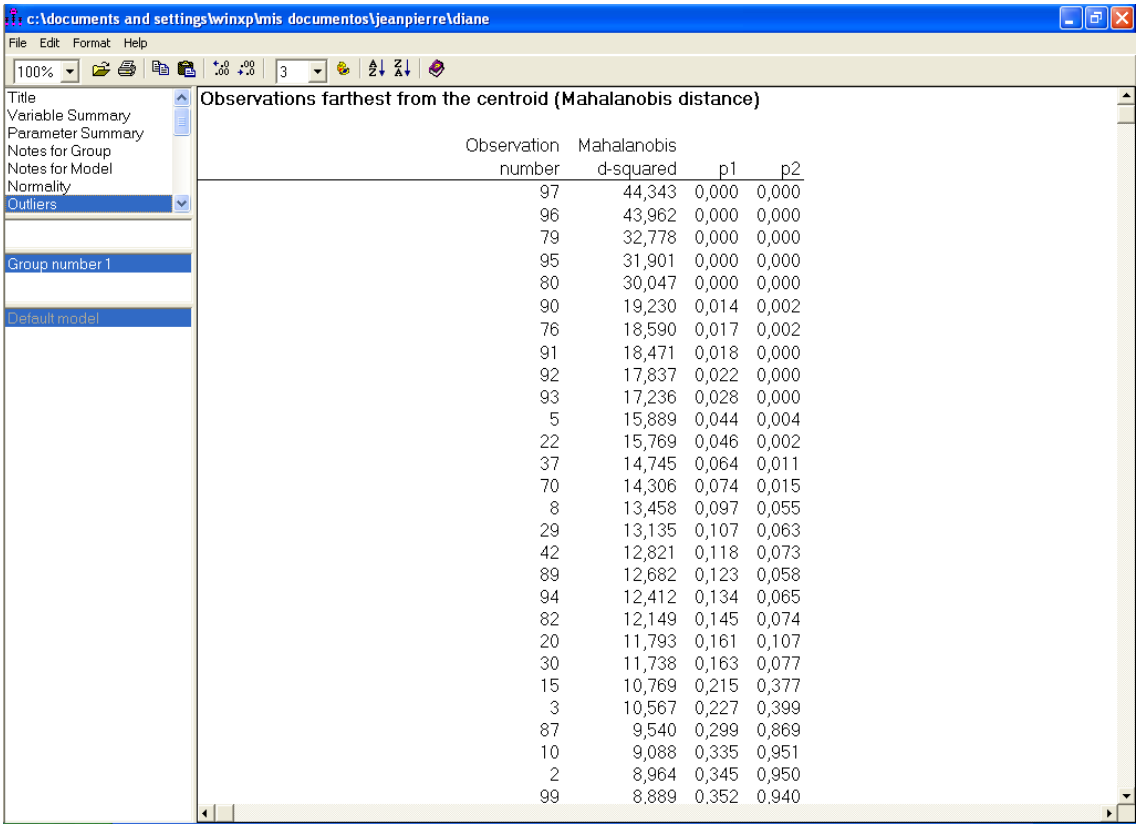
Si comparamos los resultados obtenidos por este método, veremos que no existen muchas diferencias con los logrados utilizando diagramas de caja.

Desde el punto de vista multivariante, la detección de los valores atípicos se realiza utilizando la denominada distancia de Mahalanobis, que es una medida estadística de la

distancia multidimensional de un individuo respecto al centroide o media de las observaciones, que viene dada por la siguiente expresión:

$$d_i^2 = (\mathbf{x}_i - \bar{\mathbf{x}})' \hat{\mathbf{S}}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}})$$

donde $\mathbf{x}_i$ es el vector columna con los valores de todas las variables para la observación *i*-ésima, respectivamente, $\bar{\mathbf{x}}$ es el correspondiente vector columna de medias muestrales y $\hat{\mathbf{S}}^{-1}$ es la inversa de la matriz de varianzas-covarianzas muestral. Esta distancia tiene interesantes propiedades[8] estadísticas que permiten construir pruebas de significación de la excentricidad de las observaciones. Así, ejecutando la opción *Tests for normality and outliers* del programa *AMOS* (a la que ya hicimos referencia cuando hablamos de la normalidad multivariante), y señalando el campo titulado *Outliers* dentro de la lista que aparece en la parte superior izquierda de la ventana de resultados, obtenemos la tabla siguiente:



Observations farthest from the centroid (Mahalanobis distance)

Observation number	Mahalanobis d-squared	p1	p2
97	44,343	0,000	0,000
96	43,962	0,000	0,000
79	32,778	0,000	0,000
95	31,901	0,000	0,000
80	30,047	0,000	0,000
90	19,230	0,014	0,002
76	18,590	0,017	0,002
91	18,471	0,018	0,000
92	17,837	0,022	0,000
93	17,236	0,028	0,000
5	15,889	0,044	0,004
22	15,769	0,046	0,002
37	14,745	0,064	0,011
70	14,306	0,074	0,015
8	13,458	0,097	0,055
29	13,135	0,107	0,063
42	12,821	0,118	0,073
89	12,682	0,123	0,058
94	12,412	0,134	0,065
82	12,149	0,145	0,074
20	11,793	0,161	0,107
30	11,738	0,163	0,077
15	10,769	0,215	0,377
3	10,567	0,227	0,399
87	9,540	0,299	0,869
10	9,088	0,335	0,951
2	8,964	0,345	0,950
99	8,889	0,352	0,940

[8] Cabe destacar que es invariante a cambios de escala, es euclídea, está normalizada y tiene en cuenta las correlaciones entre las variables, es decir, la redundancia entre éstas. Una explicación mas detallada se puede encontrar en CUADRAS (1988; 306-308).

Esta tabla contiene una lista con todas las observaciones ordenadas, de mayor a menor, según su distancia –de Mahalanobis– al centroide. Las dos columnas siguientes contienen sendos p-valores con la significación de esas distancias, asumiendo la existencia de normalidad multivariante, desde dos puntos de vista.

Así, para la observación más alejada, la número 97, la columna p_1 muestra la probabilidad de que una observación cualquiera se encuentre a una distancia de Mahalanobis mayor o igual que 44'343. En cambio, la columna p_2 indica la probabilidad de que la observación más alejada del centroide se encuentre a una distancia de Mahalanobis mayor o igual que 44'343. Para la siguiente observación –la segunda más alejada del centroide, la número 96– p_1 indica la probabilidad de que un individuo cualquiera esté a una distancia mayor o igual que 43'962, mientras que p_2 muestra la probabilidad de que el segundo individuo más lejano del centroide esté a una distancia mayor o igual que 43'962. El resto de p-valores se interpretarían de forma análoga.

Lógicamente, los valores de p_1 aumentan a medida que vamos descendiendo en la lista (nos acercamos más al centroide) mientras que no sucede lo mismo con los valores de p_2. Por otro lado, mientras que, en cierta medida, podemos esperar encontrar algunos valores pequeños de p_1, sin embargo, si los valores de p_2 son también muy pequeños esto indicaría que esas observaciones están improbablemente lejos del centroide bajo la hipótesis de normalidad multivariante. En cualquier caso, HAIR *et al*. (1999; 58) sugieren que "dada la naturaleza de los tests estadísticos, […] se use un nivel muy conservador, quizá 0'001, como valor umbral para la designación como caso atípico".

De acuerdo con este criterio, en nuestro ejemplo, podríamos considerar como atípicas las observaciones número 97, 96, 79, 95 y 80, las cuales ya se habían revelado como *outliers* al aplicar los procedimientos univariantes antes indicados[9].

[9] Para un estudio más detallado de los métodos de detección de valores atípicos y su tratamiento se recomienda la lectura de HAIR *et al*. (1999; 57-61)

5. Glosario

Asimetría: Rasgo característico de la forma de una distribución de frecuencias (o de probabilidad) que se presenta cuando los valores que están a la misma distancia de la media no tienen igual frecuencia (o probabilidad); se puede distinguir entre *asimetría a la derecha o positiva*, cuando los valores bajos de la variable son los más frecuentes o probables, y *asimetría a la izquierda o negativa*, en caso contrario.

Curtosis: Rasgo característico de la forma de una distribución de frecuencias (o de probabilidad) que se refiere al grado de apuntamiento de la misma en comparación con la distribución normal; se puede distinguir entre *leptocurtosis o curtosis positiva*, cuando la distribución es más apuntada y con colas menos gruesas que la normal, *curtosis negativa o platicurtosis*, si es más aplastada y con colas más gruesas que la distribución normal, y *mesocurtosis* si es igual de apuntada que la normal.

Diagrama de cajas: Representación gráfica en la que se muestran los tres cuartiles de una distribución y que permite detectar la existencia de valores extremos.

Diagrama de dispersión: Gráfico bidimensional en el que se representan, en sendos ejes de coordenadas y por medio de puntos, los pares de valores correspondientes a las mediciones de dos variables cuantitativas observadas sobre un conjunto de individuos.

Distancia de Mahalanobis: Medida estadística de la distancia multidimensional de un individuo respecto al centroide o media de las observaciones que se caracteriza por tener en cuenta las correlaciones entre las variables eliminando así las redundancias que puedan existir entre éstas.

Gráfico de probabilidad normal: Representación gráfica que consiste en enfrentar, en un mismo gráfico, los valores observados frente a los teóricos que se obtendrían de una distribución normal, de modo que, si la variable se distribuye según una normal, los puntos se concentrarán en torno a una línea recta.

Histograma: Representación gráfica de la distribución de frecuencias de una variable cuantitativa en la que los datos están agrupados en intervalos.

Outliers o valores atípicos: son individuos que presentan un valor o combinación de valores en la(s) variable(s) observada(s) que les diferencia claramente del grueso de las observaciones.

6. Bibliografía

BOLLEN, K. A. (1989): *Structural equations with latent variables*, New York: Wiley.

CUADRAS AVELLANAS, C. M. (1988): "Distancias estadísticas", *Estadística Española* 119, sep.–dic.1988, pp. 295-378.

D'AGOSTINO, R. B. (1986): "Tests for the normal distribution" en D'AGOSTINO, R. B.; STEPHENS, M. A. (eds): *Good of fit techniques*, Marcel Dekker, New York, pp. 367-419.

HAIR, J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK; W. C. (1999): *Análisis Multivariante*, 5ª ed. Madrid: Prentice-Hall Iberia.

JÖRESKOG, K. G. (1999): *Formulas for skewness and kurtosis*, accesible en http://www.ssicentral.com/lisrel/kurtosis.pdf.

LILLEFORS, H. W. (1967): "On the Kolmogorov-Smirnov test for normality with mean and variance unknown", *Journal of the American Statistical Association*, vol. 64, pp. 387-389.

MARDIA, K. V. (1970) "Measures of multivariate skewness and kurtosis with applications", *Biometrika* 57, pp. 519–530.