UNIVERSIDAD EAFIT

Técnicas robustas y no parámetricas

Taller 2

Estadística no paramétrica

Profesor:
Henry Laniado
Año Académico:
2022

Abelino Sepulveda Estrada
Camilo Oberndorfer Mejía
Luisa Toro Villegas

---

1. Estimaciones de las densidades y las correlaciones entre los activos

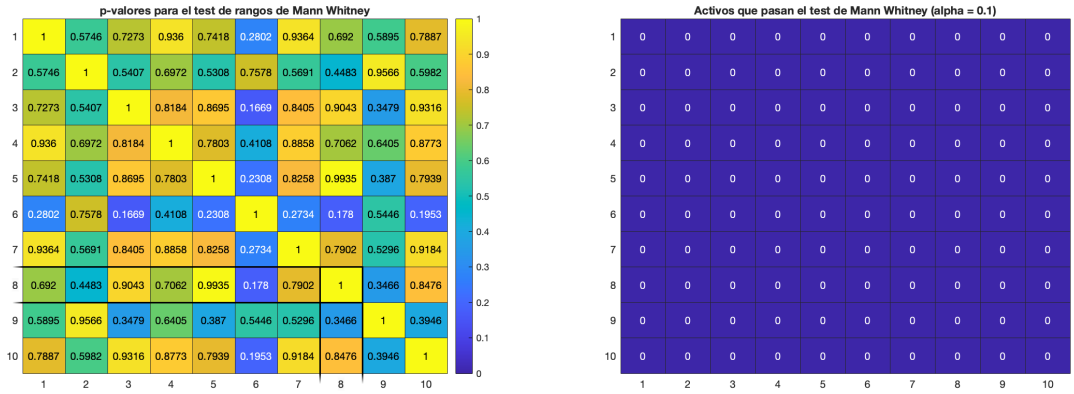
1.1. Test de rangos de Mann-Whitney

El test de *Mann–Whitney–Wilcoxon (WMW)*, también conocido como Wilcoxon rank-sum test es un test no paramétrico de rangos para determinar si dos muestras provienen de poblaciones equidistribuidas.

El test WMW contrasta que la probabilidad de que una observación de la muestra X sea mayor a una observación de la muestra Y sea igual a que una observación de la muestra X sea menor a una observación de la muestra Y. Es decir, que los valores de una población no tienden a ser mayores que los de otra. Quedando las siguientes hipótesis

$$H_0 : P(X > Y) = P(X < Y) = 0.5$$
$$H_a : P(X > Y) \neq P(X > Y) \neq 0.5$$

Se realizó el test de rangos WMW considerando los últimos 900 meses de fichero *returns.txt* obteniendo los resultados que se muestran en la Figura 1



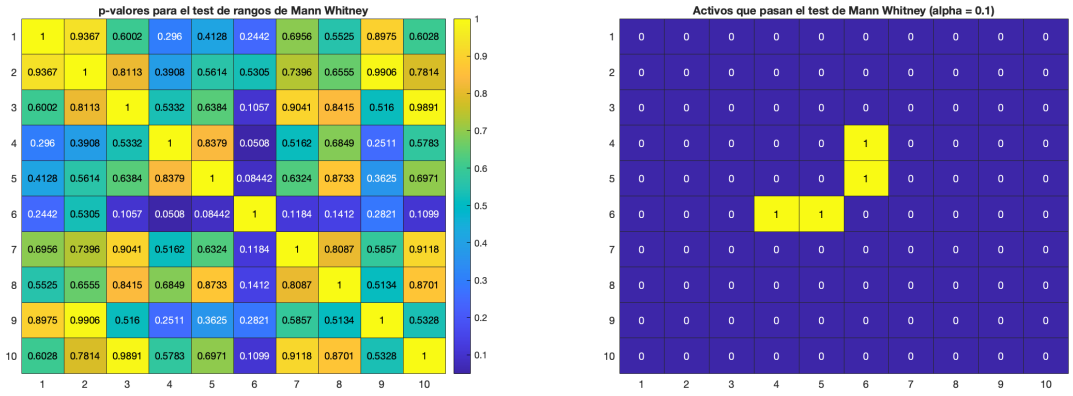
(a) p-valores

(b) Activos que no pasan el test

Figura 1: Test de Mann-Whitney-Wilcoxon para los últimos 900 meses

En la Figura 1a se evidencia el heatmap de los p-valores del test de cada par de activos del fichero y en la Figura 1b se observan los pares de activos que pasan y que no pasan el test. Si el valor es cero quiere decir que no hay evidencia suficiente para rechazar que los pares de activos provienen de una misma distribución y uno quiere decir que se rechaza la hipótesis nula, es decir, que vienen de distinta distribución.

Como observamos en 1b, ningún par de activos provienen de distinta distribución. Por este motivo, se decide cambiar el periodo de observación. Así, consideramos los primero 450 meses del fichero.



(a) p-valores (b) Activos que pasan el test

Figura 2: Test de Mann-Whitney-Wilcoxon para los primero 450 meses

Analizando los resultados de la Figura 2 se evidencia que los pares de activos (4,6), (5,6) no pasan el test, es decir, hay suficiente información para determinar que los dos pares de activos provienen de distinta distribución.

2. Densidades de los activos que según el test de Mann Whitney

Estimamos las densidades entre cada par de activos y calculamos el área en común entre las densidades de cada par de activo para determinar cuáles eran más y menos parecidos.

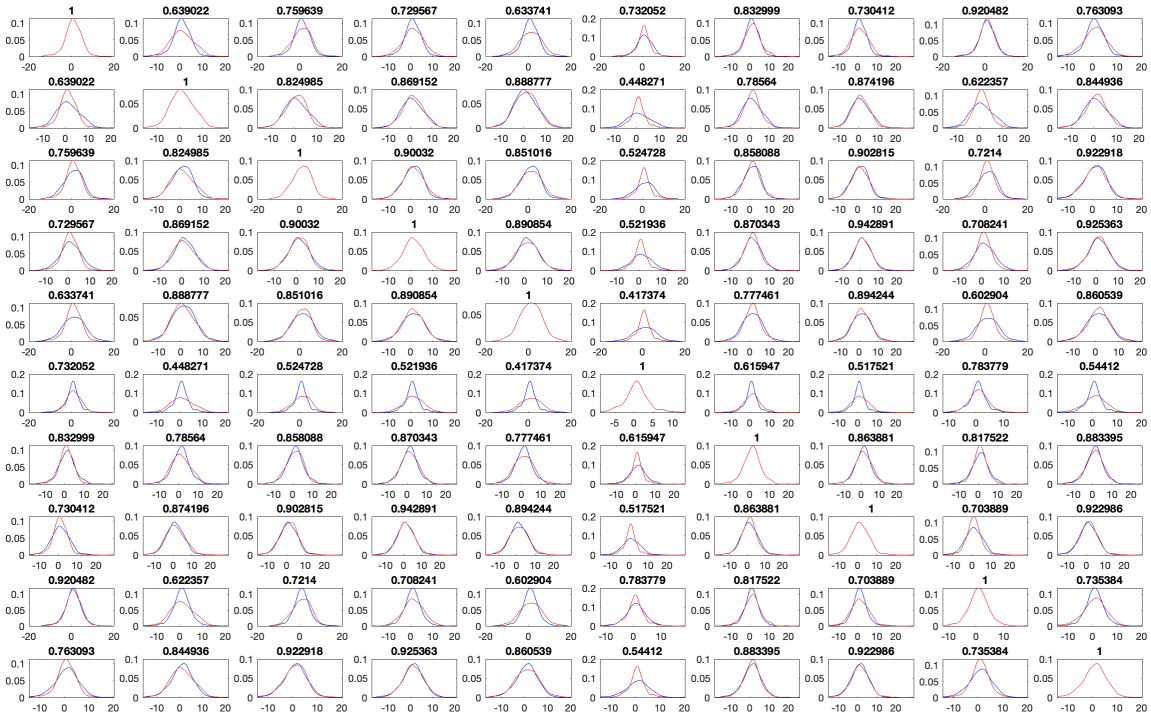


Figura 3: Estimación de las densidades de los primeros 450 meses de los activos

Las estimaciones de las densidades mostradas en la Figura 3 son por medio de un *kernel Gaussiano*. Para realizar el calculo del kernel se toma la formula:

$$\widehat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h\left(x - x_i\right) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \tag{1}$$

Una vez teniendo en cuenta que el kernel utilizado es uno Gaussiano:

$$G_{1D}(x;\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}} \tag{2}$$

En este se tomo el ancho de banda como:

$$\text{BandWidth } = \sigma * \frac{4}{3*N}^{1/5} \tag{3}$$

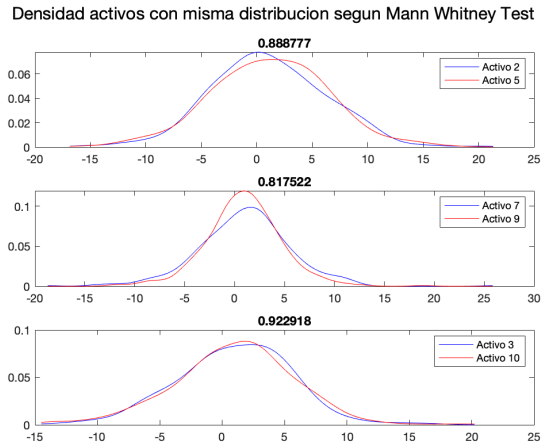
Donde N es la longitud de los datos σ se calcula mediante la estimación robusta por medio de MAD:

$$\sigma = k * median(|x_i - median(x)|), \quad k = 1.4826$$

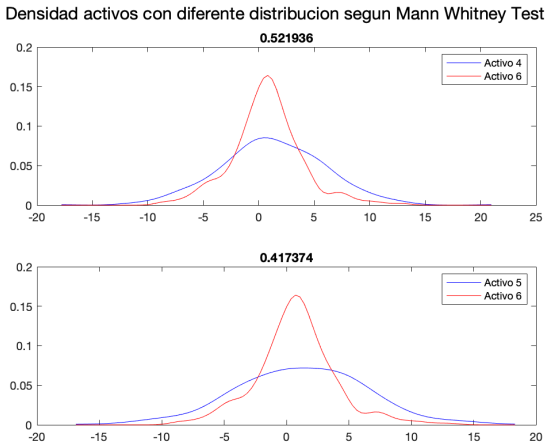
Activo	1	2	3	4	5	6	7	8	9	10
Bandwidth	1.085	1.571	1.381	1.439	1.668	0.789	1.224	1.397	1.004	1.339

Tabla 1: Bandwidth para cada activo

En la Tabla 1 se muestra los anchos de banda de las densidades de cada activo según la ecuación 3. Ahora, graficamos las densidades de los pares de activos que provienen de distinta distribución según el test de WMW mostrados en la Figura 2, adicional a esto, graficamos los pares de activos que cuyas densidades tienen mayor área en común mostrados en la Figura 3



(a) Densidades Similares



(b) Densidades Distintas

Figura 4: Densidades con Kernell Gaussiano de Activos segun el MannWhitney Test

2.1. Modelo de regresión lineal

Consideramos un par de activos que pasaron el test de WMW (en este caso el 5 y 6) y construimos un modelo de regresión lineal sobre estos.

El primer modelo de regresión fue realizado de forma paramétrica. Donde aproximamos una relación de dependencia entre la variable dependiente (*activo 5*) y la variable independiente (*activo 6*). De la forma

$$activo\ 6 = \beta_0 + \beta_1 * activo\ 5 + \epsilon, \quad \epsilon \sim N(0,1) \tag{4}$$

En la ecuación 4 podemos evidenciar de forma clara por qué el test es paramétrico y es porque asume que los residuales del modelo distribuyen normal.

Estos modelos de regresión trabajan sobre las siguientes hipótesis:

$$H_0 : \beta_0 = 0 \ \ \& \ \ \beta_1 = 0$$
$$H_a : \beta_0 \neq 0 \ \ \& \ \ \beta_1 \neq 0$$

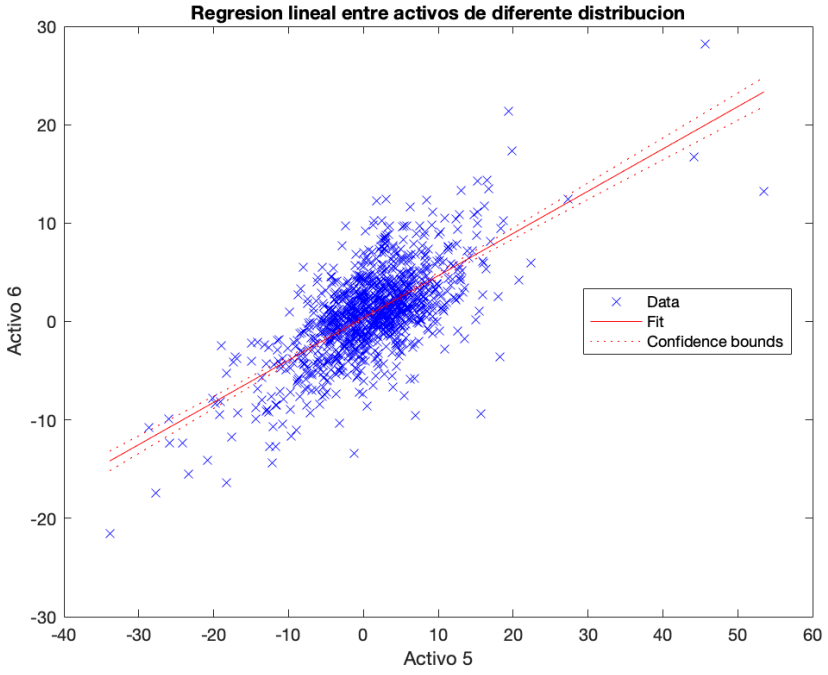


Figura 5: Modelo de Regresión entre los activos 5 y 6

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)    0.36754      0.10337    3.5556     0.00039361
    x1             0.42926     0.014274    30.074    2.4231e-144


Number of observations: 1067, Error degrees of freedom: 1065
Root Mean Squared Error: 3.34
R-squared: 0.459,  Adjusted R-Squared: 0.459
F-statistic vs. constant model: 904, p-value = 2.42e-144
```

Figura 6: Modelo de Regresión

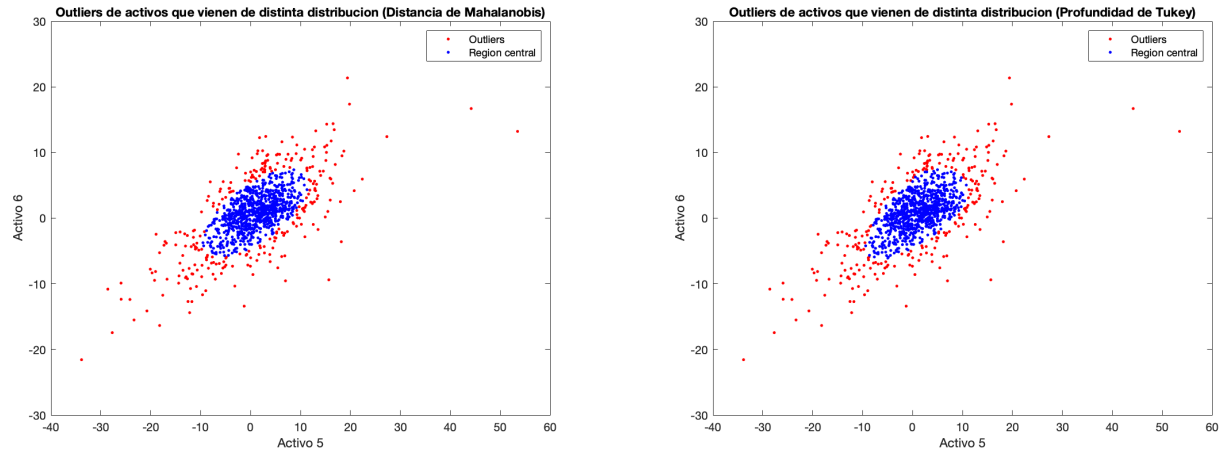
Como podemos observar en la Figura 6, se obtuvo que ambos coeficientes estimados (β_0, β_1) son significativos, es decir, sus p-values son menores a 0.05. También se evidencia que el ajuste del modelo es muy bajo, puesto que da 0.459, un ajuste muy alejado de 1. Este resultado era de esperarse dado que eran los activos que no habían pasado el test de WMW, es decir, eran un par de activos con distribución distinta.

2.2. Eliminar outliers con medidas de profundidad estadística

Adicional a la profundidad de Tukey, decidimos utilizar la distancia de Mahalanobis para para eliminar el 25% de los datos menos profundos. Estas distancias son calculadas de la siguiente forma:

$$d_M(\vec{x}, \vec{y}; Q) = \sqrt{(\vec{x} - \vec{y})^\top S^{-1} (\vec{x} - \vec{y})}. \tag{5}$$

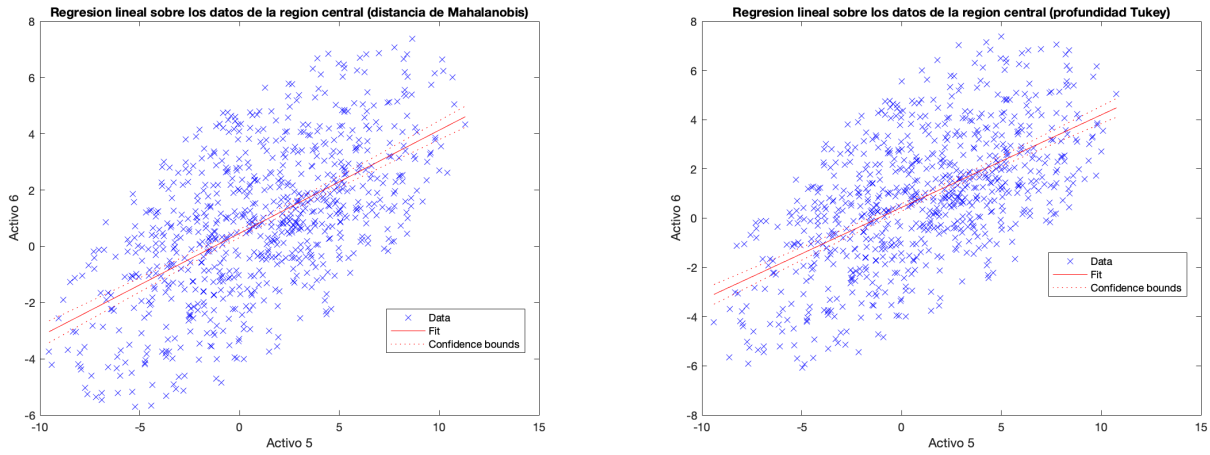
Donde S es la matriz de covarianzas.



(a) Distancia de Mahalanobis (b) Profundidad de Tukey

Figura 7: Identificación de outliers con diferentes medidas

Una vez obtenida la región central por ambas medidas de profundidad mostrada en la Figura 7, procedemos a construir un modelo regresión sobre este.



(a) Distancia de Mahalanobis (b) Profundidad de Tukey

Figura 8: Modelos de regresión lineal con regiones centrales obtenidas mediante diferentes métodos de identificación de outliers

	Modelo normal	Distancia de Mahalanobis	Profundidad de Tukey
Ajuste R2	0.459	0.345	0.360

Tabla 2: R2 ajustado de modelos de regresión de las regiones centrales

Como evidenciamos en la Tabla 2 el ajuste del R2 de la regresión en las regiones centrales por Mahalanobis y Tukey disminuye en comparación a todos los datos. Esto es debido que disminuye mucho la relación lineal entre ambos activos y la región construida entre ellos es circular y un poco dispersa.

2.3. Eliminar outliers con residuales

En un modelo de regresión, los residuales están dados por la siguiente fórmula

$$\epsilon = y - \hat{y}, \quad \hat{y} = \beta_0 + \beta_i * x_i + \epsilon_i \tag{6}$$

Una vez construido el modelo de regresión se calculan los residuales y se van eliminando uno a uno las observaciones con mayor residual, es decir, se elimina el mayor residual, se vuelve a hacer la regresión y se vuelve a eliminar el mayor residuales, así hasta haber eliminado el 25% de los datos.

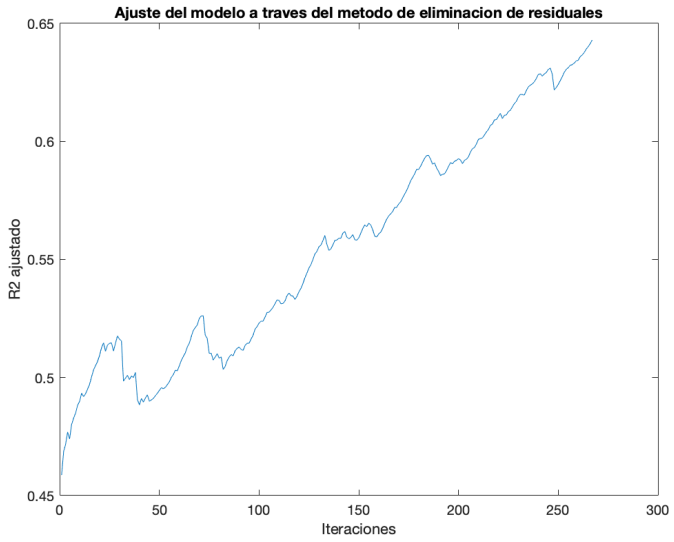
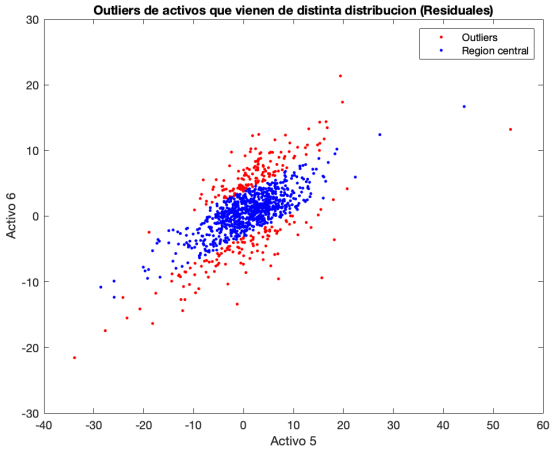
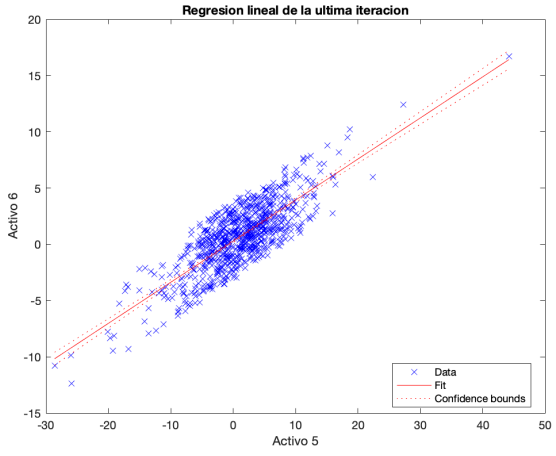


Figura 9: R2 ajustado de las iteraciones

En la Figura 9 observamos la variación de del R2 ajustado a medida que se van eliminado los mayores residuales. Este resultado era de esperarse dado que a medida que se van eliminando los mayores residuales, la relación lineal entre la variables iría incrementando.



(a) Outliers identificados



(b) Modelo de Regresión lineal en la ultima iteración

Figura 10: Regresión lineal con el método de eliminación del mayor residual

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE       tStat       pValue
                   ________    ________    ______    __________

    (Intercept)    0.27387     0.062704    4.3677     1.4212e-05
    x1             0.36503     0.009615    37.965    4.6155e-181


Number of observations: 801, Error degrees of freedom: 799
Root Mean Squared Error: 1.75
R-squared: 0.643,  Adjusted R-Squared: 0.643
F-statistic vs. constant model: 1.44e+03, p-value = 4.62e-181
```

Figura 11: R2 ajustado de la ultima iteración del método de eliminación del mayor residual

En la Figura 10 se muestra cómo queda construido el modelo de regresión paramétrico por eliminación de registros con mayor residual, donde se evidencia mayor relación lineal entre ambas variables. En la Figura 11 observamos el modelo una vez eliminados todos los estos registro y vemos cómo incrementó significativamente el R2 ajustados, pasando de 0.459 a 0.643.

2.4. Regresiones lineales robustas y no paramétricas

Para hacer las regresiones lineales no paramétricas tomamos en cuenta las fórmulas para hacer la estimaciones de los coeficientes

$$\begin{aligned} \beta_1 &= (X'X)^{-1}X'Y = \Sigma_{XX}^{-1}\Sigma_{XY} \\ \beta_0 &= \bar{Y} - \beta_1 * \bar{X} \end{aligned} \tag{7}$$

Donde se estima una versión robusta de la matriz de varianzas y covarianzas. Los métodos utilizados para la estimación de esta se representan en la siguiente tabla.

	Parametrico	Fast MCD	Kendall	Spearman	MCD y Spearman	Shrinkage
Covarianza	Σ_{XY}	[1]	$\rho^k\sigma(x)\sigma(y)$	$\rho^s p\sigma(x)\sigma(y)$	$\Sigma_{XX} = fastMCD$ $\Sigma_{XY} = \rho^s p\sigma(x)\sigma(y)$	Rao-Blackwel

Tabla 3: Estimación de la matriz de covarianzas para cada método

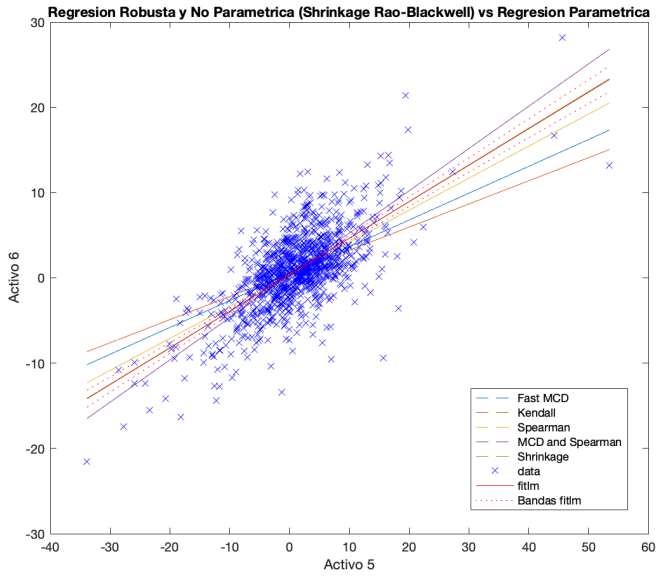


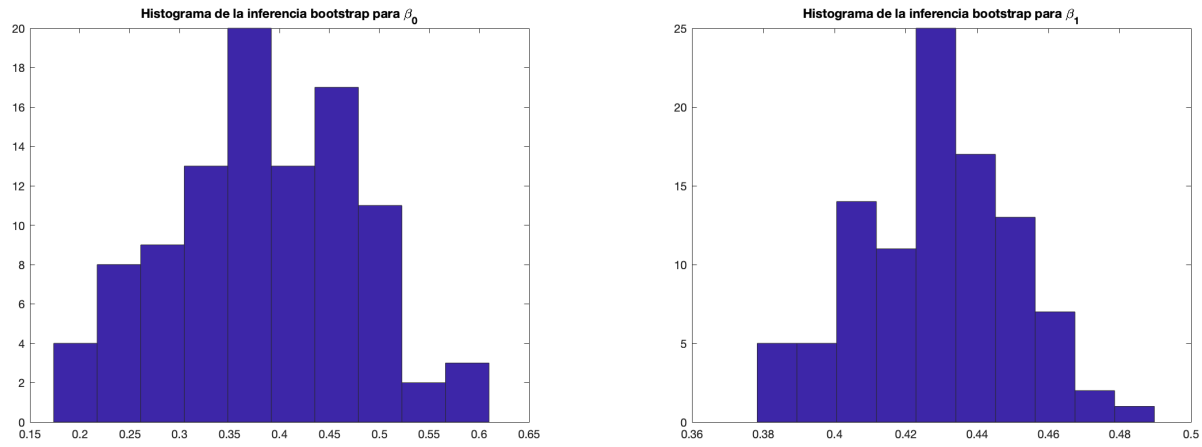
Figura 12: Regresiones robustas

Método de regresión	R2 ajustado
Paramétrico (fitlm)	0.459
Fast MCD	0.427
Kendall	0.397
Spearman	0.452
MCD y Spearman	0.448
Shrinkage	0.459

Tabla 4: R2 ajustado el tipo de regresión

Se puede observar como en la tabla 4 las regresiones robustas nunca dan un R2 mejor al parámetrico. Esto se puede esperar a que pase, porque en este caso los outliers contienen información, y en general como los datos no parecen tener una relación lineal entonces se dificulta un poco mas la regresión lineal, por lo que en general se espera que no tengan un mejor R2 las técnicas robustas, pues estas pueden llegar a perder información.

2.5. Intervalo de confianza para la regresión lineal



(a) Histograma β_0 (b) Histograma β_1

Figura 13: Histogramas de los coeficientes estimados por regresión paramétrica por medio de inferencia Bootstrap

En la Figura 13 vemos los histogramas de los coeficientes por medio de inferencia bootstrap con 100 repeticiones para el modelo de regresión paramétrico, por medio de estos se puede interpretar los intervalos de confianza, donde se evidencia que ambos coeficientes son significativos, dado que estos intervalos no contienen al cero.

Los intervalos de confianza para los modelos de regresión robustos y no paramétricos se muestran a continuación.

Método de regresión	β_1	β_2
Paramétrico (fitlm)	0.392, 0.463	0.145, 0.528
Fast MCD	0.225, 0.690	0.261, 0.371
Kendall	0.246, 0.304	0.346, 0.749
Spearman	0.329, 0.404	0.233, 0.614
MCD y Spearman	0.417, 0.575	0.099, 0.536
Shrinkage	0.392, 0.463	0.145, 0.528

Tabla 5: Intervalos de confianza para β_1 y β_0 usando inferencia bootstrap

En los intervalos de la tabla 5 intervalos lo primero que podemos observar es que en ninguna parte se observa que se contenga un cero, esto entonces nos indica preliminarmente que en ninguna de las regresiones se negaba

la influencia de la variable de explicacion X. Aunque hay intervalos que son muy cercanos a cero por lo que podemos decir que la variable con la que esperabamos explicar a Y no tiene mucha significancia sobre este.

3. XY

Suponga que $x^a = \left(x_1^a, \ldots, x_{n_a}^a\right)$ son n_a realizaciones de una uniforme en el intervalo $[6, 8]$ y $x^b = \left(x_1^b, \ldots, x_{n_b}^b\right)$ son n_b realizaciones de una uniforme en el intervalo [2, 10]. Considere la siguiente mixtura de normales

$$Y \sim 0.2N\left(-2x^a + 10, 1\right) + 0.8N\left(2x^b + 4, 1\right)$$

$$X \sim 0.2U[6, 8] + 0.8U[2, 10]$$

3.1. Densidades

Considerando un $n_a = 10000$ y un $n_b = 5000$ se realizan las mixturas y se estiman las densidades de estas para comprobar. Estas densidades se muestran en la siguiente figura.

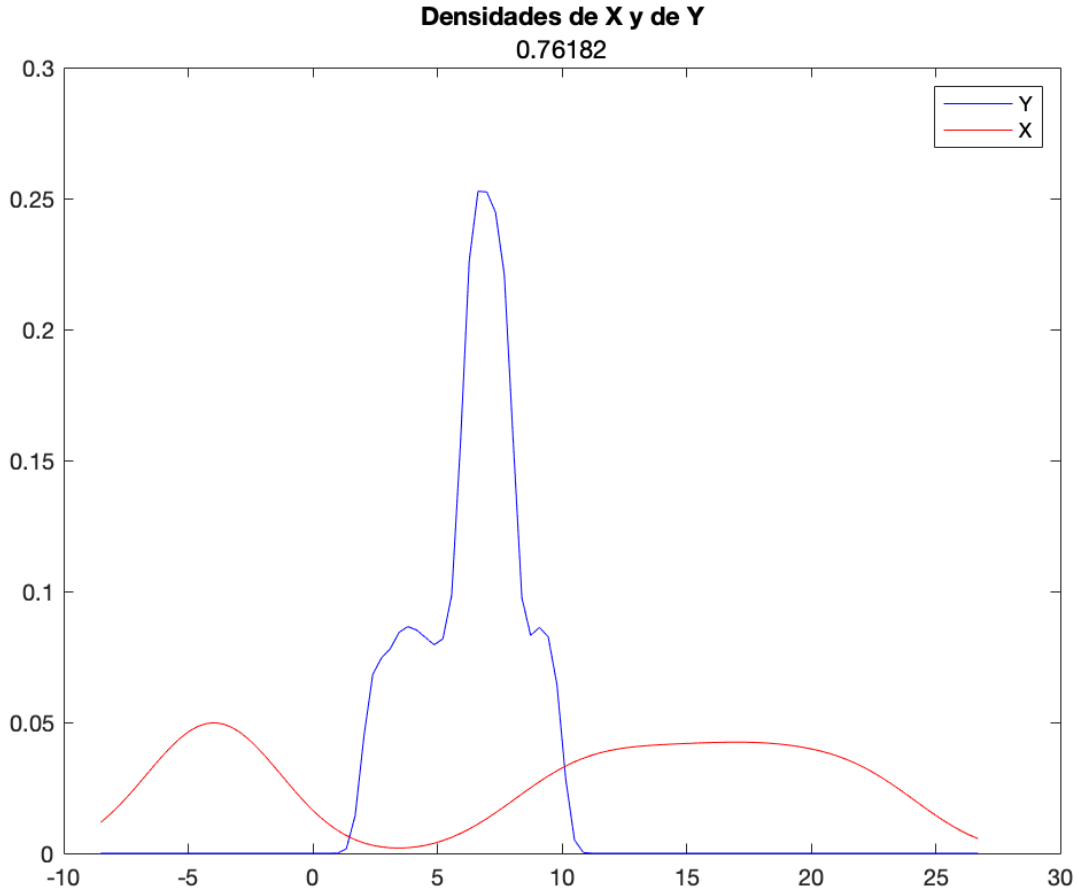


Figura 14: Densidades XY

3.2. Modelo de regresión

Una vez obtenidas las mixturas, realizamos un modelo de regresión lineal paramétrica sobre estos. El resultado de esta se puede observar en la Figura 15. Debido a que hay dos cúmulos de datos muy separados, para el modelo lineal es muy difícil obtener un buen ajuste. Por eso, se le aplican métodos de eliminación de outliers para intentar mejorarlo.

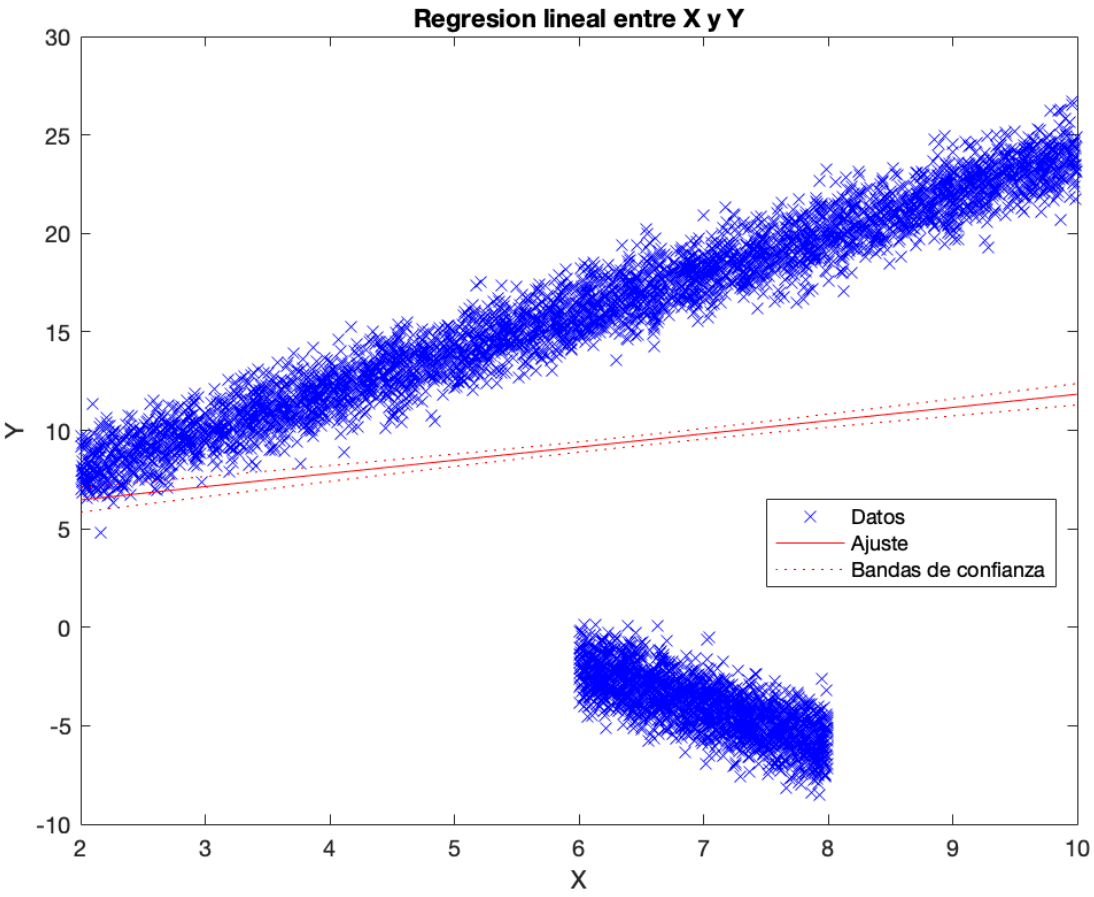


Figura 15: Regresión XY

3.3. Profundidad

La profundidad de Tukey se define como el punto z con la minima masa de probabilidad cargaad por cualquier semiespacio conteniendo z. Esto es:

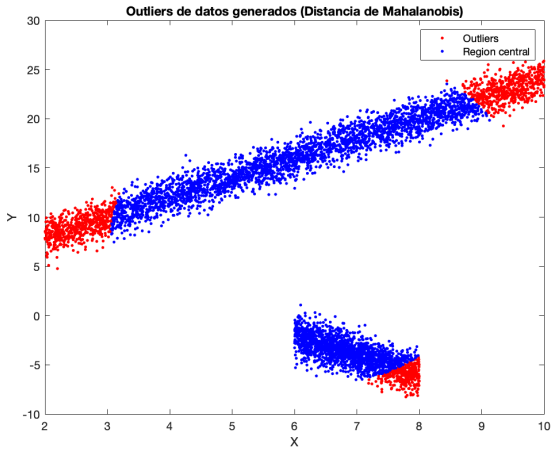
$$D(z \mid P) = \inf_{u \in \mathcal{S}^{p-1}} P\left(u^\top X \leq u^\top z\right)$$

Donde $\mathcal{S}^{p-1} = \{v \in R^p : \|v\| = 1\}$. Para n p-dimensiones observaciones. $\mathcal{X}^n := \{X_i\}_{i=1}^n$, su versión muestral es correspondientemente:

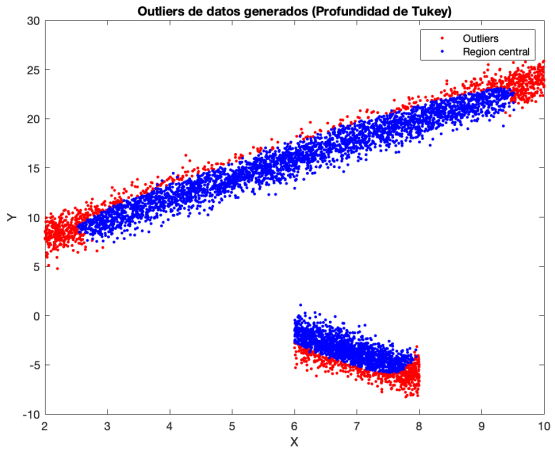
$$D_n(z) := D\left(z \mid P_n\right) = \inf_{u \in \mathcal{S}^{p-1}} P_n\left(u^\top X \leq u^\top z\right)$$

donde P_n denota la distrubucion emperica de $\mathcal{X}^n$.

Se haya la profundidad de Tukey de cada dato y la distancia de Mahalanobis. Luego, se remueven el 25% de los datos (en el caso de Tukey, serán los menos profundos, es decir aquellos que tengan menor profundidad de Tukey y en el caso de Mahalanobis aquellos con mayor distancia). Esto se muestra en la Figura 16:



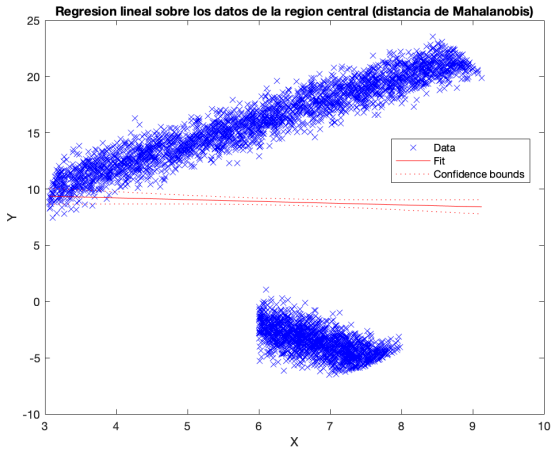
(a) Distancia de Mahalanobis



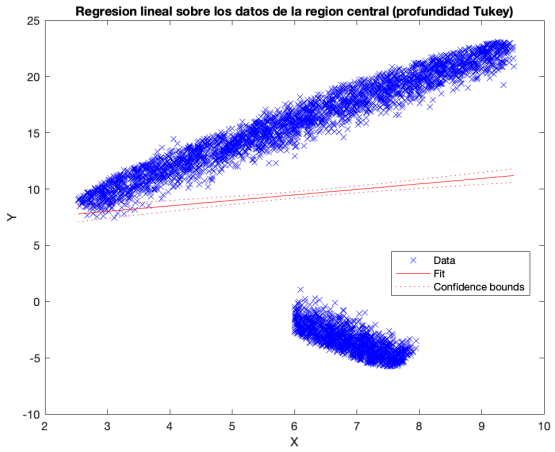
(b) Profundidad de Tukey

Figura 16: Identificación de outliers con diferentes medidas

Luego, similarmente a la sección pasada, realizamos una regresión lineal a los datos sin los outliers (los datos de la región central). Esto se muestra en la Figura 17:



(a) Distancia de Mahalanobis



(b) Profundidad de Tukey

Figura 17: Modelos de regresión lineal con regiones centrales obtenidas mediante diferentes métodos de identificación de outliers

Visualmente, podemos observar que la línea del ajuste se alejó de la región de arriba y se centralizó un poco. Si la región de abajo fueran outliers, estos no están siendo identificados correctamente por este método. Para verificar que tanto mejoró o empeoró la estimación, incluímos los $R2$ ajustados de cada conjunto de datos (completos, removiendo outliers con la distancia de Mahalanobis y con la profundidad de Tukey). Estos se muestran en la Tabla 6:

	Modelo normal	Distancia de Mahalanobis	Profundidad de Tukey
Ajuste R2	0.459	0.345	0.360

Tabla 6: R2 ajustado del modelo de regresión en la región central.

La tabla confirma lo que se percibe visualmente, y es que el modelo de regresión lineal captura mejor el comportamiento con los datos originales que con los datos recortados (debido al comportamiento de los mismos).

3.4. Residuales

Repetimos el análisis de residuales anterior para el conjunto de datos generados (X, Y). En la Figura 18 vemos como mejora el ajuste de la regresión lineal con cada iteración de remover aquel dato con mayor valor en su residual.

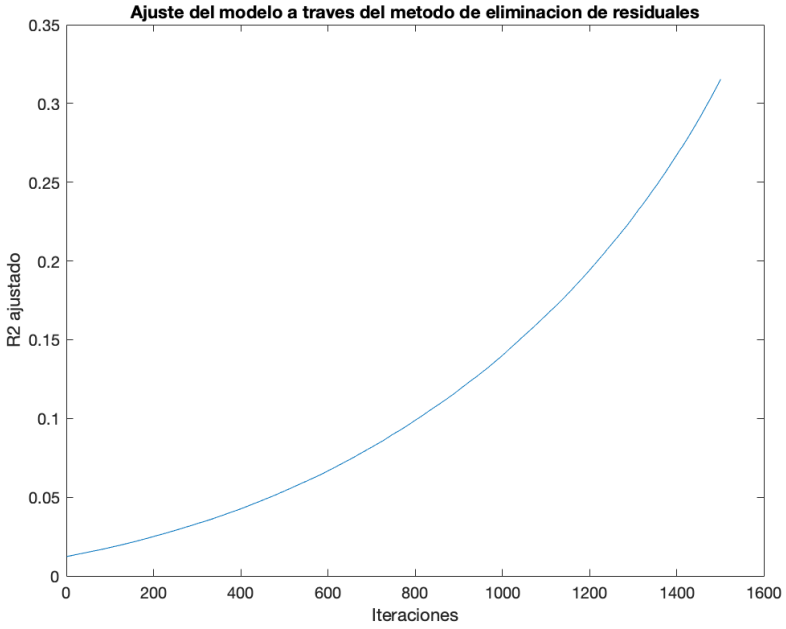
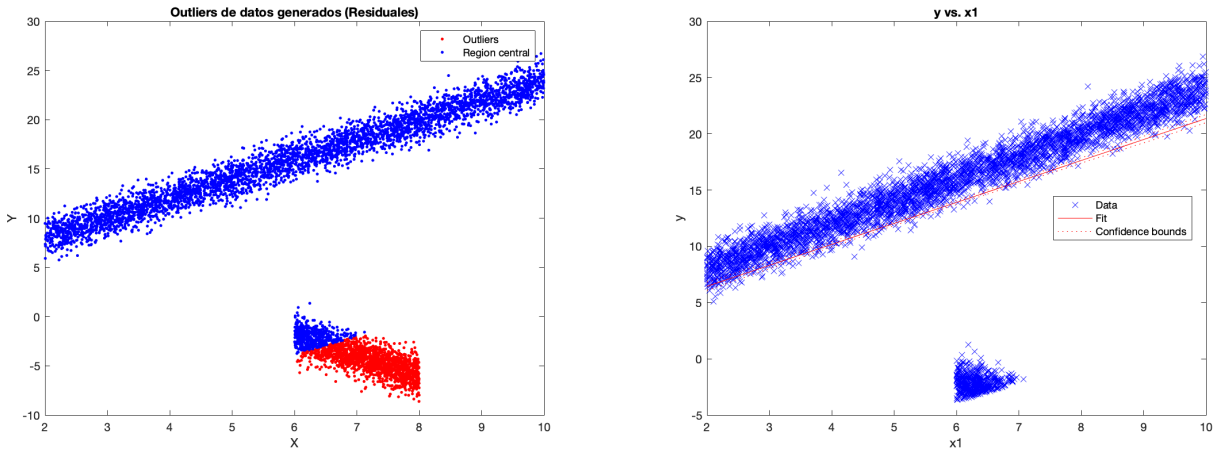


Figura 18: R2 ajustado de las iteraciones de eliminación de outliers con residuales para los datos generados (X,Y)

Podemos observar en la figura que el método ayudó mucho a este modelo ya que paso de tener un ajuste menor a 0.05 a un ajuste un poco mayor a 0.3. En la Figura 19a observamos los datos que el método considera como outliers de color rojo, y los que considera como la región central en azul. Si fuéramos a remover más que el 25% de los datos, se esperaría que el método capturara completamente el cúmulo de datos de abajo, es decir aquellos que se considerarían outliers y así se tendría un ajuste mucho más alto.



(a) Outliers identificados mediante el metodo de eliminacion de residuales

(b) Modelo de Regresión para la región central por eliminación de residuales

Figura 19: Eliminación de registros con mayor valor en su residual

El modelo de regresión mostrado en la Figura 19b presenta un R2 ajustado de 0.32 el cual tiene un mayor

desempeño que los mostrados en la Tabla 6. Lo cual nos indica que este método presenta un mejor desempeño que los mostrados por las profundidades de Tukey y Mahalanobis,

3.5. Regresiones lineales robustas y no paramétricas

Para el conjunto de datos generado, se utilizan los mismos datos utilizados anteriormente para los activos. Utilizando diferentes estimaciones de la matriz de covarianzas, obtenemos versiones diferentes de la estimación. Se puede visualizar en la Figura 22b:

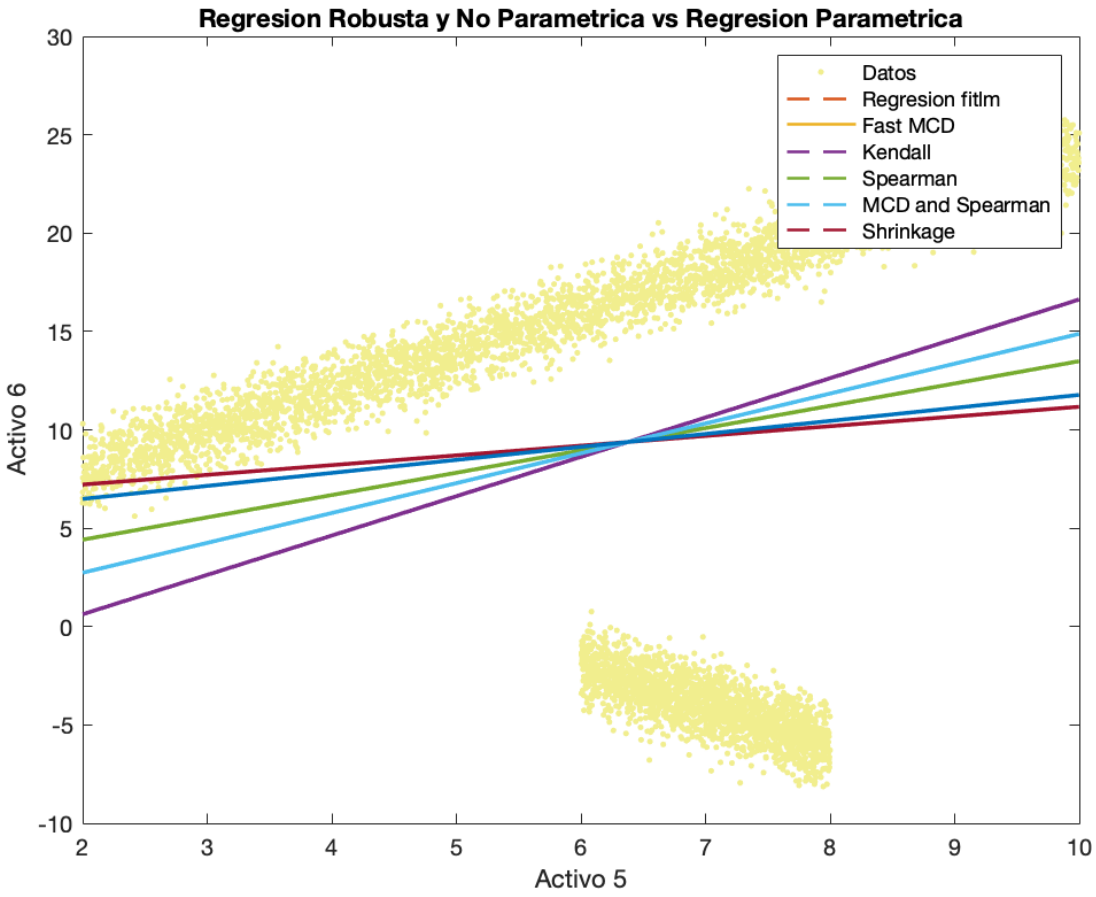


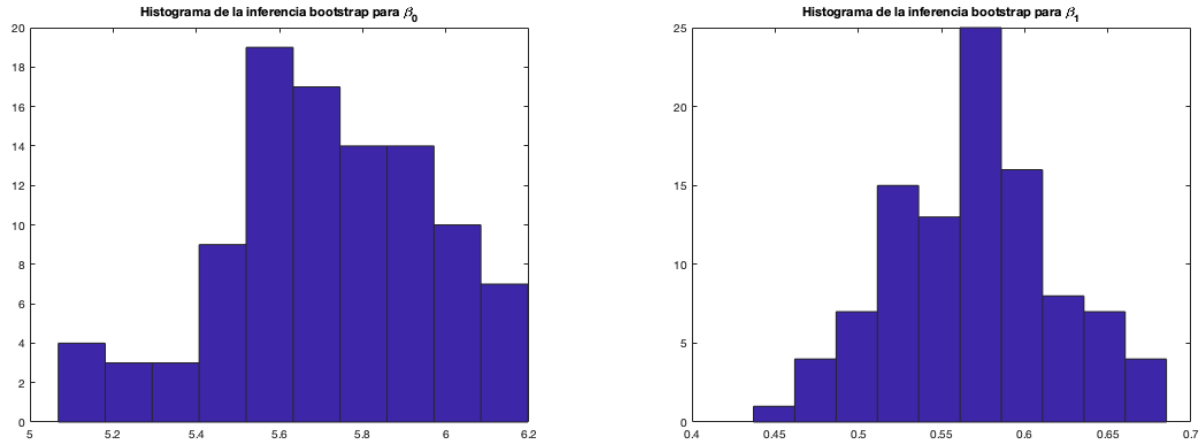
Figura 20: regresiones robustas XY

De aquí podemos observar que la estimación con Kendall logra capturar la misma pendiente a los datos de la región central (el cúmulo de arriba). Sin embargo, su punto de corte con el eje Y está muy desfasado. Sucede algo similar con la combinación de fast MCD y Spearman. Lo contrario pasa con Shrinkage, ya que captura el punto de corte pero no la pendiente. Ninguno le atina al verdadero comportamiento de la región central.
Para entender mejor el comportamiento de cada estimación, en la Tabla 7:

Método de regresión	R2 ajustado
Paramétrico (fitlm)	0.015
Fast MCD	-0.048
Kendall	0.008
Spearman	-0.010
MCD y Spearman	0.015
Shrinkage	0.016

Tabla 7: R2 ajustado según el tipo de regresión

Ninguna en realidad logra un buen ajuste.



(a) Histograma de β_0 paramétrico

(b) Histograma de β_1 paramétrico

Figura 21: Histogramas de los coeficientes por medio de inferencia Bootstrap

Método de regresión	β_0	β_1
Paramétrico (fitlm)	0.480, 0.669	5.189, 6.301
Fast MCD	1.984, 2.009	-3.529, -3.024
Kendall	0.967, 1.181	1.958, 3.012
Spearman	1.264, 1.522	-0.282, 1.155
MCD y Spearman	0.415, 0.528	6.192, 6.491
Shrinkage	0.440, 0.673	5.289, 6.367

Tabla 8: Intervalos de confianza para β_1 y β_0 usando inferencia bootstrap para los datos generados

En los intervalos de la tabla 8 se logra observar como el $beta_1$ de Spearman tiene a cero dentro de su intervalo de confianza, esto entonces nos indicaría que muy posiblemente se rechaza la hipótesis nula de esta regresión lineal y que tiene un nivel de significancia mayor a 0.05. De igual manera, dado a los ajustes tan malos que se obtuvo en la tabla 7, se hubieran esperado ver una mayor cantidad de ceros en los intervalos de confianza.

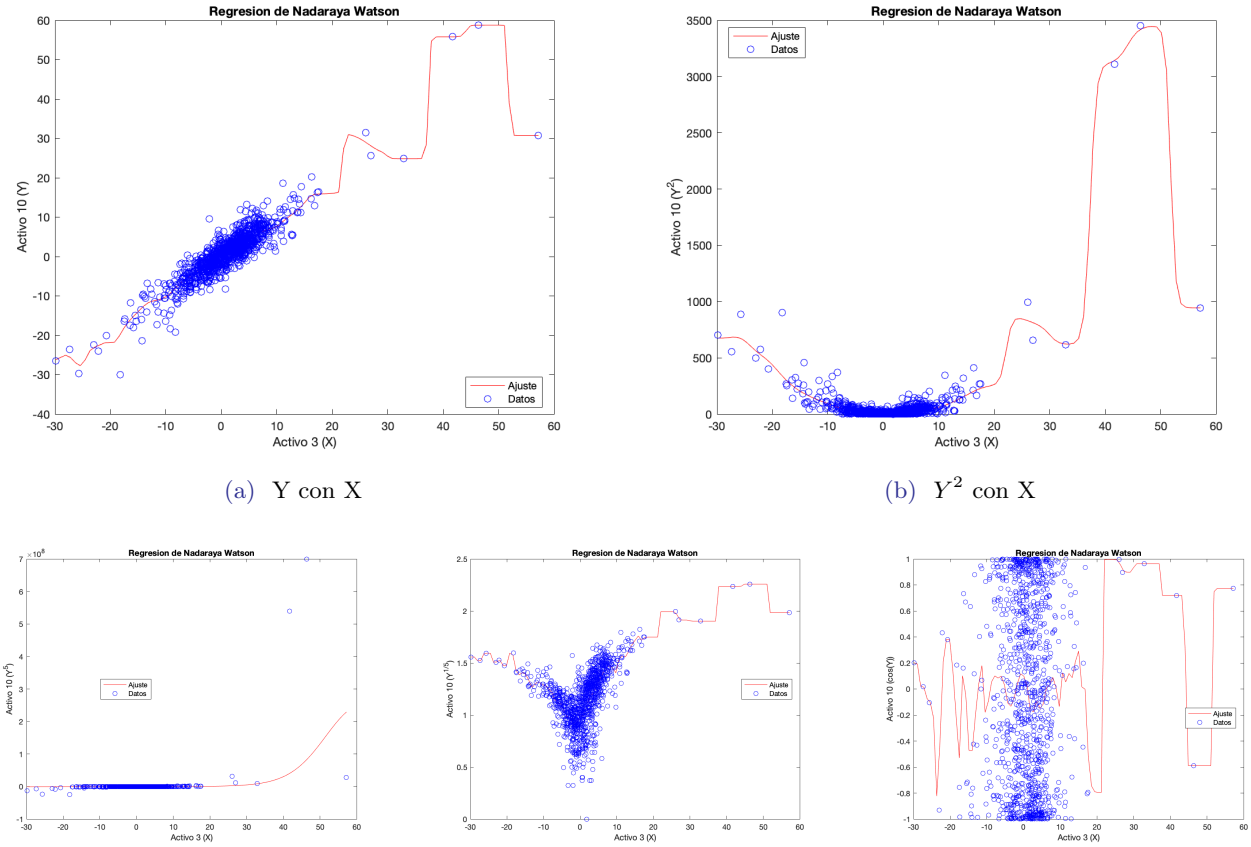
4. Nadaraya Watson

Nadaraya y Watson, en 1964 , propusieron un estimado m como un amedia ponderada local, utilizando un kernel como una funcion de ponderacion. El estimador de Nadaraya-Watson es:

$$\widehat{m}_h(x) = \frac{\sum_{i=1}^{n} K_h\left(x - x_i\right) y_i}{\sum_{i=1}^{n} K_h\left(x - x_i\right)} \tag{8}$$

donde K_h es un kernel con ancho de banda h.

X es el activo 3 y Y es el activo 10



(a) Y con X

(b) Y^2 con X

(c) Y^5 con X

(d) $Y^{1/5}$ con X

(e) $cos(Y)$ con X

Figura 22: Modelos de regresión por Nadaraya-Watson

	(X,Y)	(X,Y^2)	(X,Y^5)	(X,$Y^{1/5}$)	(X, cos(Y))
Bandwidth (h)	1.209	2.199	14.310	0.374	0.571

En la figura 22 se observa como la regresión de Nadaraya-Watson explica el comportamiento de las gráficas. En general es capaz de captar el comportamiento de cada una de las gráficas relativamente bien excepto en la figura 22e. El problema es que se observa mucho overfitting en la regresión, por ejemplo en la regresión de X con Y 22a se observa como la linea de la regresión genera un overfitting solo para poder explicar los datos atípicos que están un poco separados del cumulo de datos centrales.

En la figura 22e se observa como se capta muy vagamente la forma de los datos, pues aunque capta la tendencia sinoidal de los datos no logre encontrar un ajuste apropiado. Esto se debería de esperar especialmente por la naturaleza repetitiva y periódica de la función coseno.

5. Vida depth depth plot

Para elegir las dos familias en las que se partieron los datos se genero un entero aleatorio entre 1 y el tamaño de los datos.

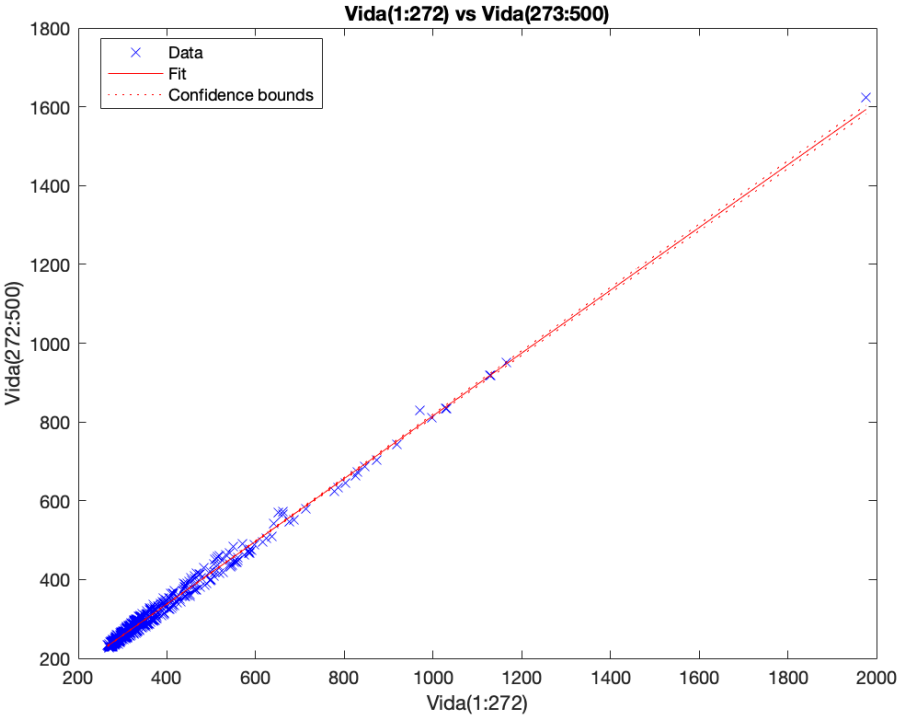


Figura 23: regresiones vida

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    _________    ______    __________

    (Intercept)     19.368        1.4378    13.471    1.7346e-35
    x1             0.79636     0.0035405    224.93             0


Number of observations: 500, Error degrees of freedom: 498
Root Mean Squared Error: 11.9
R-squared: 0.99,  Adjusted R-Squared: 0.99
F-statistic vs. constant model: 5.06e+04, p-value = 0
```

Figura 24: regresiones vida

Se observa como se forma una linea completamente recta como los puntos, lo que generalmente indicaría que los datos son dependientes entre ellos y por fuera de esto también se creería tanto por el p-valor como por la gráfica que estos datos provienen de la misma distribución.

6. Returns depth depth plot

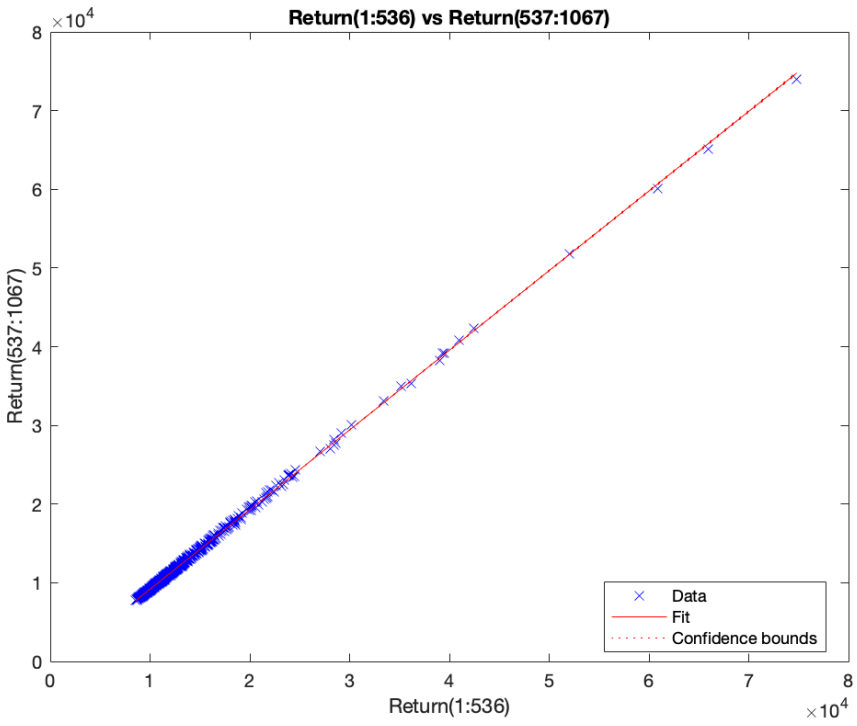


Figura 25: regresiones vida

```
Linear regression model:
    y ~ 1 + x1

Estimated Coefficients:
                   Estimate        SE         tStat     pValue
                   ________    __________    _______    ______

    (Intercept)    -935.18         12.908    -72.452      0
    x1               1.012      0.00097533    1037.6      0


Number of observations: 1067, Error degrees of freedom: 1065
Root Mean Squared Error: 162
R-squared: 0.999,  Adjusted R-Squared: 0.999
F-statistic vs. constant model: 1.08e+06, p-value = 0
```

Figura 26: regresiones vida

Al igual que en la seccion anterior se observa como el ajuste de la regresion a los puntos es muy buena con un nivel de significancia muy alto, esto entonces al igual que en la seccion anterior nos lleva a concluir que posiblemente todos los datos que se tomaron sean identicamente distribuidos.

7. Identificación de poblaciones idénticas para imágenes

Para identificar si dos imágenes provienen de una misma población se puede comparar las imágenes en escala de grises con un test de Kolmogorvo-Smirnov para ver si las imágenes tienen una misma distribución en la escala de grises.

Claramente si dos imágenes son idénticas, exceptuando un ruido leve, entonces van a tener una distribución de

la escala de grises casi idéntica. Por esto es que la distribución de la escala de grises de las fotos es útil para averiguar si dos imágenes provienen de una misma distribución.

Es importante tener en cuenta que cualesquiera dos fotos van a estar sujetas a ruido, entonces aunque vengan de una misma distribución pueden fallar el test. Por tanto este test solo nos da suficiente información para rechazar la hipótesis nula.

References

[1] Peter J. Rousseeuw and Katrien Van Driessen. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, 41(3):212–223, 1999.