Determinación de puntos atípicos en estudios longitudinales, mediante el análisis biplot de datos sin transformaciones previas

Maura Vásquez de Ramírez [1], Guillermo Ramírez [1], Mercedes López de Blanco [2], Purificación Galindo Villardon [3], Virgilio Bosch [4], Jorge Vazquez [5]

RESUMEN. En este trabajo se propone un procedimiento basado en la aplicación de la técnica Biplot a datos sin transformaciones previas, el cual se complementa con un índice que permite detectar puntos atípicos en un estudio longitudinal, y simultáneamente explicar su comportamiento. El método es aplicado al estudio de los valores de triglicéridos séricos en una submuestra del Estudio Longitudinal de Caracas, formada por varones adolescentes entre 12 y 16 años. Los resultados permitieron distinguir individuos cuyos valores de triglicéridos estaban muy por encima, o por debajo, del valor promedio de su grupo a lo largo de casi todos los períodos. Además, fue posible identificar individuos cuyos perfiles no guardan paralelismo con respecto al patrón promedio de su grupo, lográndose también una aproximación a la forma de evolución temporal del fenómeno. *An Venez Nutr 1999; 12(1):10-15.*

Palabras clave: Estudios de crecimiento longitudinal, análisis Biplot, estudios longitudinales, análisis estadístico.

Introducción

La determinación de puntos atípicos respecto de una o varias variables, ha sido objeto de numerosas propuestas en la literatura estadística. En el ámbito del modelo de regresión, se han desarrollado una gran variedad de métodos gráficos y analíticos para estos efectos (1). Algunas aproximaciones al problema, utilizando el Análisis de Componentes Principales (ACP), han sido propuestas por Jollife (2). Menos numerosas son las investigaciones que refieren el problema a estudios longitudinales multivariantes, razón por la cual nos hemos propuesto como objetivo fundamental de este trabajo, ensayar la aplicación de un procedimiento especialmente diseñado para este tipo de datos, basado en el Análisis Biplot (AB) (3), aplicado a datos sin transformaciones previas, complementado con un índice que sirve de ayuda para identificar e interpretar la atipicidad (4).

No existe una definición formal ampliamente aceptada de punto atópico. Algunos autores ofrecen aproximaciones intuitivas al concepto, señalando que estos puntos son observaciones cuyo comportamiento tiende a ser diferente, en algún sentido, al de las restantes (5). En un análisis univariante, por lo tanto, una observación considerada atípica será sinónimo de un valor extremo en la variable bajo estudio. Una extensión de este planteamiento al ámbito p-variante, indicaría que un punto atípico debe aparecer alejado del resto de las observaciones en el espacio p-dimensional. De acuerdo con esta idea, un mecanismo usualmente utilizado para detectar puntos atípicos multivariantes, consiste en calcular la distancia de Mahalanobis de cada observación al centro de gravedad, o equivalentemente, utilizar la razón de varianza generalizada (6). Una limitación que se impone a la efectividad de este procedimiento está asociada con el hecho de que una observación multivariante puede que no sea extrema en ninguna de las variables que la caracterizan, y sin embargo, ser atípica porque no se adecua a la estructura de las correlaciones de los datos restantes. Por ejemplo, si se estudia la concentración de triglicéridos séricos en un niño, un valor de 110 mg/dl es aceptable a los 12 años, y 40 mg/dl también lo es a los 13, pero la combinación (110,40) es virtualmente atípica puesto que combina un valor muy alto con uno muy bajo. Es decir, se viola el patrón general de correlación serial positiva entre esas variables. Cuando el número de variables en un estudio se

1. Escuela de Estadística, FACES, EECA, UCV, Venezuela.
2. Postgrado de Nutrición, USB, Venezuela.
3. Departamento de Estadística y Matemática Aplicadas, Universidad de Salamanca, España.
4. Instituto de Lipidología, Facultad de Medicina, U.C.V., Venezuela.
5. Escuela Luis Razzetti, Facultad de Medicina, U.C.V., Venezuela.

Solicitar copia a: Maura Vásquez de Ramírez, Aptdo. 65680, Caracas. 1066-A.

incrementa, puede dificultarse la interpretación de la clase de observaciones atípicas que rompen la estructura general de las correlaciones entre las variables. En ese caso, los gráficos resultantes del ACP (7) o del AB (8), pueden ser de gran utilidad.

Los estudios longitudinales diseñados para estudiar fenómenos de crecimiento y maduración en niños y adolescentes plantean, entre otros de sus objetivos, la construcción de normas de referencia para uso clínico, lo que usualmente se lleva a cabo a través de la determinación de curvas que identifican patrones de comportamiento promedio o percentilar. En esta situación, la presencia de puntos atípicos que rompen la estructura de las correlaciones seriales, descritos por individuos que pierden su canal inicial y no regresan a él, conocidos como descanalizdos, o por otros cuyos valores pierden solo transitoriamente su canal, designados como recanalizados (9), puede tener efectos desproporcionados sobre la construcción de las curvas. Es por ello, que en este tipo de investigaciones resulta una necesidad obligante el realizar un análisis preliminar para detectar puntos con estas características.

Materiales y métodos

Materiales

Los datos utilizados en esta investigación forman parte de la información recogida por el Estudio Longitudinal del Area Metropolitana de Caracas (10), llevado a cabo entre 1976 y 1982, cuyo diseño corresponde al de un estudio prospectivo mixto e imbricado sobre tres cohortes de niños que inician a los 4, 8 y 12 años respectivamente, con un seguimiento a lo largo de 5 años. El muestreo fue realizado en 2 etapas, en la segunda de las cuales se seleccionaron niños que cumplían con las siguientes características: pertenecer a los estratos sociales I y II de acuerdo con el Método Graffar modificado para Venezuela (11), madre venezolana o latinoamericana, venezolano por nacimiento, producto de un embarazo normal, peso al nacer adecuado para su edad gestacional y estar aparentemente sano. El tamaño de la muestra fue determinado siguiendo recomendaciones del Programa Biológico Internacional (12). Particularmente en esta investigación se estudia la concentración de triglicéridos séericos (TS) sobre la submuestra (n=31) correspondiente a los varones que iniciaron el seguimiento a los 12 años, con mediciones interanuales completas hasta los 16 años. Esta variable fue escogida para el análisis por ser, entre las bioquímicas, la que mayor variabilidad presenta en niños venezolanos para las edades estudiadas (10).

Métodos

El método JK-Biplot es aplicado sobre los valores de TS registrados en los niños participantes en esta investigación, a lo largo de las edades 12 a 16 años. Las mediciones del perfil de evolución temporal para el i-ésimo niño genérico se organizan sobre el vector $I_i = (x_{i,12}, x_{i,13}, x_{i,14}, x_{i,15}, x_{i,16})$, y el valor promedio de TS para ese niño, a lo largo del período se denota mediante $\bar{x}_i$. La técnica Biplot utiliza como herramienta de trabajo un dispositivo gráfico, denominado plano biplot, sobre el cual se representan simultáneamente: los perfiles mediante puntos, y la variable TS, en las distintas edades, mediante vectores denotados por Trigj (j=12,...,16), los cuales son denominados ejes biplot (13). Para efectos comparativos, se define un niño ideal cuyos valores coinciden con los promedios de TS en las diferentes edades, el cual se denomina perfil promedio (Iprom), también se consideran como referencia los percentiles 10, 25, 50, 75 y 90, estimados sobre la muestra completa de varones del ELAMC (Tabla 3). Las representaciones se efectúan sobre un plano cuyo origen ha sido trasladado para hacerlo coincidir con el perfil promedio, éstas aproximan con alta calidad la información contenida en los perfiles originales, en particular, se preserva la distancia euclídea usual (14), y por consiguiente se obtienen buenas aproximaciones para la distancia entre perfiles, para la distancia de un perfil cualquiera respecto del origen (d^2_i), y también para los sumandos de la descomposición:

$$d^2_i = \sum_{j=12}^{16} (x_{ij} - \bar{x}_i)^2 + 5\,\bar{x}_i^2$$

La partición anterior es un agregado de dos componentes, el primero de los cuales refleja esencialmente la variabilidad temporal de un perfil respecto de su nivel promedio, y el segundo describe el efecto debido propiamente al promedio. Esta descomposición, es de interés cuando tiene sentido el cálculo de promedios por fila en una matriz de datos, adquiriendo especial importancia en el caso de observaciones obtenidas al medir una variable sobre los mismos individuos, en diferentes oportunidades. La aproximación a la componente de variabilidad temporal de un perfil, sobre el plano biplot, se obtiene en la forma:

$$\sum_{j=12}^{16} (I_i^t(2)\,(\mathrm{Trig}_j(2) - \mathrm{Trig}(2)))^2 \approx \sum_{j=12}^{16} (x_{ij} - \bar{x}_i)^2$$

siendo $I_i(2)$ el marcador de un punto sobre el plano que aproxima el perfil del niño i, $\mathrm{Trig}_j(2)$ el marcador del vector que identifica a los triglicéridos séricos en la edad j, y Trig(2) la aproximación al promedio de esa variable en el período considerado.

Los elementos básicos para la interpretación de las representaciones sobre el plano biplot están fundamentados en las propiedades siguientes: (a) La cercanía entre dos puntos del plano es un indicador de semejanza entre los perfiles correspondientes, y viceversa en relación con la lejanía; y (b) La proyección ortogonal del perfil de un niño sobre un eje biplot, que identifica los TS a una edad determinada, reproduce el valor de la variable para ese niño, en esa edad. En este sentido, el método constituye una herramienta poderosa para sugerir: agrupaciones de niños con perfiles similares; ordenamientos de niños de acuerdo a su valor en los TS a una edad determinada; y formas en los patrones de evolución temporal de los perfiles. La interpretación se complementa con la ayuda del siguiente índice de variabilidad temporal :

$$\text{Ivariab(i)} = \frac{\sum_{j=12}^{16} (I_i^t{}_{(2)} (\text{Trig}_{j(2)} - \text{Trig}_{(2)})^2}{d^2_{(2)}}$$

El procesamiento de los datos fue llevado a cabo utilizando el programa computacional BIPLOT, diseñado por los autores especialmente para estos efectos.

Resultados

En la Tabla 1 se presentan los valores de TS que describen el perfil de evolución de la variable para cada uno de los participantes en el estudio, para las edades comprendidas entre los 12 y los 16 años, también se muestran los valores promedio, que definen al niño ideal.

Tabla 1
Concentración de triglicéridos séricos (mg/dl) en varones entre 12 y 16 años. ELAMC.

Niño	Trig12	Trig13	Trig14	Trig15	Trig16
I1	90	99	106	121	108
I2	142	83	79	56	78
I3	173	117	70	100	152
I4	58	120	167	171	150
I5	74	72	79	110	129
I6	66	36	63	92	71
I7	43	48	60	71	73
I8	66	93	86	74	66
I9	106	103	92	78	51
I10	92	87	77	77	138
I11	113	92	76	78	85
I12	72	92	129	78	67
I13	164	125	112	125	165
I14	69	78	93	64	86
I15	76	108	115	45	76
I16	102	93	115	91	53
I17	125	91	73	113	118
I18	45	50	53	52	69
I19	27	71	101	81	62
I20	20	40	51	34	68
I21	47	42	45	78	94
I22	130	133	130	111	70
I23	93	50	35	48	90
I24	43	33	41	96	85
I25	115	53	19	91	148
I26	39	43	67	144	90
I27	179	99	100	108	87
I28	51	115	26	58	32
I29	73	38	39	56	50
I30	95	76	146	72	35
I31	50	42	37	37	25
Iprom	85.1	78.1	80.1	84.2	86.2

En la Tabla 2 puede observarse que, la proporción de información captada por el primer plano biplot (95.8%) es determinada en casi su totalidad por el primer eje principal (92.5%), cuyo vector director es proporcional al perfil promedio. Con base en estas consideraciones lo más relevante de las configuraciones sobre el plano (Gráfico 1) consiste en la posibilidad de detectar, hacia el extremo derecho del primer eje principal (Eje 1), niños caracterizados como atípicos debido a que sus niveles de TS son marcadamente superiores a los del perfil promedio en casi todas las edades; y hacia el extremo izquierdo, opuestos al grupo anterior, se ubican niños con valores marcadamente bajos en la variable, también para casi todas las edades. Adicionalmente, el segundo eje principal (Eje 2) puede considerarse como un indicador de formas que contrapone perfiles extremos de recanalización. De esta manera, hacia el extremo superior se encuentran niños que inician el estudio con valores altos de TS, pierden su canal en las edades intermedias, y terminan regresando al canal inicial hacia las etapas finales del estudio; mientras que hacia el extremo inferior se observan niños con un patrón de recanalización invertido respecto del anterior.

Por consiguiente, las representaciones sobre el plano (Gráfico 1) permiten, en líneas generales, describir niños con diferentes patrones en la evolución del perfil de TS, distinguiéndose algunas agrupaciones, cuyas características pueden ser precisadas más o menos claramente:

(i) Niños cuyos perfiles están representados por puntos dentro de la elipse, para quienes los valores de TS están dentro de un rango aceptado como normal, sin cambios marcados en el valor de la variable entre una edad y otra. En particular, sus valores se encuentran entre los percentiles 10 y 90 de TS, estimados para los varones de la muestra completa ELAMC .

(ii) Niños ubicados hacia la extrema derecha del Eje 1: (I13, I4, I3), los cuales presentan valores altos de TS a lo largo de casi todos los períodos. En particular, los niños I13 e I3 presentan en las edades extremas valores de TS bastante mayores que el percentil 90 de su población, y en las edades intermedias toman valores en torno al percentil 90. Por su parte, el primer valor de la serie para el perfil de I4, está cercano al percentil 25, presentando un fuerte incremento en las edades subsiguientes, que ubica sus niveles de TS en torno, o por encima, del percentil 90 de la muestra ELAMC.

(iii) Niños ubicados hacia la extrema izquierda del Eje 1: (I20, I31), que presentan los valores más bajos de la serie, entre los percentiles 10 y 25 de la muestra completa, para casi todas las edades.

(iv) Niños ubicados hacia la región inferior del plano (I23, I25, I3), cuyas proyecciones sobre los ejes biplot sugieren perfiles que evolucionan en forma de "U", lo que determina un patrón de recanalización en el cual los valores de TS son más altos al principio del estudio, con

un descenso hacia las edades intermedias, y retomando al final el valor inicial.

(v) Niños en la región superior del plano (I19, I30, I12) cuyas proyecciones indican un patrón de recanalización invertido en relación con el del grupo anterior, esto es, con valores de TS más altos en las edades intermedias y más bajos al inicio y al final del estudio.

Tabla 2
Información captada por el plano JK-Biplot

Proporción de información en captada por el plano biplot	$\frac{\sum_{i=1}^{2}\lambda_i}{\sum_{i=1}^{5}\lambda_i} = 0.9577$
Proporción de información en captada por el primer eje	$\frac{\lambda_i}{\sum_{i=1}^{5}\lambda_i} = 0.9254$
Relación entre el vector director Del primer eje y el perfil promedio	$v^1 = (0.47,0.42,0.43,0.45,0.46) \cong \frac{1}{185.12}(85.1,78.1,80.1,84.2,86.2)$

Tabla 3
Percentiles de concentración de triglicéridos (mg/dl), por grupos de edad. Muestra Completa ELAMC

Grupos de Edad	PERCENTILES Per3	Per10	Per25	Per50	Per75	Per90	Per97
11-12.99	32	43	57	74	98	124	164
13-14.99	26	37	53	78	103	126	174
15-16.99	37	47	60	78	111	138	171

El índice de variabilidad temporal (Variab), que se muestra en la Tabla 4, enfatiza las características de los niños pertenecientes a las dos últimas agrupaciones. La magnitud del mismo en los niños referidos, indica que sus posiciones sobre el plano están altamente determinadas por los cambios que presentan sus niveles de TS, entre período y período. En el Gráfico 2, pueden visualizarse los perfiles que describen la evolución de los triglicéridos para cada uno de los niños en el estudio, y también para el individuo promedio (Iprom). Como puede observarse, el patrón de perfiles para los niños de estos dos grupos, efectivamente corresponde a un proceso de recanalización, y está en concordancia con la proyección de los mismos sobre el plano biplot. Particularmente destacan los individuos I25, I23, I30, I3, I19,e I12, con los mayores cambios en el nivel de TS de una edad a otra.

Gráfico 1
Concentración de triglicéridos séricos (mg/dl)
Plano JK-Biplot de datos sin transformar, con origen trasladado al perfil promedio

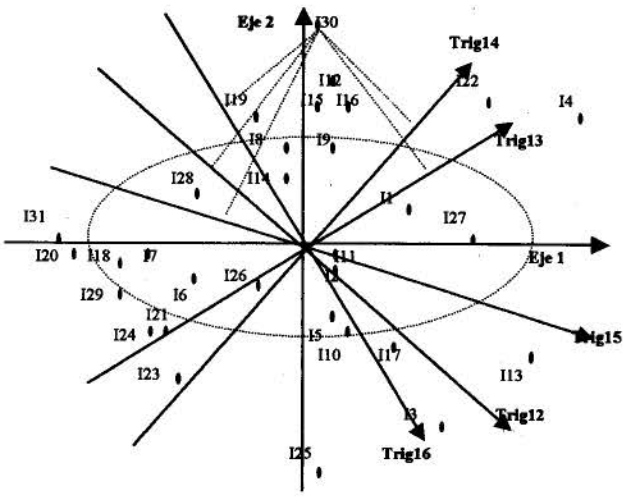
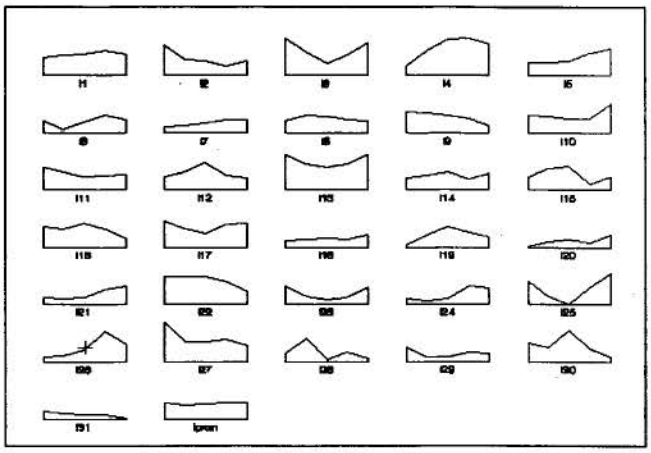


Tabla 4
Indice de variabilidad temporal

Niño	Variab	Niño	Variab	Niño	Variab	Niño	Variab
I1	0.42	I9	1.66	I17	3.39	I25	22.37
I2	1.44	I10	3.62	I18	0.80	I26	0.89
I3	6.92	I11	0.93	I19	6.99	I27	1.06
I4	1.76	I12	5.81	I20	0.45	I28	1.00
I5	2.47	I13	2.60	I21	5.24	I29	3.07
I6	1.25	I14	0.67	I22	2.28	I30	10.89
I7	0.57	I15	3.86	I23	10.97	I31	0.59
I8	1.77	I16	3.26	I24	5.13	Iprom	0.002

Gráfico 2
Perfiles de la concentración de triglicéridos séricos en niños de 12 a 16 años. ELAMC



Discusión

Los métodos multivariantes cuya álgebra y geometría están basados en la Descomposición en Valores Singulares, no son generalmente aplicados al tratamiento de datos sin transformar, debido fundamentalmente a que los efectos de ciertas medias o varianzas, pueden encubrir aspectos de importancia para el análisis de los resultados. Los criterios relacionados con este tema son controversiales: Investigadores señalan que en general, el modelo de Componentes Principales de tres víias aplicado a datos brutos produce descripciones imprecisas acerca de los datos (15y16); por otra parte, en el ámbito ecológico los investigadores suelen ser fervientes partidarios de su utilización. Noy-Meier (17) construyó una matriz de datos artificiales en la cual era posible distinguir claramente dos grupos de individuos, y encontró que en los resultados del ACP sobre datos sin transformar se evidenciaba claramente la separación de los grupos, mientras que cualquier tipo de centramiento la hacía borrosa.

Siempre que sea posible, consideramos que es altamente recomendable efectuar los análisis con y sin las transformaciones, puesto que de no hacerlo así el analista de datos podría incurrir en una pérdida importante de información. Este planteamiento puede ilustrarse fácilmente, si se supone un juego de datos con características muy peculiares respecto del conjunto de variables, como pueden ser las que se presentan frecuentemente en fenómenos biométricos: (a) medias en valor absoluto muy diferentes de cero, (b) heterocedasticidad muy marcada y (c) correlaciones de importancia moderada; bajo estas condiciones, la dirección de la recta que mejor se ajusta a las filas de una matriz de datos brutos seguramente será dominada por un vector que sencillamente refleja el alejamiento de la nube de individuos respecto de su centro de gravedad. Por otro lado, si los datos se centran, este efecto desaparece, y es muy probable que la dirección de la recta pase a ser dominada por las variables con mayor dispersión. En una situación como ésta, la no utilización de los datos brutos, podría conllevar pérdida de la información útil para detectar y caracterizar puntos atípicos.

En particular, los resultados obtenidos en esta investigación ponen de manifiesto que, en la mayoría de los varones investigados, el valor de la concentración de triglicéridos séricos es bastante estable, como es de esperarse, en edades entre los 12 y los 16 años. Sin embargo, la técnica utilizada fue capaz de determinar la existencia de subgrupos de niños atípicos con: (a) Niveles de triglicéridos séricos que pueden considerarse muy altos, como ocurre con los niños I3 e I4; (b) Niveles muy bajos, como en los niños I31 e I20; (c) Patrones de cambios bruscos en los valores de triglicéridos séricos, que los posicionan en niveles normales cuando entran a la adolescencia, con incrementos inusuales en las edades subsiguientes, y con un regreso al nivel inicial hacia el final del estudio, como es el caso de los niños I30, I19, e I12; y (d) Patrones invertidos, respecto del anterior como el que se pone de manifiesto en el niño I25, entre otros. Es decir, la metodología utilizada es capaz de detectar niños con alteraciones en los niveles de triglicéridos séricos por exceso, o por déficit, pero además también constituye un indicador de aquellos grupos de niños con valores de triglicéridos séricos que rompen con la estructura de las correlaciones seriadas, como es el caso de ciertos niños en los cuales se recanaliza la variable, que pudieran ser de alto riesgo para la aparición de enfermedades crónicas.

La metodología utilizada puede considerarse de relevancia y utilidad en la investigación clínica, si se toma en cuenta que concentraciones de triglicéridos séricos que excedan al límite superior aceptable, así como cambios bruscos en los niveles de triglicéridos, a edades tempranas, pueden constituir factores de riesgo que favorezcan la aparición de enfermedades crónicas.

La mayor importancia reside en que el método de determinación de puntos atípicos durante el seguimiento permite identificar riesgo o ausencia de riesgo, de una manera mas eficiente al relacionar al individuo con su mismo grupo, en contraste con los métodos tradicionales utilizados en clínica que usan puntos de corte de percentiles o de valores de disfusión.

Referencias

1. Seber GA. Linear Regression Analysis, Wiley, New York. 1977.
2. Jolliffe IT. Principal Components Analysis, Springer-Verlag, New York. 1986.
3. Gabriel KR. The biplot graphic display of matrices with application to principal component analysis, Biometrika, 1971;58,453-467.
4. Vasquez de Ramírez M. Aportaciones al Análisis Biplot. Un enfoque algebráico, Tesis Doctoral, Universidad de Salamanca, España. 1994.
5. Barnett V y Lewis T. Outliers in Statistical Data, Chichester, Wiley. 1978.
6. Jobson JD. Applied Multivariate Data Analysis, Springer-Verlag, New York. 1992.
7. Gnanadesikan R y Kettenring JR. Robust estimates, residuals and outliers detection with multiresponse data. Biometrics, 1972;28,81-124.
8. Gabriel KR y Zamir S. Lower rank approximation of matrices by least squares with any choice of weights. Technometrics, 1979;21,489-498.
9. Berkey C et al. Longitudinal analysis of incomplete adolescent data. Annals of Human Biology, 1991;18(4), 311-326.
10. López de Blanco M, Izaguirre I, Macias de Tomei C, Cevallos J, Bosch V, Saab L, Fossi M, Angulo N, Méndez M. Estudio Longitudinal del Area Metropolitana de Caracas. Informe Final. CONICIT. Caracas. (Mimeo).1995.
11. Méndez Castellanos H y Méndez MC. Estratificación Social y Biología Humana. Arch. Ven. Puer. Ped., 1986;49, 93-104.
12. WeinerJS y Lourie SA. Human Biology: A Guide to Field Methods, IBP Handbook Oxford Blackwell Scientific Publications, 9. 1969.
13. Greenacre M. Correspondence Analysis in Practice, Academic Press, London. 1993.
14. Galindo MP. Una alternativa de representación simultánea: HJ-BIPLOT. Questiió, 1986;10: 1,13-23.
15. Kroonemberg P. Three-mode principal components analysis. DSWO Press, Leiden. 1991.
16. Teer BraakCJ. Principal Components biplots and alpha and beta diversity, Ecology, 1983;64, 454-462.
17. Noy-MeierE Y. Data transformation in ecological ordination. I. Some advantages of non-centring. J Ecol, 1973;61, 329-41.

Determining atypical points in longitudinal studies using biplot analysis

SUMMARY. A proceedure based in the application of the Biplot analysis to non transformed data, that detects and explains why observations of a longitudinal variant nature, can be considerad atypical. The method is applied to data that describe the triglycerid pattern of in a sample of the Caracas Longitudinal boys study between 12 and 16 years of age. Results allow the identification of individuals whose triglyceride levels was consistently above the mean as well as individuals with a pattern similar to the "common" pattern. *An Venez Nutr 1999; 12(1):10-15.*

Key words: Longitudinal studies in growth, Biplot analysis, longitudinal studies, atypical points.