Universidad EAFIT

Escuela de ciencias aplicadas e ingeniería

Estadística II

Tareas

Semestre 2023-1

Nombre: Julieth Stefanny Escobar Ramírez y Sara Gallego Villada

1 Detección de outliers

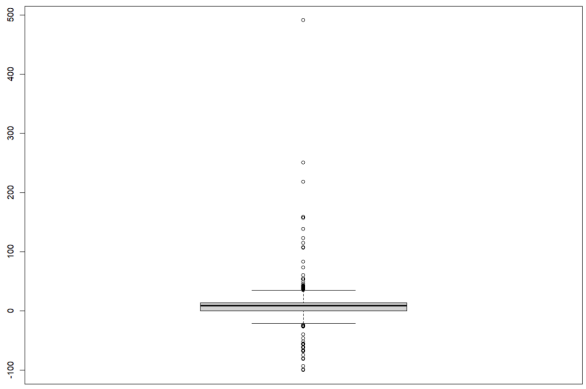
1.

Boxplot:

El boxplot es un tipo de gráfico que muestra un resumen de una gran cantidad de datos en cinco medidas descriptivas(el mínimo, el primer cuartil, la mediana, el tercer cuartil y el máximo.)
Consiste en una gráfica, en la cual dibujamos una caja desde el primer cuartil hasta el tercer cuartil. Además de una línea vertical que atraviesa la caja en la mediana. Y los bigotes van desde cada cuartil hasta el mínimo o el máximo.
Este tipo de gráficos nos permite identificar valores atípicos y comparar distribuciones. Además de conocer de una forma cómoda y rápida como el 50% de los valores centrales se distribuyen.

Implementación en R:

La base de datos con al que trabajamos son las rentabilidades de los Fondos de Inversión Colectiva (FIC).Para el boxplot usamos la rentabilidad semestral.



Boxplot rentabilidad semestral

De este boxplot podemos observar que la mediana no se encuentra en el centro de la caja sino un poco más arriba, lo que nos índica que la distribución no es simétrica. También podemos observar que en los datos tomados se encuentran muchos datos atipicos, estos valores atípicos pueden afectar nuestros resultados en proximos análisis por lo tanto es

muy importante tenerlos en cuenta.

2.Tomar dos variables y hacer un bag-plot,hacer paso a paso, elementos y aplicaciones

Bag-plot:

El bag-plot también llamado trama starburst, es un método en estadísticas sólidas para la visualización de dos o datos estadísticos tridimensionales, análoga a la unidimensional diagrama de caja .fueron propuestos inicialment por Rousseeuw, Ruts y Tukey en como una extension al caso multivariante del boxplot. El bag-plot permite visualizar la ubicación, la dispersión, la asimetría y los valores atípicos de un conjunto de datos.

Construcción:.
Sea $x \in R^2$ un punto y $ldepth(x, X)$ su profundidad de Tukey con respecto a una muestra $X = \{x1, x2, ..., xn\}$.
Definimos la region de profundidad Dk como el conjunto de puntos x tales que $ldepth(x, X) \leq k$ con $0 \leq k \leq 1$. Estas zonas son en realidad polígonos convexos con la propiedad de que $D_{k+1} \subset D_k$. La mediana muestral sabemos que es el punto T^* que maximiza $ldepth(x, X)$. En el caso de que no fuese unico basta elegir el centro de gravedad de los candidatos.

El bag-plot consta de tres polígonos anidados , llamados "bolsa", "valla" y "bucle". La mochila se construye de la siguiente manera:

- El polígono interior, llamado bolsa, se construye sobre la base de la profundidad de Tukey, el menor número de observaciones que puede contener un semiplano que también contiene un punto dado. Contiene como máximo el 50% de los puntos de datos.
  Sea D_k el numero de puntos muestrales contenidos en D_k. Primero se calcula el valor de k tal que $D_k \leq \frac{n}{2} \leq D_{k1}$ y luego se interpola linealmente entre D_k y D_{k1} (relativo al punto T^*), esta la denotaremos como B.

- El más exterior de los tres polígonos, llamado valla, no se dibuja como parte de la gráfica de bolsas, pero se usa para construirlo. Se forma inflando la bolsa por un cierto factor (generalmente 3).
  La valla se obtiene inflando B (en relación a T^*), por un factor, que usualmente es $\rho = 3$ -valor experimental obtenido a través de simulaciones- aunque se puede usar también $\rho = 2.58$ ya que cuando los datos siguen una distribución gaussiana se encuentran en un 99 %. Los puntos fuera de la valla son considerados como outliers.

- Las observaciones que no están marcadas como valores atípicos están rodeadas por un bucle, el casco convexo de las observaciones dentro de la cerca.

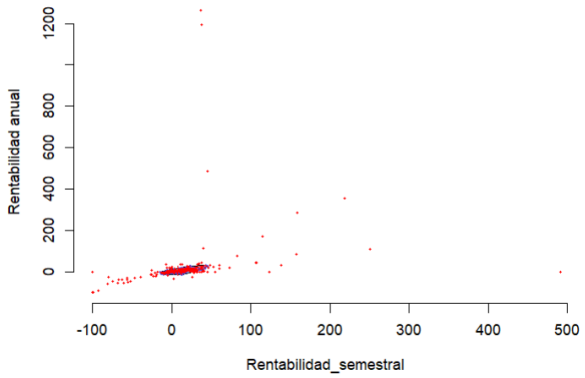
Propiedades:

Una propiedad que posee el diagrama de mochila, es que debido a la invariabilidad de la profundidad de Tukey por transformaciones afines, si trasladamos o rotamos la muestra, el bagplot no cambia. Por lo tanto el gráfico de bolsas es invariante bajo transformaciones

afines del plano y robusto frente a valores atípicos.

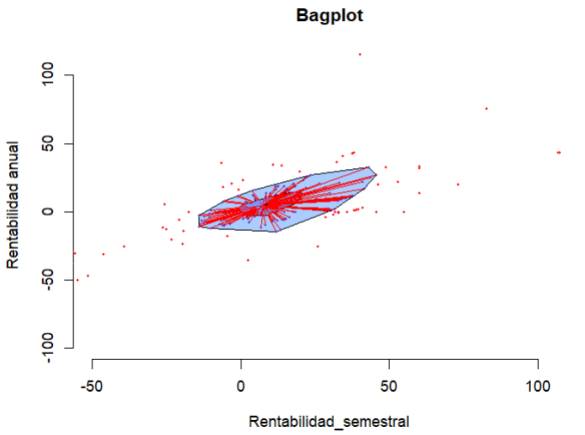
Implementación en R:

Para la implementación del bagplot usamos la misma base de datos, pero esta vez se usaron dos variables distintas las cuales son, la rentabilidad semestral y la rentabilidad anual.
Primero se hizo un bagplot tomando en cuenta todos los datos de las variables tomadas, de lo cual se obtuvo la siguiente gráfica:



Bag-plot rentabilidad semestral y rentabilidad anual

De este bagplot se obtiene una muy mala visualización de los datos, esto debido a la presencia de tantos outliers.
Por lo tanto para visualizar mejor los datos se realizó otra gráfica en la que se reducieton los límites en los ejes x,y, de esta forma se logra ignorar la presencial de algunos outliers en la gráfica, de la cual el resultado fue:



Bag-plot rentabilidad semestral y rentabilidad anual

3.Con $p \geq 2$ hallar la distancia de Mahalanobis y el determinante de covarianza mínima.

Definición: El determinante de covarianza mínima es un estimador muy útil para la detección de outliers. Es especialmente en más de dos variables. El conjunto de datos

analizados, es una submuestra cuyo tamaño es $h < n$, de tal forma que esta submuestra se caracterice por poseer la matriz de covarianza con mínima determinante de todas las matrices de covarianzas calculadas.La idea detrás de esto es que los valores atípicos en los datos afectarán la matriz de covarianza y, por lo tanto, aumentarán su determinante. Por lo tanto, al elegir la submuestra con la menor covarianza posible, se espera que se eliminen los efectos de los valores atípicos en la matriz de covarianza.
En resumen, este método busca obtener una estimación precisa y confiable de la matriz de covarianza a partir de una submuestra de datos. Referencia

Definamos $\mathbf{X}$ como una matriz $n \times p$, siendo n el número de objetos y p el número de variables donde $\mathbf{X} = (x_1, x_2, ...,_n)^t$, siendo $\mathbf{x_i} = (x_{i1}, x_{i2}, ..., x_{ip})^t$ la $i - ésima$ observación.

Sea x una submuestra de tamaño h tomada desde $\mathbf{X}$ tal que:

$$h = \left\lfloor \frac{n+p+1}{2} \right\rfloor$$

Donde $\lfloor . \rfloor$ es la parte entera del cociente con $n/2 \leq h < n$, y cuanto mayor sea h más robusto es el estimador.

Ahora asumimos que las observaciones son una muestra de una distribución unimodal elípticamente simétrica con parámetros desconocidos μ y $\mathbf{\Sigma}$, donde μ es un vector con p componentes y Σ es una matriz $p \times p$ definida positiva.

Decimos que una distribución es elípticamente simétrica y unimodal si existe una función decreciente estricta g tal que la densidad puede ser escrita de la siguiente forma

$$f(x) = \frac{1}{\sqrt{|\Sigma|}} g(d^2((\mathbf{x}, \mu, \mathbf{\Sigma})) \tag{1}$$

Ahora, definamos $\hat{\mu}$ (vector de medias submuestra):

$$\hat{\mu} = \frac{1}{h} \sum_{j \in x} x_j \qquad \text{con } j = 1, ..., h$$

También tenemos $\hat{\Sigma}$ que seria la matriz de covarianza para nuestra submuestra:

$$\hat{\Sigma} = \frac{c(h)s(h,n,p)}{h-1} \sum_{j \in x} (x_j - \hat{\mu})(x_j - \hat{\mu})^T \qquad \mathbf{con} j - 1, ..., h$$

Ahora sabiendo que X es elíptica, simétrica y unimodal, decimos que $c(h)$ dado por:

$$c(h) = \frac{h/n}{P(\chi^2_{p+2} < \chi^2_{p,1-h/n})}$$

y $s(h, n, p)$ es un factor de corrección para muestras finitas.
Finalmente se hace uso de algúm algoritmo para definir h en el ques e obtinen diferentes H_i y el proceso sería el siguiente:

- En primer lugar, se establece $\bar{\mu}$ y $\bar{\Sigma}$ de un conjunto inicial de datos que contiene h observaciones.
- Luego, se utiliza el vector de medias y la matriz de covarianza establecidos anteriormente para calcular la distancia Mahalanobis de cada observación.
- A continuación, se elige una submuestra de datos que tenga las observaciones con las distancias Mahalanobis más pequeñas de entre todas las disponibles en el paso anterior.
- Finalmente los pasos 1, 2 y 3 se repiten de forma hasta encontrar una submuestra que tenga las distancias Mahalanobis(Robusta) más pequeñas en el paso 2. Es decir, se ajusta el vector de medias y la matriz de covarianzas de la nueva submuestra, se recalculan las distancias Mahalanobis para cada observación disponible y se selecciona una submuestra de observaciones con las distancias más pequeñas. Este proceso se repite varias veces hasta encontrar la submuestra óptima.

Donde la distancia de Mahalanobis esta dada por la distancia:

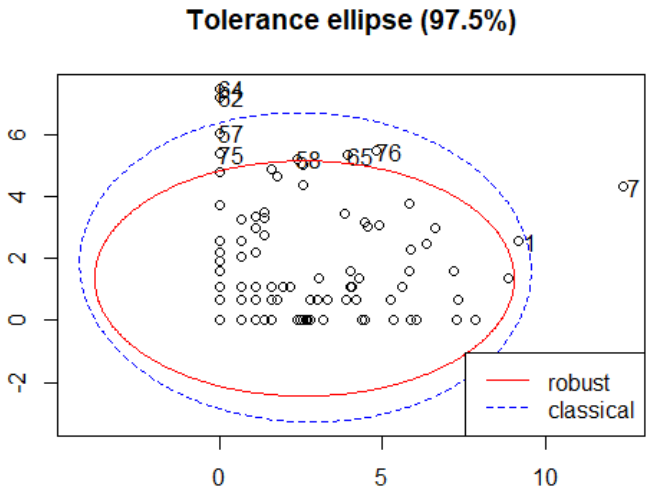
$$d^2(\mathbf{x}, \mu, \mathbf{\Sigma}) = (\mathbf{x} - \mu)'\mathbf{\Sigma^{-1}}(\mathbf{x} - \mu) \tag{2}$$

Esta distancia la podemos definir como:

$$d^2(\mathbf{x}, \bar{\mathbf{x}}, \mathbf{S}) = (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S^{-1}}(\mathbf{x} - \bar{\mathbf{x}}) \tag{3}$$

Que es la que define la elipse de tolerancia, y este es igual a $\chi^{\mathbf{2}}_{\mathbf{p},\alpha}$ donde esto es el cuantil α de la distribución χ^2_p, donde S es la matriz de covarianza de la muestra y $\bar{x}$ la media de la muestra. Esta distancia nos dice que tan lejos esta $\mathbf{x_i}$ del centro de la nube de datos con respecto al tamaño y forma.

Resultados implementación Utilizamos las columnas de rentabilidad diaria y numero de inversionistas obtuvimos los siguientes resultados con una muestra $n = 99$
La implementación se realizó con la librería de R *DETMCD* Referencia implementación

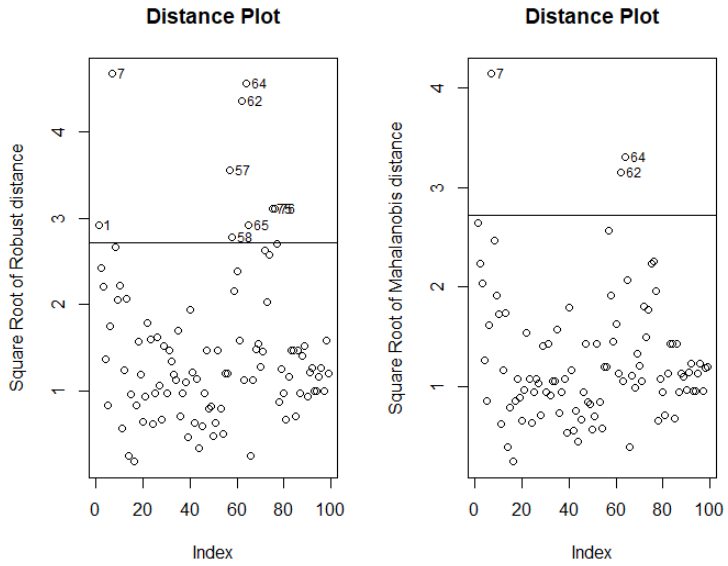


Elipse tolerancia 97.5%

Vemos una clara diferencia entre la forma robusta de identificar outliers lo cual hace que la distancia de Mahalanobis clásica tenga más datos atípicos como datos dentro de la media que verdaderamente estarían afectando el resumen de los datos.

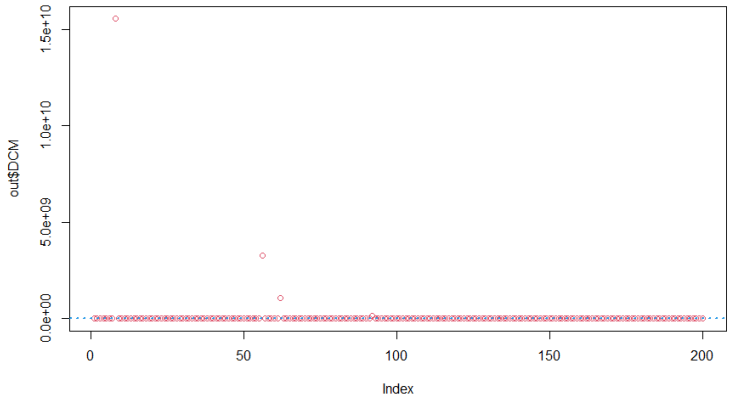
En la siguiente gráfica se percibe con más facilidad la diferencia entre las distancias, la robusta siendo el MCD y la de clásica que seria la Mahalanobis, al robusto ayuda a que nuestros datos sean más precisos y no en entorpezcan los resultados.

Referencias



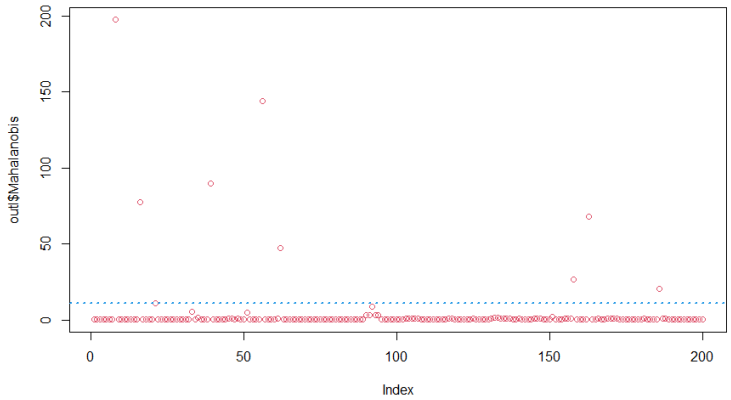
Comparación distancia MCD y Mahalanobis clásica

$P > 2$
Ahora, tomamos las variables *numero inversionistas, rentabilidad diaria,aportes recibidos valor unidad operaciones dia t* En la cual comparamos los resultados con el MCD y la distancia de Mahalanobis.
En este caso utilizamos *cov.mcd* para la implentación del algoritmo.
Se obtuvo los siguientes resultados:



Det MCD 4 varibales

Ahora para Mahalanobis

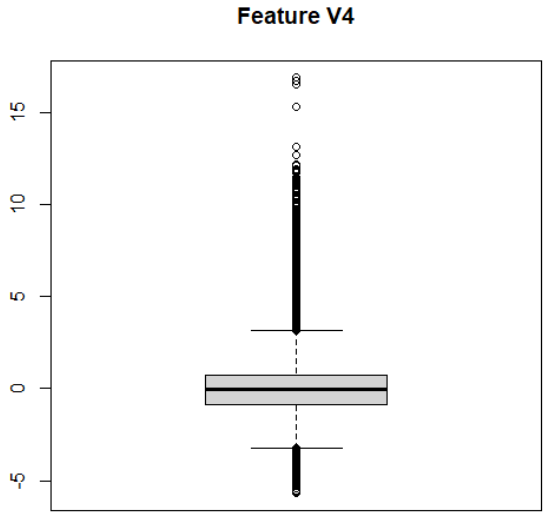


Mahalanobis 4 varibales

En los resultados vemos que la distancia de Mahalanobis tiene más datos atípicos, esto se debe a que es más sensible pues al no ser robusto como el DCM, la matriz de covarianza también es afectada por datos atípicos. Mientras que el método DETMCD utiliza una estimación, que es menos sensible a valores atípicos que la estimación de la matriz de covarianza utilizada por el método de Mahalanobis lo que hace que sea menos afectado por la dimesión del problema.

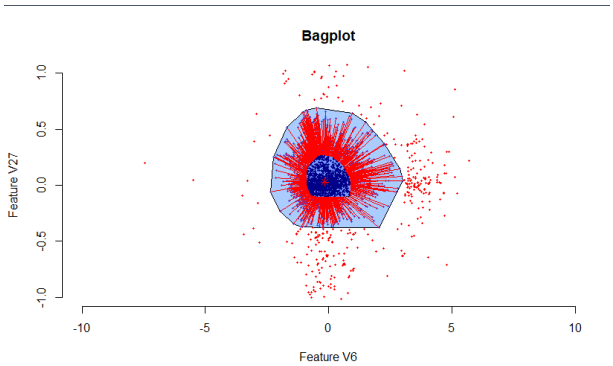
Los datos para esta entrega fueron tomados de: datos abiertos

Dado a la forma de nuestros datos se repitió el proceso con una base de datos para detectar fraudes. Estos datos son simulados Credit Card Fraud Detection
Se obtuvo los siguientes resultados.
En este caso utilizamos las variables $V4$ que es una de las características de los clientes para detectar fraudes, sin embargo no se especifíca cual en la base de datos por temas de privacidad.



Bloxplot

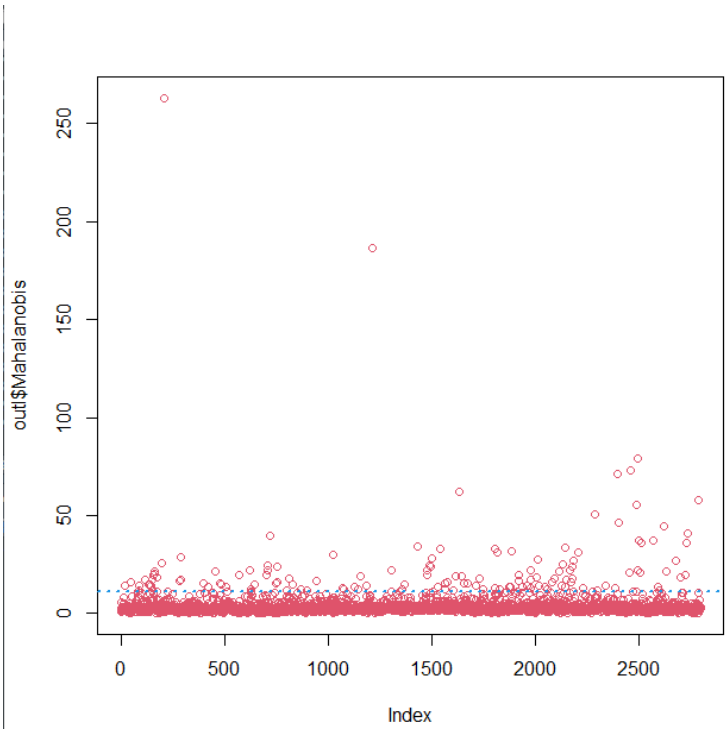
Vemos que en nuestra muestra de 2800 de 284000 datos hay outliers especialmente por encima del 5 cualtil
Ahora el bagplot con las variables *V6* y *V27*:



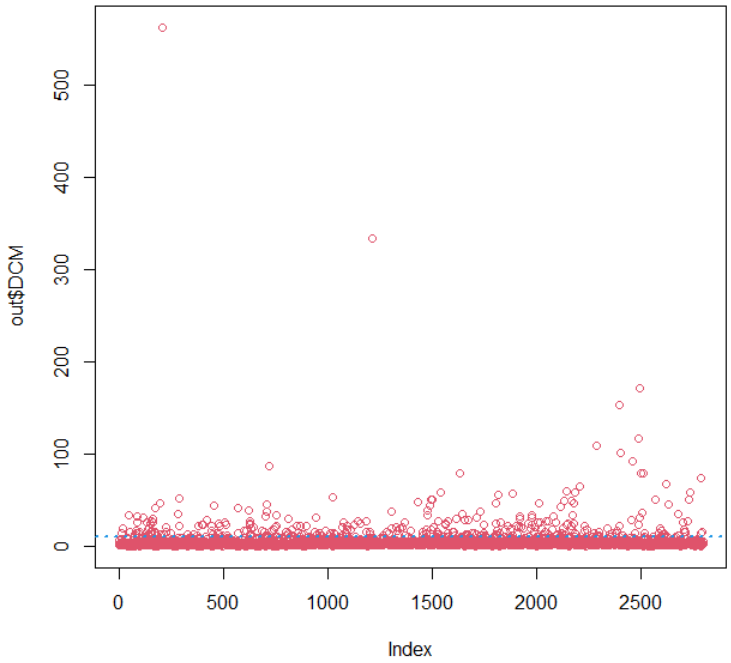
Bagplot Fraude tarjetas de crédito

En este caso es más clara la identificación de outliers.

De igual manera implementamos el Determinante de covarianza mínima y la distancia de Mahalanobis y se obtuvo esto:



Mahalanobis Fraude tarjetas de crédito



CMD Fraude tarjetas de crédito

En este caso vemos que los outliers de la distancia robusta son más, es decir, por encima del estadístico que es la linea punteada azul hay más "puntos". Análogamente a los datos previos esto sucede por la característica del muestreo y matrices de covarianzas.

Para ver los resultados se ajunta el acceso directo a carpeta con códigos y datos