BAGPLOT, UN GRAFICO BIVARIADO BASADO EN LA PROFUNDIDAD

1-Introducción

El "bagplot" es una generalización al caso bivariado del box-plot. Fue propuesto por Rousseeaw y Ruts en 1997. Permite visualizar locación, dispersión, correlación, simetría y outliers en un conjunto de datos bivariados independientemente de su distribución y del sistema de coordenadas elegido.

En su construcción se utiliza el concepto de profundidad (depth) o semiespacio de profundidad (halfspace depth) introducido por Tukey en 1974 que permitió extender el concepto de rango al caso multivariado para un conjunto finito de datos.

2- GENERALIZACIÓN DEL CONCEPTO DE PROFUNDIDAD

En 1974 Tukey introdujo el concepto de **profundidad (depth)** o **semiespacio de profundidad (halfspace depth)** de un punto relativo a un conjunto de datos multivariado.

Para un conjunto unidimensional de datos $X = \{X_1, \ldots X_n\}$ se define la profundidad (depth) de un valor x relativo a un conjunto X como el mínimo número de datos a la derecha y a la izquierda de x.

$$\boxed{depth_1(x, X) = mín(\#\{i : X_i \le x\}, \#\{i : X_i \ge x\})}$$

Ejemplo:

$$X = \{1,3,5,7,9,11,13,15,17,19,21\}$$

Si queremos calcular $depth(6, X)$, hacemos:

$$\#\{i : x_i \le 6\} = \#\{1,3,5\} = 3$$
$$\#\{i : x_i \ge 6\} = \#\{7,9,11,13,15,17,19,21\} = 8$$

Luego $depth(6, X) = mín(3,8) = 3$

En el caso d-dimensional, Donoho y Gasko , definen la profundidad como:

$$\boxed{depth_d(x, X) = \min_{|u|=1} depth_1(u^T x; \{u^T X_i\}) = \min_{|u|=1} \#\{i : u^T X_i \ge u^t x\}}$$

Como $u \in R^d$ es un vector unitario, entonces el conjunto $\{u^T X_i\}$ es la proyección ortogonal del conjunto de datos X sobre un espacio de dimensión 1.

Es decir la profundidad de un punto $x \in R^d$ relativo a un conjunto X de R^d es la menor de las profundidades de x relativas a las proyecciones ortogonales del conjunto de datos X sobre espacios unidimensionales.

Tukey sugirió que a partir del concepto de profundidad (depth) se podría definir un concepto análogo al rango para el caso multivariado.

De hecho en el caso unidimensional (si no hay datos repetidos) el valor mínimo y máximo de la muestra son los puntos con $ldepth = 1$, los cuartiles 1 y 3 son los puntos con $ldepth \approx n/4$ y la mediana tiene $ldepth \approx n/2$.

Si calculamos las profundidades para cada uno de los elementos del ejemplo, obtenemos:

$$ldepth(1, X) = ldepth(21, X) = 1 \qquad ldepth(7, X) = ldepth(15, X) = 4$$
$$ldepth(3, X) = ldepth(19, X) = 2 \qquad ldepth(9, X) = ldepth(13, X) = 5$$
$$ldepth(5, X) = ldepth(17, X) = 3 \qquad ldepth(11, X) = 6$$

Observamos:

$$med = 11 \; y \; ldepth(11, X) = 6 \approx \frac{n}{2}$$
$$Q_1 = 5 \quad Q_3 = 17 \; y \; ldepth(5, X) = ldepth(17, X) = 3 \approx \frac{n}{4}$$
$$ldepth(1, X) = ldepth(21, X) = 1$$

1 y 21 son los valores mínimo y máximo respectivamente del conjunto.

3- MEDIANA DE TUKEY

La propuesta inicial de Tukey dio lugar a un número interesante de posibilidades.

Una de ellas es que permite definir la mediana en el caso multivariado. Como en el caso unidimensional la mediana es el valor de "máxima profundidad", en más dimensiones sería razonable pensar que el punto $x \in R^d$ con máxima profundidad será la mediana multivariada.

En el caso multivariado se define la **mediana de Tukey** $T_*(X)$ para $X = \{X_1, ..., X_n\}$ en $\mathbb{R}^d$ como el punto de mayor profundidad, es decir

$$T_*(X) = \arg \max_x ldepth(x, X)$$

Es invariante por transformaciones afines (por ejemplo traslaciones, rotaciones), es decir selecciona el mismo punto del espacio independientemente del sistema de coordenadas. Además es un estimador resistente.

4- REGIONES DE PROFUNDIDAD

Tukey consideró además los contornos de profundidad (**contours of depth)**, que en realidad son regiones de profundidad para indicar la forma de un conjunto de datos en R^2

El contorno de profundidad $n/4$ es una región convexa cuya forma indica la escala y la correlación de los datos.

Se define la **región de profundidad** k como

$$D_k = \{x \in R^d : ldepth(x, X) \geq k\}$$

Las regiones de profundidad forman una sucesión decreciente de conjuntos convexos encajados, es decir $D_{k+1} \subset D_k$.

Una pregunta que podemos hacernos es:

¿Cuál es el valor máximo de profundidad (depth) para un conjunto de datos?

Si el conjunto de datos es casi simétrico, el máximo valor de profundidad estará muy próximo a $n/2$.

¿Qué forma tienen las regiones de profundidad?

Esto dependerá del conjunto de datos.

5- BAGPLOT

El **bagplot** es una versión bivariada del box-plot, propuesta por Rousseeaw y Ruts. El box-plot está basado en rangos, pues la caja va desde la observación con rango $\lfloor n/4 \rfloor$ hasta la observación con rango $\lceil 3n/4 \rceil$, y la línea en el interior de la caja es la mediana, por lo tanto es necesario generalizar el concepto de rango al caso multivariado.

Como vimos antes una natural generalización del rango al caso multivariado es la noción de profundidad introducida por Tukey en 1974.

Luego el **bagplot** se basa en las regiones de profundidad y en la mediana de Tukey definidas anteriormente.

En las figuras 5.1 y 5.2 se presentan dos versiones del bagplot para un mismo conjunto de datos.

Estos dos gráficos corresponden a la primera versión escrita en 1997 por Rousseeaw y Ruts en Fortran 77 e implementada en S-plus 2000.

El bagplot está formado por una "bolsa" que contiene al 50% central del conjunto de los datos (representada por el polígono oscuro), una zona alrededor de la bolsa que denominaremos vecindad (representada por el área más clara) que contiene los inliers más extremos y una valla (cuyo gráfico es opcional y se representa con línea punteada en la figura 5.2) que separa los inliers de los outliers.

La mediana de Tukey es indicada por una cruz y los outliers están indicados por asteriscos.

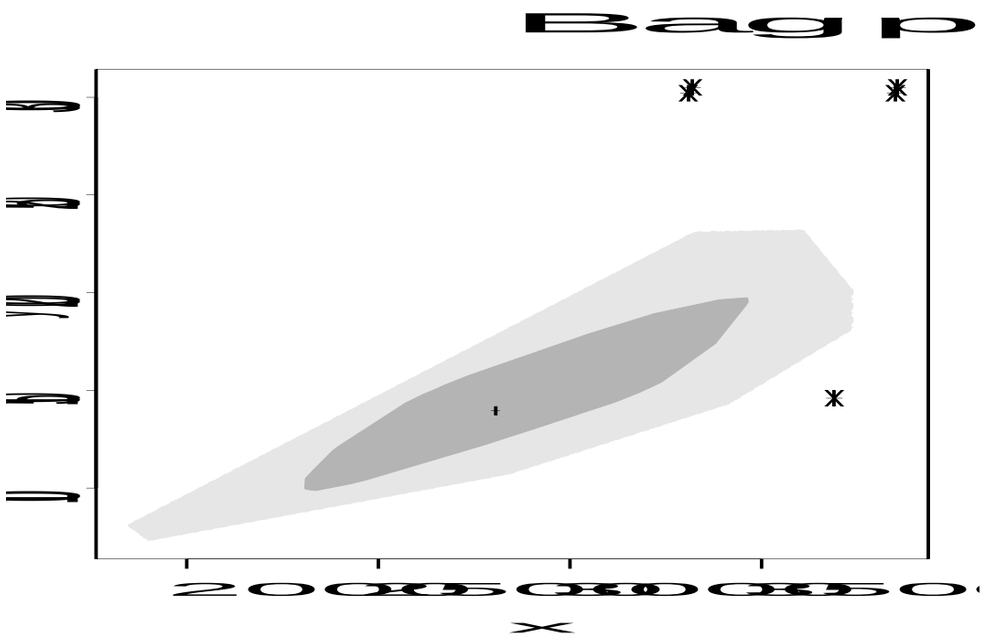


Figura 5.1

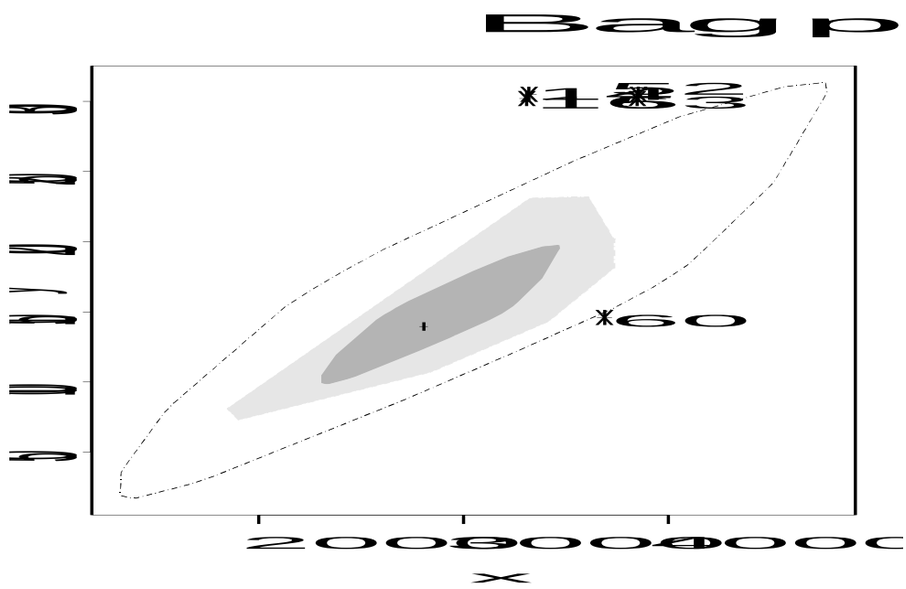
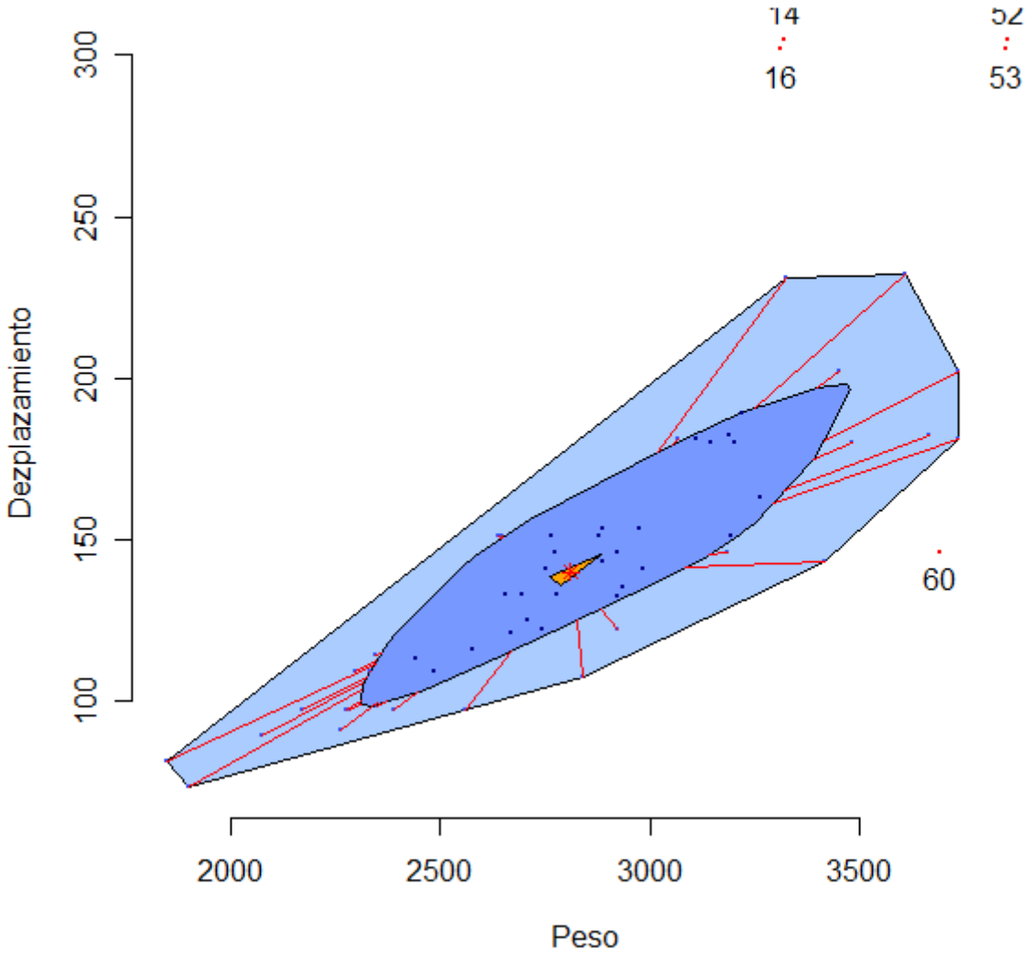


Figura 5.2

El siguiente gráfico (Figura 5.3) corresponde a la versión del 2007 de Peter Wolf para R.



En las dos figuras fueron incluidos todos los puntos muestrales, y en la figura 5.2 se identificaron los outliers.

La figura 5.1 es la salida que se obtiene por defecto para una muestra con 15 o más datos; pero el programa permite diferentes opciones, como por ejemplo la inclusión de todos los puntos de la muestra y su identificación (clickeando sobre ellos).
Esto corresponde a la primera versión.

El programa también calcula las coordenadas de la mediana de Tukey, que para nuestro ejemplo son (2806.63, 139.513).

El conjunto de datos usado para el ejemplo pertenece a la librería del S-PLUS. La variable x es el peso y la variable y el desplazamiento del motor de 60 autos chicos, medianos y grandes.

Como el boxplot en el caso univariado, el bagplot nos permite visualizar diferentes características del conjunto de datos: su **posición** (la mediana de Tukey), la **dispersión** (el tamaño de la bolsa), la

correlación (la orientación de la bolsa), la **simetría** (la forma de la bolsa y la vecindad) y las **colas** (la amplitud de la vecindad y los outliers).

4.2 Construcción del bagplot

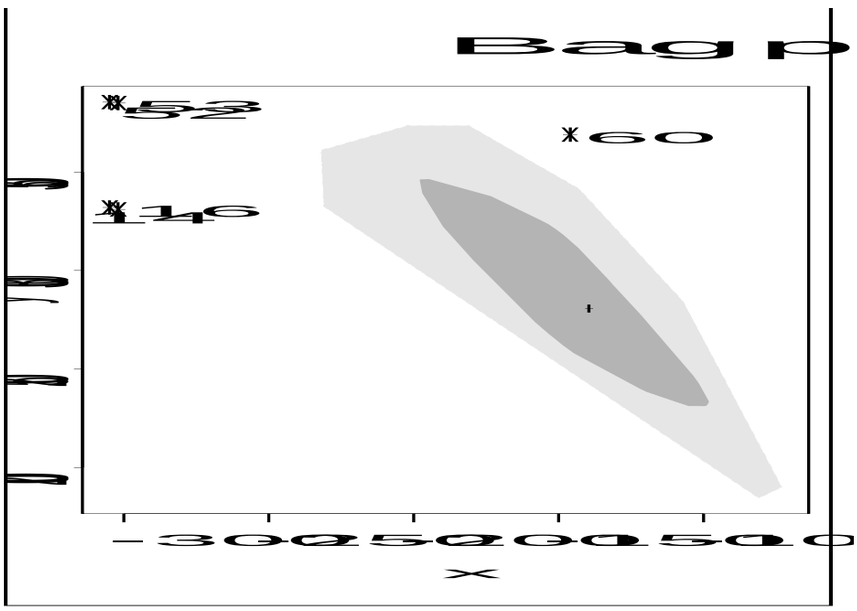
Para construir la **bolsa** B**,** primero se determina el valor de k tal que $\#D_k \le \lfloor n/2 \rfloor < \#D_{k-1}$ y se interpola linealmente en forma relativa a la mediana de Tukey T_*. La bolsa es un polígono convexo.

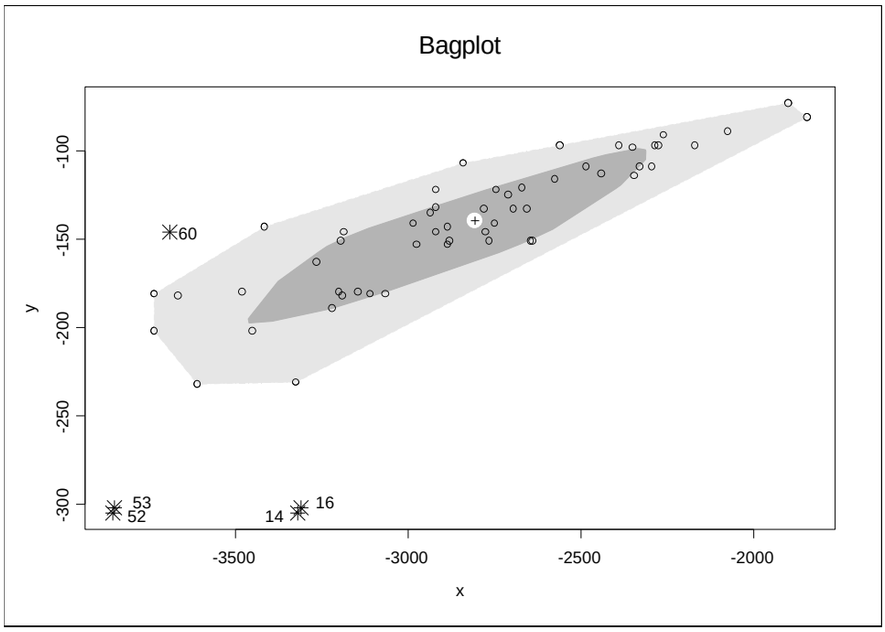
La **valla** es obtenida expandiendo la bolsa por un factor 3 relativo a T_*. Este valor 3 fue obtenido en base a simulaciones.

La **vecindad** contiene todos los datos entre la bolsa y la valla, es decir rodea a la bolsa y contiene a los puntos que no son outliers.

Cuando las observaciones están sujetas a una traslación o una transformación lineal no singular (por ejemplo una rotación), el bagplot es transformado en forma acorde, pues la profundidad es invariante por transformaciones lineales afines y los conjuntos convexos se transforman en conjuntos convexos. Entonces los puntos dentro de la bolsa siguen siendo interiores y los outliers siguen siendo outliers. Para observar ésto al conjunto de datos del ejemplo de la figura 5.1 se le aplicó una rotación de 90º y 180º, los bagplots obtenidos son los de la figura 5.4 y 5.5 respectivamente.

Comparando los dos gráficos con el gráfico de la figura 5.1 podemos observar que los outliers siguen siendo los casos 14, 16, 52, 53 y 60.





Visualización de estructuras de correlación: ejemplos

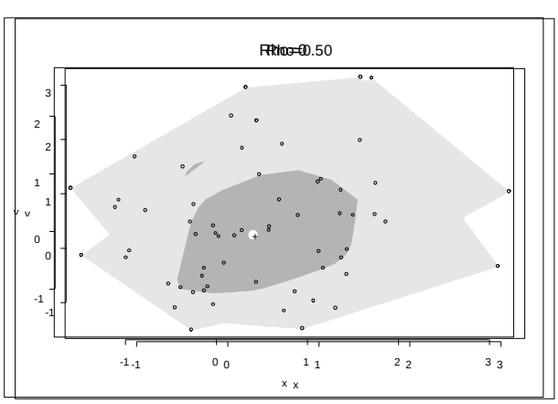
Para observar cómo se reflejan en el bagplot diferentes estructuras de correlación, se generaron muestras aleatorias, cada una con 50 datos, de normales bivariadas con distintos coeficientes de correlación ρ programando una función en S-PLUS .

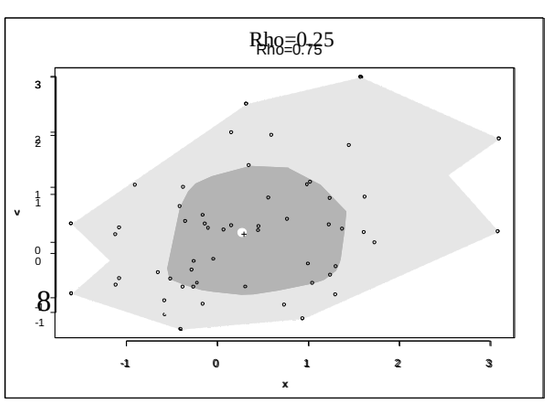
Los bagplots obtenidos se muestran en las figuras a y b Podemos observar que cuando $\rho = 0$ la bolsa es casi un círculo que se va deformando con forma casi elíptica hasta llegar a un boxplot univariado cuando $\rho = 1$ o $\rho = -1$.

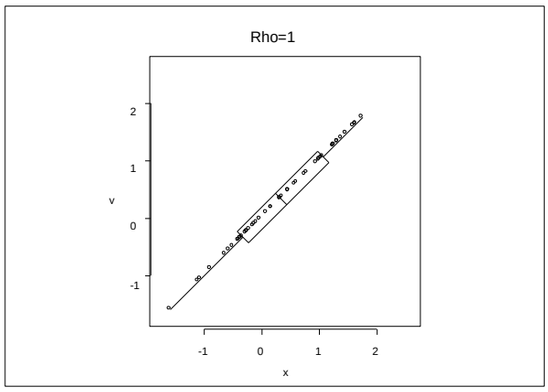
La orientación de izquierda a derecha es ascendente cuando $\rho > 0$ y descendente cuando $\rho < 0$.

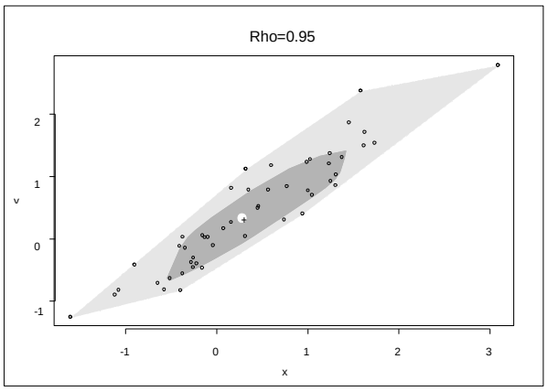
Podemos observar también en todos los gráficos la simetría de los datos por la forma de la vecindad.

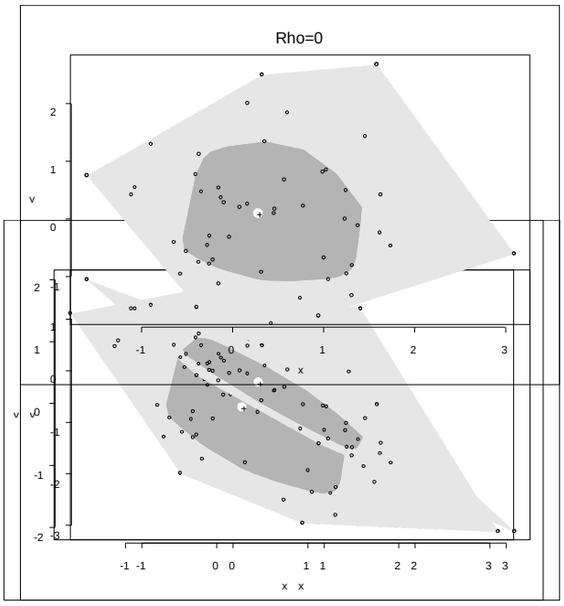
En estos ejemplos no hay outliers.

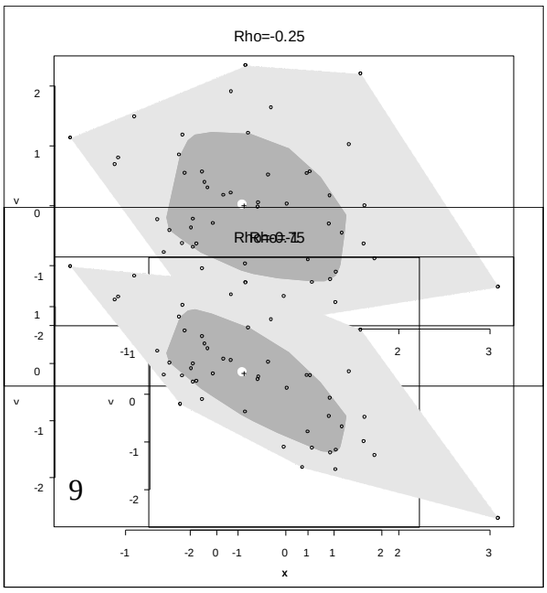





Rho=1
v
x


Rho=0.95
v
x


Rho=0
v
x

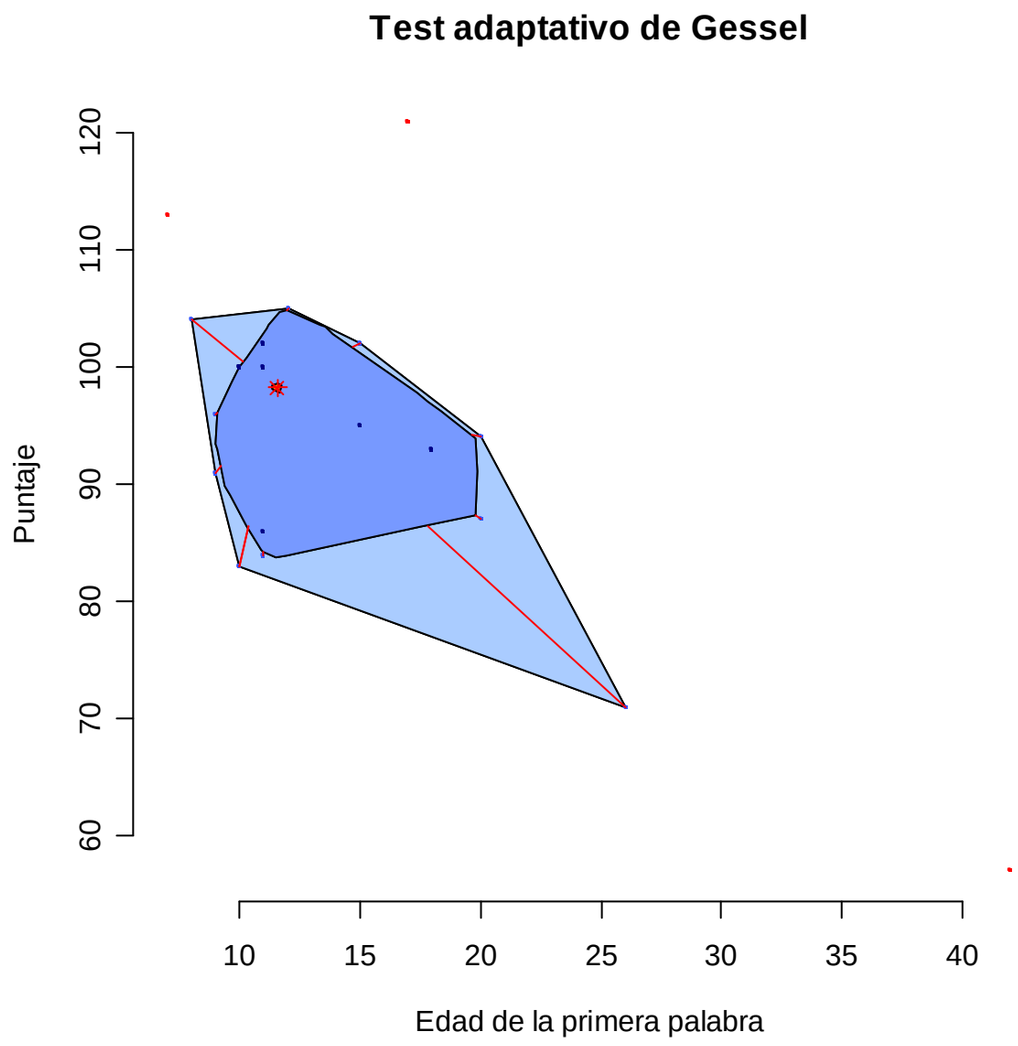

Rho=-0.25
v
x

Otros ejemplos

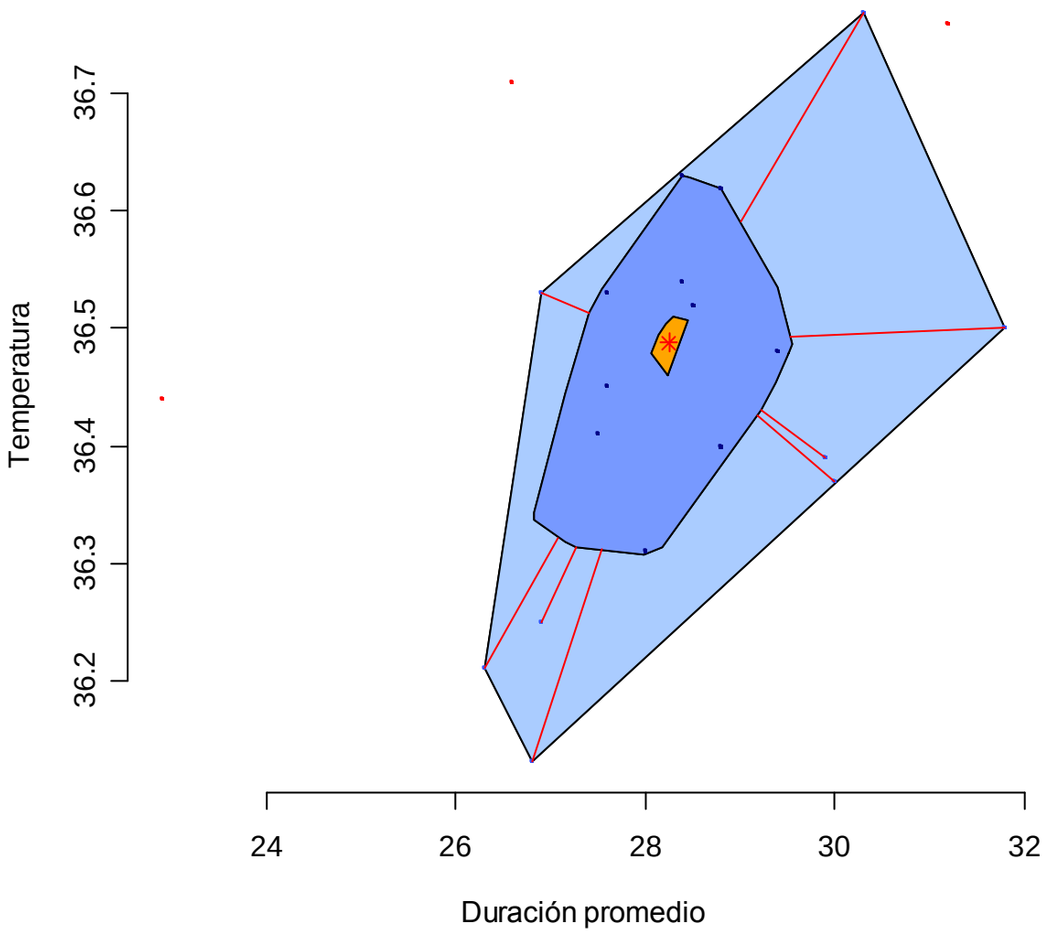
El siguiente ejemplo corresponde a una muestra de 21 niños a los que se les registró la edad en meses a la cual cada uno dijo su primera palabra y el puntaje obtenido en la prueba de Gessell tomada mucho despues. El propósito del trabajo fue determinar si la edad de la primera palabra (x) podría predecir el puntaje posterior del test (y). Los datos se encuentran en el archivo Gessel.txt

Las coordenadas de la mediana de Tukey son (11.5594 , 98.262)

Si observamos podemos decir que existe una correlación negativa entre las variables x e y , el conjunto de datos es asimétrico, pues la mediana de Tukey se encuentra en la parte superior izquierda de la bolsa donde la vecindad es muy estrecha, su forma además no sugiere una distribución elíptica y se observan 3 outliers, correspondientes a los casos 11, 18 y 19.



El último ejemplo corresponde a Royston y Abrams (1980) que midieron la duración promedio del ciclo menstrual en días (x) y la temperatura basal preovulatoria promedio en ºC (y) de un grupo de 21 mujeres sanas que estaban usando un método de planificación familiar natural. Los datos se encuentran en el archivo ejtuckey2.txt



BIBLIOGRAFIA

Donoho, D. L. y Gasko, M., (1992), "Breakdown Properties of Location Estimates Based on Halfspace Depth and Projected Outlyingness", *The Annals of Statistics*, **20**, 1803-1827

Rousseeaw,P. J., Ruts, I. , y Tukey J. W., (1999), "The Bagplot: A Bivariate Boxplot", *TheAmerican Statistician*, **53**, 382-387