UNIVERSIDAD EAFIT

# Técnicas robustas y no parámetricas

Examen Final

Estadística no paramétrica

---


**Profesor:**
Henry Laniado
**Año Académico:**
2022

**Abelino Sepulveda Estrada**

**Luisa Toro Villegas**


---

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

El fichero return.txt contiene rentabilidades de 12 activos. Elimine la primera columna que es la fecha. Construya una etiqueta (1) si media de la fila es positiva y (-1) si la media de la fila no es positiva.

### 1.1. Test de rangos de Mann-Whitney

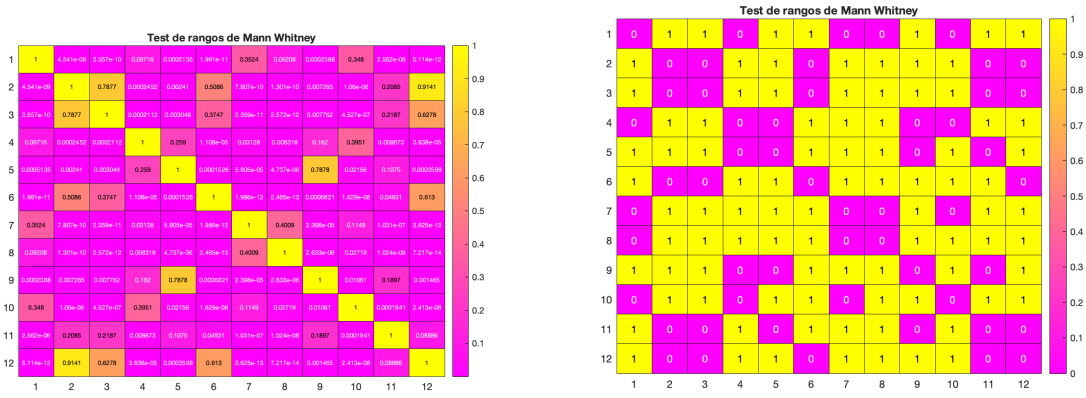
Pase un test no paramétrico de rangos para identificar qué activos de etiqueta (-1) provienen de distinta distribución. Realice una matriz de $12X12$ con el valor 1 si se rechaza H0 y valor 0 si no se rechaza $H_0$. Estime y grafique las densidades en un mismo plano para un par de activos que no pasen el test. ¿Observa congruencia entre el resultado del test y las gráficas de las densidades? Explique el tipo de kernel y el ancho de banda utilizado para las estimaciones.

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 las filas de datos con etiqueta -1

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, la mayoría de los pares de activos vienen de distinta distribución, por lo que también visualizaremos las densidades y las áreas en común entre cada par para escoger los que son más diferentes.

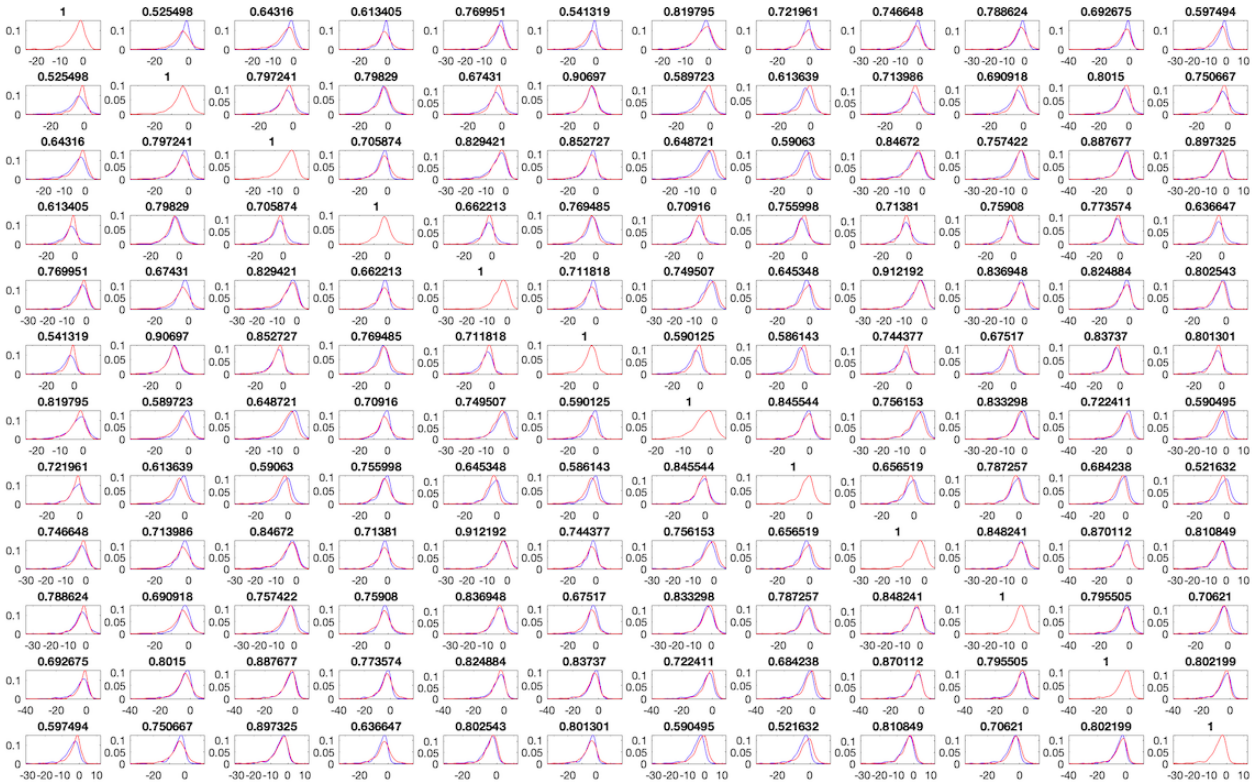


Figura 2: Estimación de las densidades para las filas de datos con etiqueta -1

Según la Figura 2 los activos más diferentes son el 1 y el 2, por eso los visualizamos en la Figura 3. Es evidente que son muy diferentes.

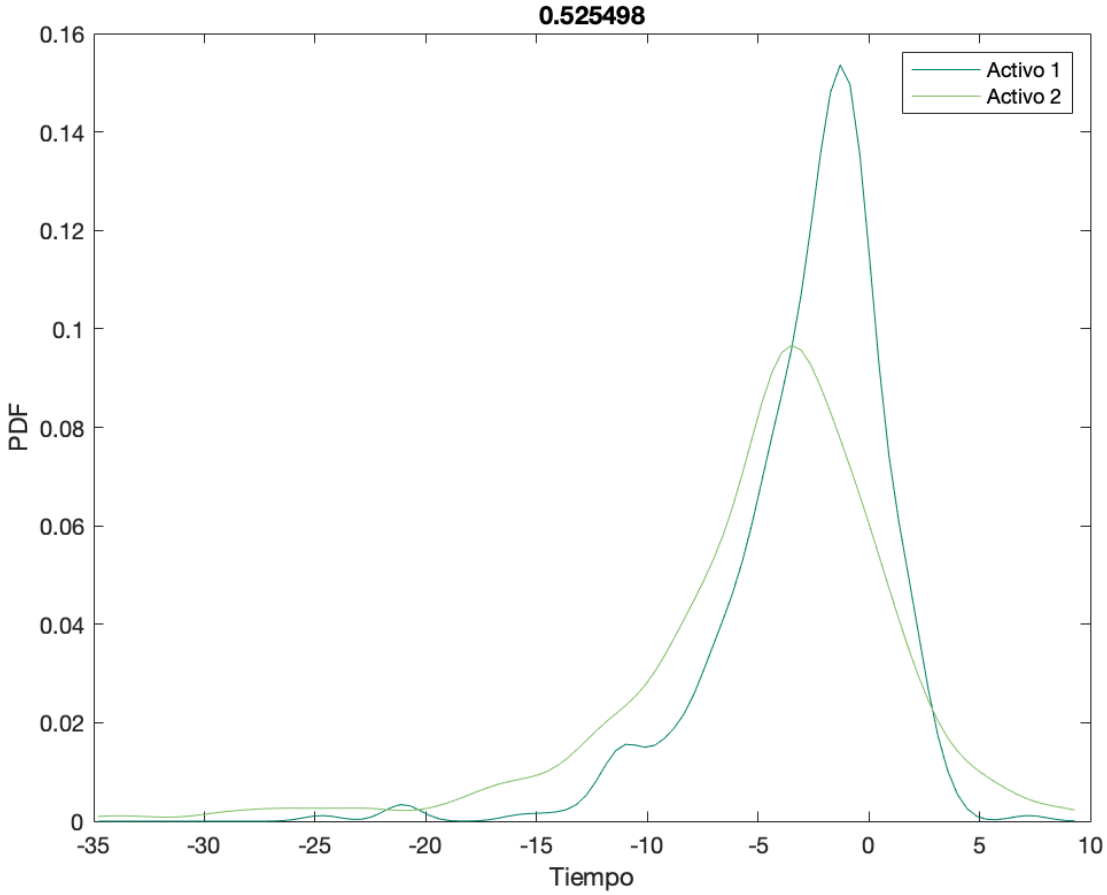



Figura 3: Estimación de las densidades con distribuciones diferentes

Las estimaciones de las densidades mostradas en la Figura 2 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 $\sigma$ se calcula mediante la estimación robusta por medio de MAD:

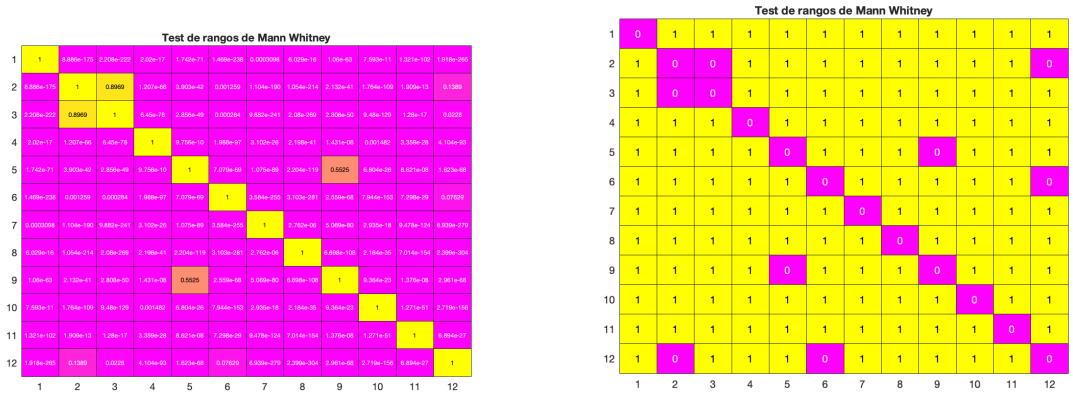
$$\sigma = k * median(|x_i - median(x)|), \quad k = 1.4826$$

| Activo | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| h | 1.027 | 1.716 | 1.406 | 1.640 | 1.254 | 1.616 | 1.313 | 1.470 | 1.254 | 1.289 | 1.465 | 1.360 |

Tabla 1: Bandwidth para cada activo

### 1.2. Datos generados

Genere 10000 datos sintéticos que vengan de la misma población de los datos crudos con etiqueta (-1). Realice el punto anterior con los datos simulados. Hay diferencias en la matriz $12X12$ con del punto anterior.




(a) p-valores

(b) Activos que pasan el test

Figura 4: Test de Mann-Whitney-Wilcoxon para los datos nuevos simulados

Analizando los resultados de la Figura 4 muestra resultados muy diferentes a la anterior. Esto es de esperarse ya que como se generan datos según los activos y no según las observaciones, estas propiedades no se conservan.

### 1.3. Datos transformados

A cada columna del fichero completo, realice la transformación $(xmin)/(max - min) + j/12$. Donde $j$ es el índice de columna. Sobre los datos transformados, realice la gráficas en un mismo plano de la distribución empírica. Explique lo que observa. Pase un test de normalidad sobre la que según la distribución empírica tiene mejores rendimientos. Así no pase el test, estime los para metros de la normal y grafique en un mismo plano la empírica y la teórica, comente los resultados.

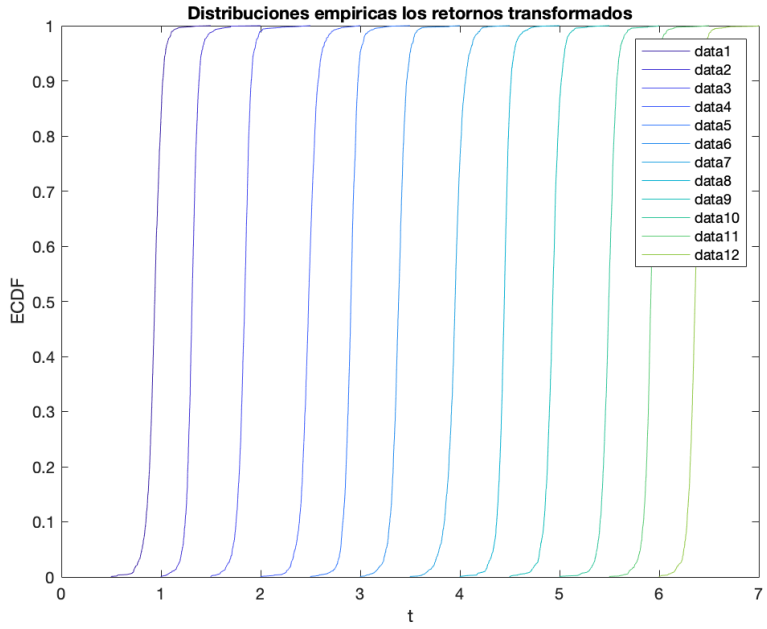



Figura 5: Distribuciones empíricas de las distribuciones empíricas

En la Figura 5 se observan las empíricas de los activos, y mientras mayor índice de activo, más desplazado a la

derecha está (esto por la naturaleza de la transformación). El más desplazado a la derecha será el último activo, el activo 12. Este, a su vez, será el de mayores rendimientos. Por eso, se escoge este para el ánalisis siguiente.

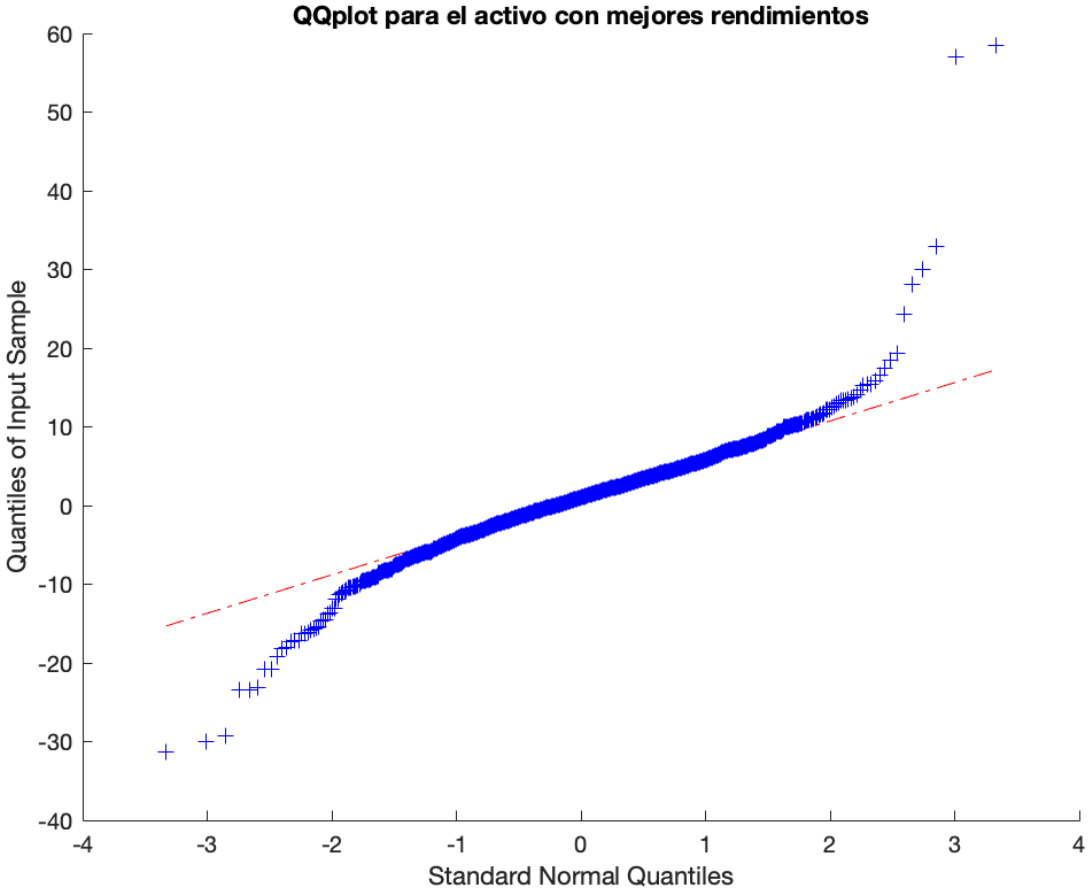


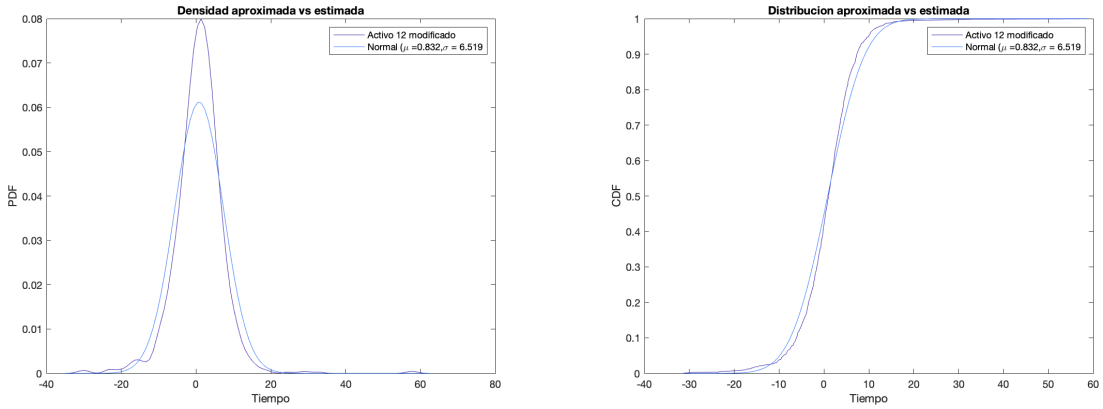

Figura 6: Quantile-quantile plot

En la Figura 6 se muestra la gráfica de cuántil a cuántil donde se verifica la normalidad de una muestra. A pesar de que los datos centrales se comportan como los cuantiles normales, las colas se comportan muy diferentes. Esto nos da indicios se que no es normal, procedemos a realizar otros tests.

| Test | Jarque Bera | Kolmogorov Smirnov |
|---|---|---|
| h | 1 | 1 |
| p-value | 1.0000e-03 | 2.3702e-165 |

Tabla 2: Pruebas de Normalidad

En la Tabla 2 mostrados los resultados del test de Jarque Bera y el de Kolmogorov Smirnov. Ambos tests confirman que el activo 12 no es normal. Dado que la hipótesis nula de ambos test es que el activo sigue un distribución normal, pero en base en los p-valores obtenidos, tenemos información suficiente para rechazar la hipótesis nula.




(a) Densidades (b) Distribuciones

Figura 7: Distribuciones y densidades comparadas con la normal ajustada

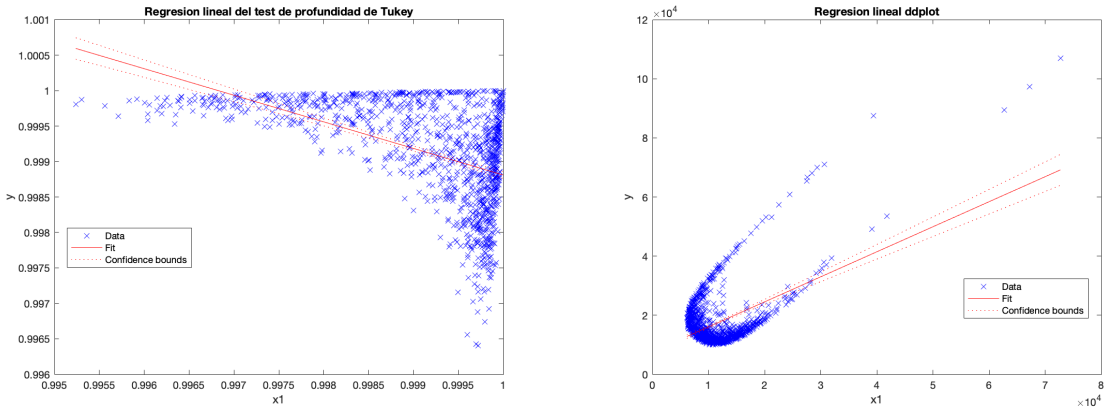
En la Figura 7 se sobreponen la densidad y la distribución de la normal con la del activo 12, con $\mu = 0.832$ y $\sigma = 6.519$ obtenidos a partir de un ajuste a la normal. Aunque en la distribución se ven muy similares, en la densidad se distingue más claramente las diferencias lo cual nos ayuda a confirmar la sospecha en contra de la normalidad.

### 1.4. Test de homogeneidad

Determine estadísticamente si la muestra conjunta original de los datos con etiqueta (-1) vienen de la misma población de la muestra conjunta de los datos con etiqueta (1). Explique con detalle el procedimiento.
Para esto, se consideró el test de tukey y el ddplot. Para el primero se consideran las siguientes pruebas de hipotesis:

$$H_0 : P_x = P_y$$
$$H_1 : P_x \neq P_y$$

Para este mismo se consideraron 200 vectores aleatorios de una distribución uniforme generados en la hiperesfera.




(a) Test de Tukey (b) Depth-depth plot

Figura 8: Tests de homogeneidad entre muestras

Como observamos en la figura 8 en ambas pruebas evidenciamos que ambas muestras provienen de distinta distribución. Es decir, según las pruebas realizadas los datos con etiquieta 1 y etiquetas $-1$ estadísticamente provienen de distinta población. Estos resultados eran de esperarse, dado que estos rendimientos presentaban cambios o valores distintos en esos meses.

### 1.5. Regresión de Nadaraya Watson


Figura 9: Regresión de Nadaraya Watson

### 1.6.

Considere los datos originales, calcule un intervalo de confianza bootstrap para estimar el parámetro $\theta = \frac{asimetria+1}{kurtosis+1}$ de la columna 12. ¿Cómo usaría este intervalo para concluir normalidad o no? Estime el sesgo del estimador.
Se consideran 10000 iteraciones para el Bootstrap para el parámetro $\theta$. El parámetro nos dió $\theta = 0.1106$ y el intervalo de confianza $[0.0476, 0.1410]$. Observamos que el parámetro se encuentra dentro del intervalo, y además no incluye al cero (es significativo).

## 2.

Sobre el fichero original, considere como variable respuesta la columna 12, y como explicativas las otras 11.

### 2.1. Modelo lineal

Contruya un modelo de regresión lineal habitual y explique la significancia del modelo y su ajuste. Exponga el modelo sólo con las variables significativas.
Realizamos el modelo lineal usando diferentes métodos. Para esta realización, se consideró el siguiente modelo:

$$y = \beta_0 + \beta^T X + \epsilon \tag{4}$$

Donde los parámetros $\beta_0, \beta$ se estiman de la siguiente manera:

$$\beta = (X'X)^{-1}X'Y = \Sigma_{XX}^{-1}\Sigma_{XY}\beta_0 = \bar{Y} - \beta'\bar{X}$$

Utilizamos el método de mínimos cuadrados como punto de comparación, y también usamos Pearson, Kendall, Spearman, fastMCD y Shrinkage. En la Tabla 4 se muestran los $\beta$ obtenidos y en la Tabla**??** los ajustes para cada método.
Para la regresión por los métodos robustos anteriormente mencionados, consideramos las siguientes estimaciones de la matriz de covianzas. Representado en la siguiente tabla

| | Parametrico | Fast MCD | Kendall | Spearman | MCD y Spearman | Shrinkage |
|---|---|---|---|---|---|---|
| Covarianza | $\Sigma_{XY}$ | | $\rho^k\sigma(x)\sigma(y)$ | $\rho^s p\sigma(x)\sigma(y)$ | $\Sigma_{XX} = fastMCD$<br>$\Sigma_{XY} = \rho^s p\sigma(x)\sigma(y)$ | Rao-Blackwel |

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

| Método de regresión | OLS | p-value OLS | Pearson | Kendall | Spearman | fastMCD | Shrinkage |
|---|---|---|---|---|---|---|---|
| b0 | -0.15725 | 0.030235 | -0.1895 | -0.2053 | -0.2121 | -0.1924 | -0.1590 |
| b1 | 0.1275 | 0.00088639 | 0.1184 | 0.1089 | 0.1332 | 0.1204 | 0.1263 |
| b2 | -0.01601 | 0.36464 | -0.0168 | 0.0562 | 0.0148 | -0.0010 | -0.0137 |
| b3 | 0.65253 | 1.0027e-69 | 0.6816 | 0.3680 | 0.6070 | 0.5599 | 0.6372 |
| b4 | 0.018149 | 0.26721 | 0.0167 | 0.0551 | 0.0369 | 0.0640 | 0.0203 |
| b5 | -0.13768 | 5.9703e-06 | -0.1290 | 0.0195 | -0.0926 | -0.0626 | -0.1296 |
| b6 | 0.045474 | 0.027216 | 0.0438 | 0.0823 | 0.0689 | 0.0526 | 0.0485 |
| b7 | 0.057918 | 0.012665 | 0.0562 | 0.0560 | 0.0443 | 0.0758 | 0.0573 |
| b8 | 0.004055 | 0.84429 | 0.0043 | 0.0142 | 0.0099 | -0.0195 | 0.0041 |
| b9 | 0.031497 | 0.2756 | 0.0303 | 0.1063 | 0.1056 | 0.1344 | 0.0333 |
| b10 | -0.002787 | 0.90268 | -0.0026 | 0.0142 | -0.0331 | -0.0340 | -0.0028 |
| b11 | 0.2102 | 3.3092e-15 | 0.2193 | 0.1432 | 0.1427 | 0.1335 | 0.2110 |

Tabla 4: Coeficientes estimados ante los distintos modelos de regresión

| Método de regresión | OLS | Pearson | Kendal | Spearman | fastMCD | Shrinkage |
|---|---|---|---|---|---|---|
| Ajuste R2 | 0.865 | 0.8662 | 0.8550 | 0.8634 | 0.8614 | 0.8674 |

Tabla 5: R2 ajustado ante los distintos modelos de regresión

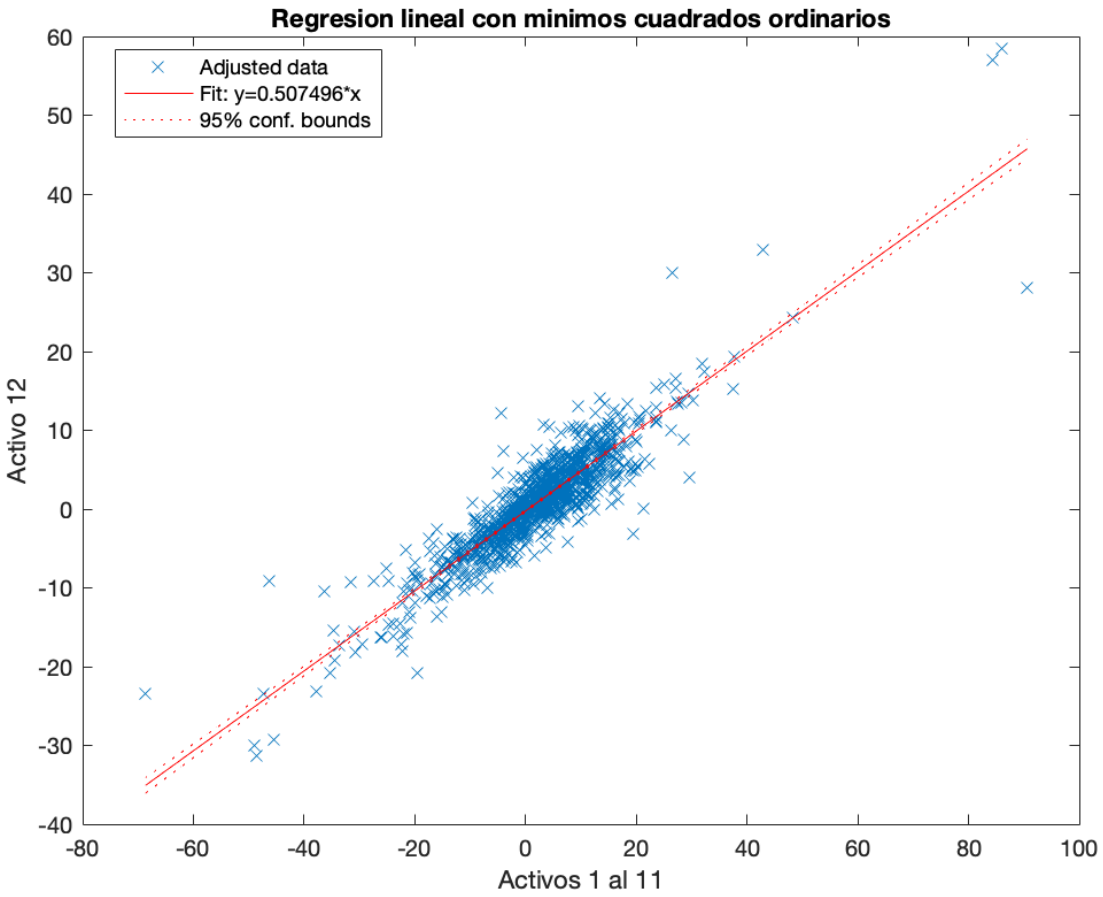



Figura 10: Regresión lineal de los activos 1 al 11 explicando al activo 12

En la Figura 10 observamos el modelo considerando sólo las variables significativas. El ajuste R2 antes daba 0.86 y luego, sin estas variables, da 0.81. Esto se debe ya que al quitar variables explicativas perdemos información, o puede deberse a correlaciones internas de estas.

### 2.2. Eliminación de outliers con la distancia de Mahalanobis

Con todas las variables, utilice la profundidad de Mahalanobis para eliminar uno a uno el 20% de los datos menos profundos. Sobre la región central, realice el punto anterior. Comente los resultados.

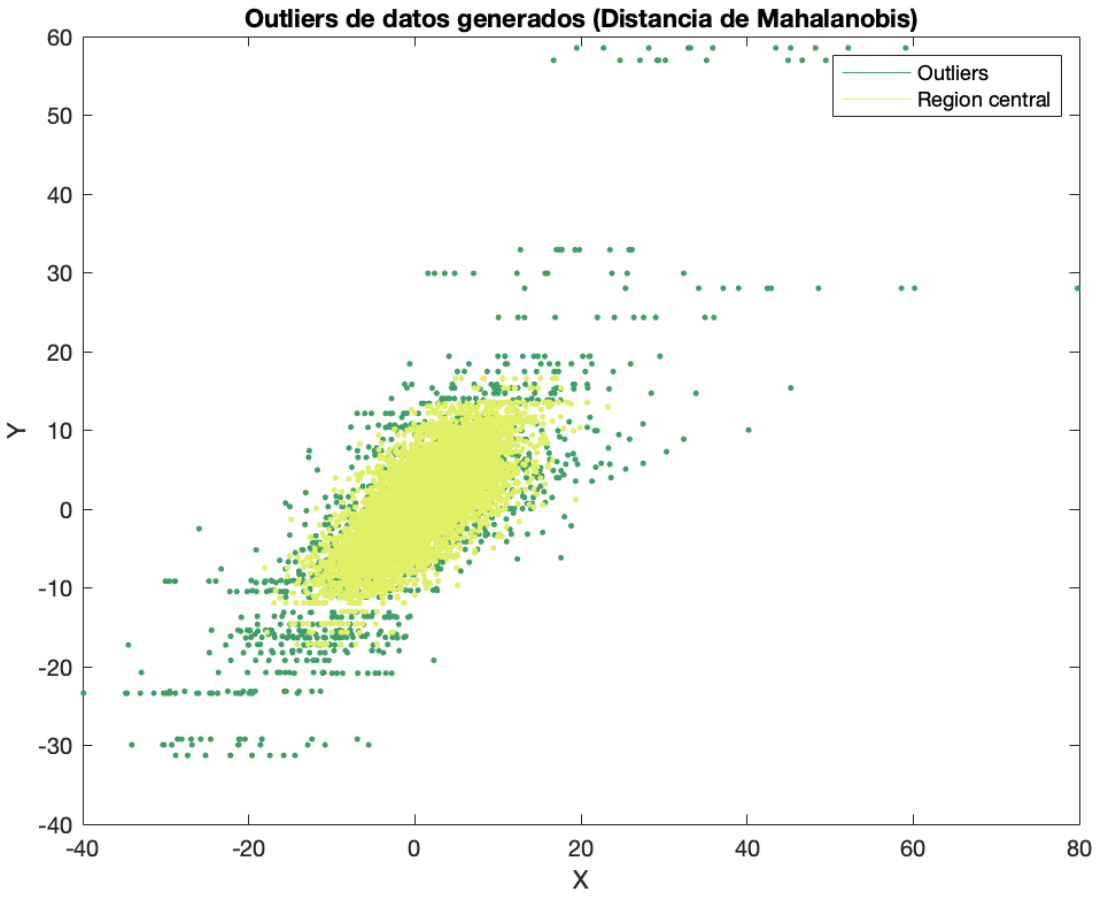



Figura 11: Identificación de outliers con la distancia de Mahalanobis

Luego de remover outliers, se obtiene un ajuste de 0.860. Si hacemos este método iterativamente hasta que obtengamos un ajuste de 0.9, lo cual toma un poco más de 400 iteraciones. Este proceso lo vemos en la Figura 12

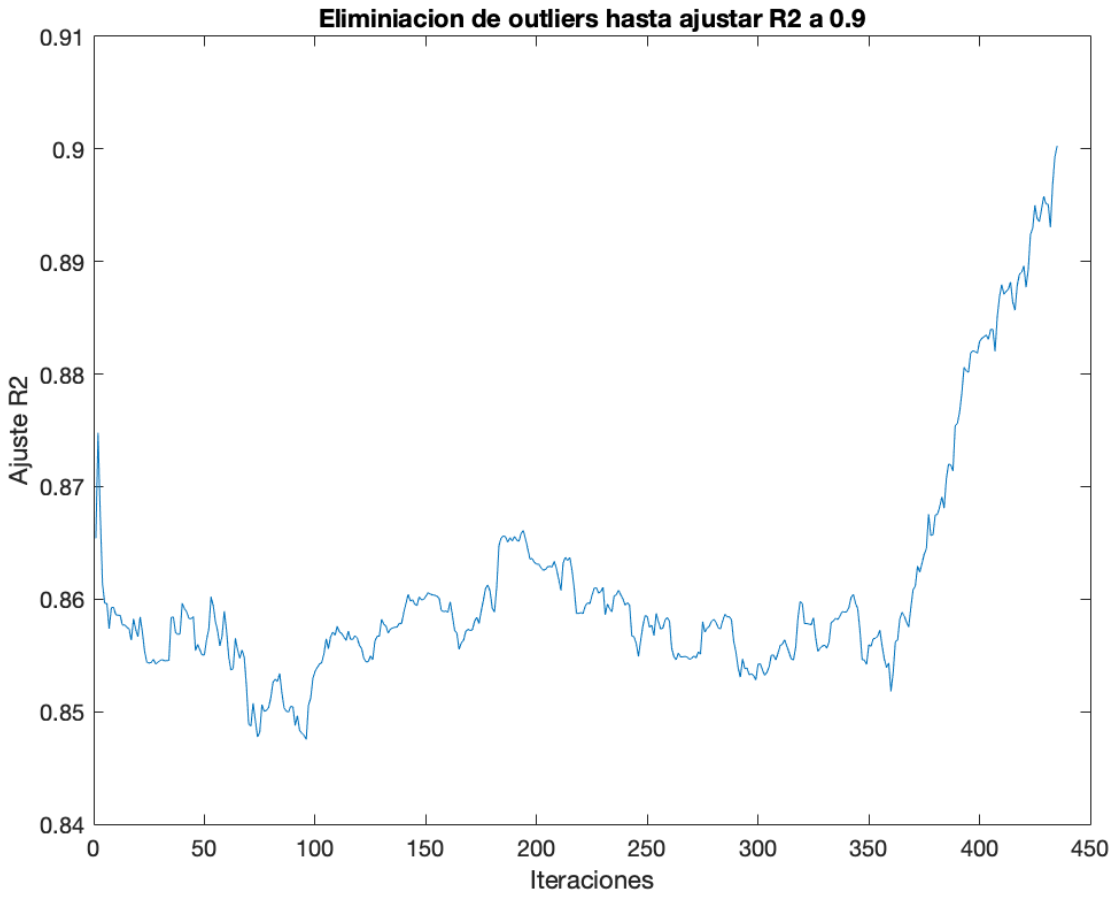



Figura 12: R2 ajustado a través de las iteraciones de eliminación de outliers

### 2.3. Eliminación de outliers con Residuales

Con todas las variables, con el método de eliminación de registros con mayor valor en su residual, elimine registros uno a uno hasta obtener un ajuste al menos del 90%. Sobre la región central construya de nuevo el modelo sólo con variables significativas. Comente los resultados.
En un modelo de regresión, los residuales están dados por la siguiente fórmula

$$\epsilon = y - \hat{y}, \quad \hat{y} = \beta_0 + \beta_i * x_i + \epsilon_i \tag{5}$$

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 20% de los datos.

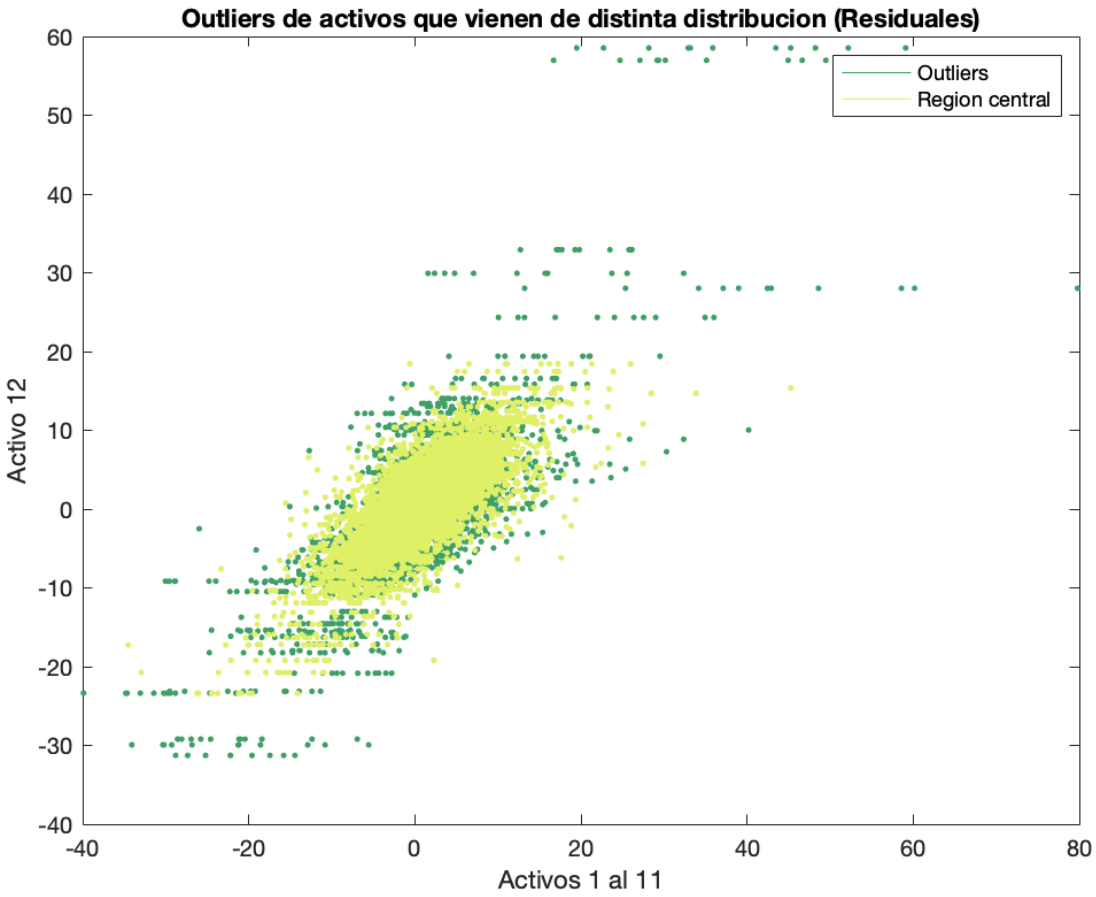



Figura 13: Identificación de outliers con residuales

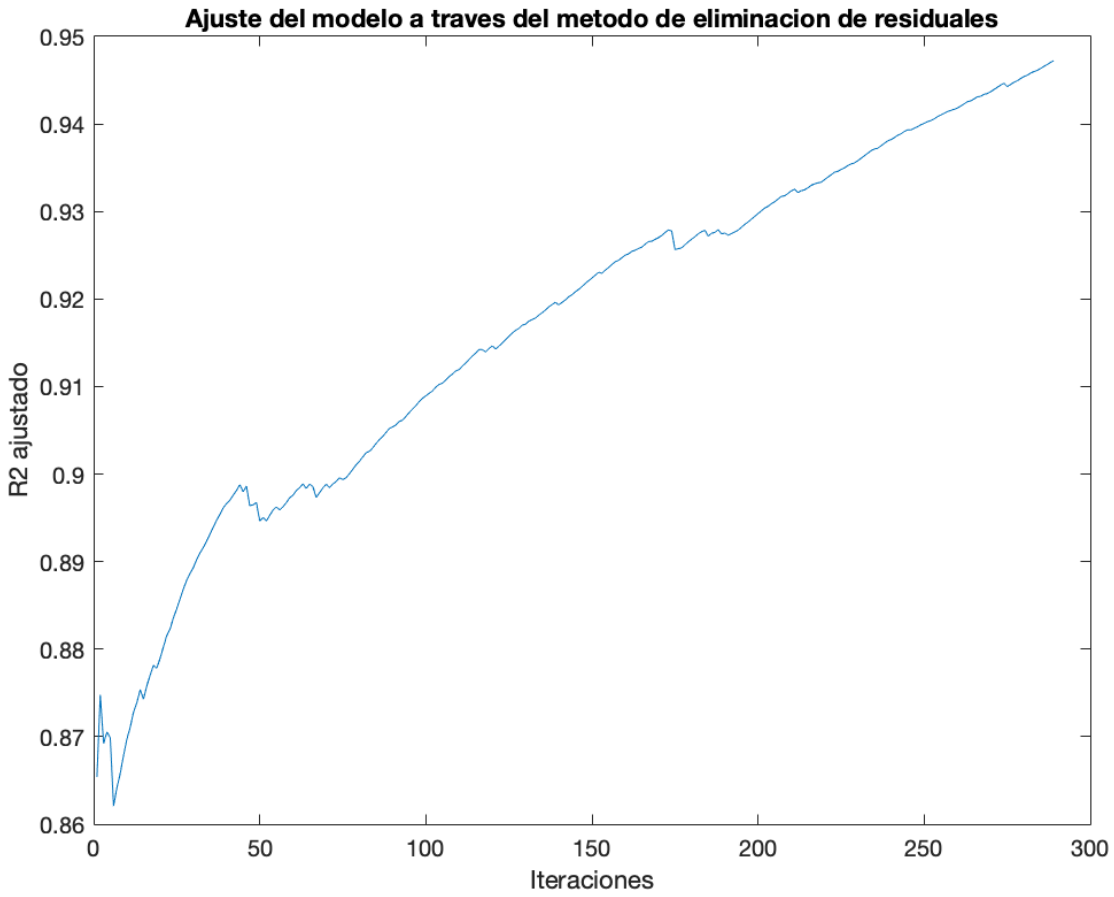



Figura 14: R2 ajustado a través de las iteraciones de eliminación de outliers

En la Figura 14 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.

### 2.4. Regresión robusta

Se realiza la matriz de correlaciones con el promedio de Pearson, Spearman y Kendall. Eso tiene un ajuste de 0.721. Se muestra en la Figura 15.

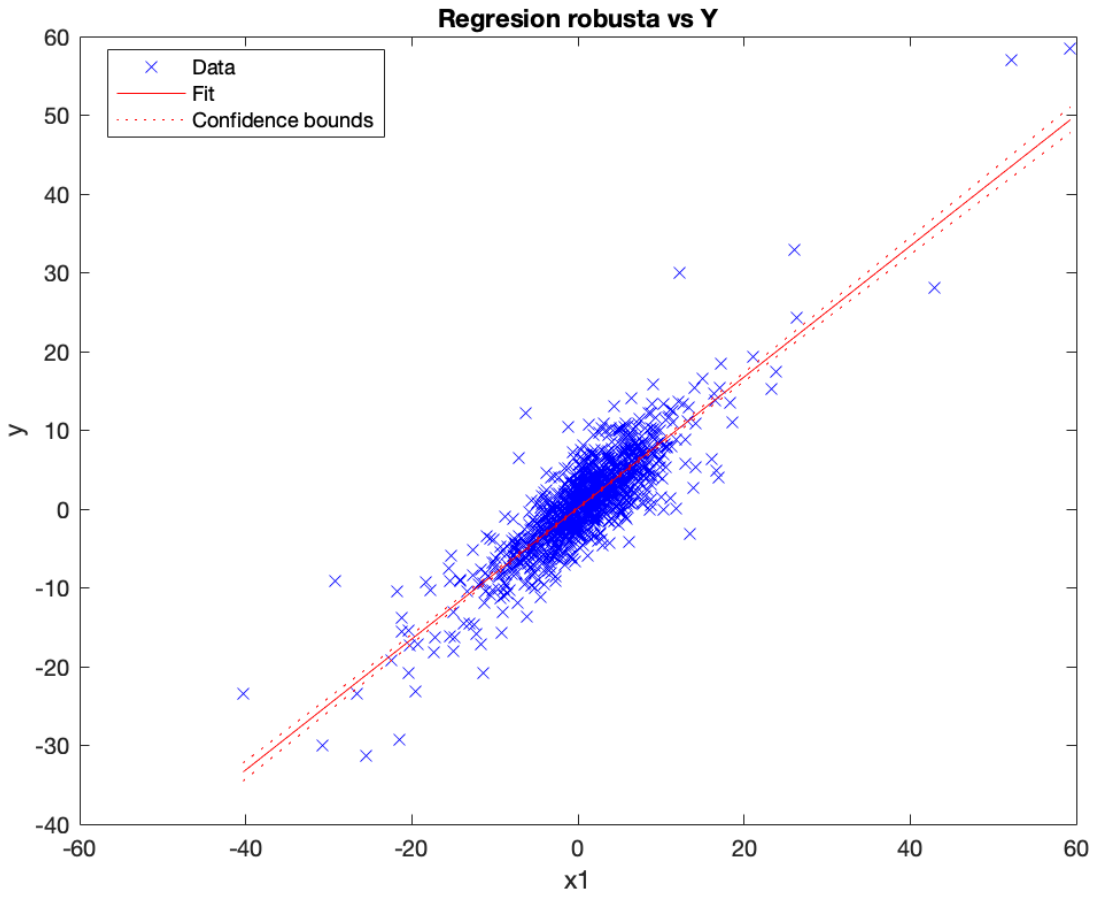



Figura 15: Regresión robusta

Los intervalos de la inferencia bootstrap de $\beta_1$ dan basicamente 0, $^{[}-0.11, -0.11] * e^{-15}$

### 2.5. Comparación entre los métodos anteriores

| | OLS | Mahalanobis | Residuales | Robusta |
|---|---|---|---|---|
| Ajuste R2 | 0.865 | 0.900 | 0.947 | 0.722 |

Tabla 6: Comparación entre métodos

En la Tabla 6 vemos los R2 ajustados para cada método. Escogemos el método de eliminación de outliers mediante residuales ya que, inicialmente tiene un buen ajuste, pero va creciendo con cada iteración. Es un método muy fácil de implementar y en general, da muy buenos resultados. Es un método apropiado combinado con la regresión lineal que ya ajusta los datos longitudinalmente y no transversalmente, como los demás.

## 3.

### 3.1.

Sean $x_1, x_2, \ldots, x_n \sim$ Bernouli(p) y sea $T = \frac{1}{n}\sum_{i=1}^{n}\sqrt{x_i}$ queremos ver que converge en probabilidad a P.
Veamos que $T \xrightarrow{p} p$.
Supongamos que $x_1, \ldots, x_n$ son independientes

$$E(T) = \frac{1}{n}E(\sum_{i=1}^{n}\sqrt{x_i}) = \frac{1}{n}E(\sum_{i=1}^{n}\sqrt{x_i})$$

como $x_i \sim B(p), \sqrt{x_i} = x_i, as,$

$$E(T) = \frac{1}{n} E\left(\sum_{i=1}^{n} xi\right) = \frac{1}{n} np$$

$$\lim_{n\to\infty} E(T) = p$$

$$\mathrm{Var}(T) = \frac{1}{n^2} \mathrm{var}\left(\sum_{i=1}^{n} \sqrt{x_i}\right) = \frac{1}{n^2} \sum_{i=1}^{n} \mathrm{var}\left(\sqrt{x_i}\right)$$

$$= \frac{1}{n^2} \mathrm{Var}\left(\sum_{i=1}^{n} \mathrm{var}\left(x_i\right)\right) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

$$\mathrm{Lim}_{n\to\infty} \mathrm{Var}(T) = 0$$

, por lo tanto $\sqrt{x_i}$ converge en probabilidad a p.

### 3.2. Estadístico ordenado

Calcule la distribución del primer estadístico ordenado para una muestra de n variables aleatorias exponenciales de parámetro 1. Calcule el sesgo teórico de $X[1]$. Genere 1000 variables exponenciales de parámetro 1 y estime el sesgo Jacknife de $X[1]$ y comparelo con el sesgo teórico.
Se generaron 1000 variables exponencias del parámetro 1 y se encontraron los siguientes resultados
ej: $x_1, \ldots, x_n \sim e(\lambda)$ $FX_{[1]}(t) = 1 - (1-t)^n \leftarrow$ estimador natural de 0. $\hat{\theta} = X_{[1]}$ eptimador de $\theta = 0$ veamos el sesgo (bias)

$$E(\hat{\theta}) - \hat{\theta}$$

$$fx_{[1]}(t) = n(1-t)^{n-1}$$

$$E(x[1]) = \int_0^1 tn(1-t)^{n-1} dt$$

$$= n \int_0^1 t(1-t)^{n-1} dt$$

$$= \frac{1}{n+1} \quad \text{cuando } n \to \infty \quad \text{Fz(t)} \ \to 0$$

asintóticamente inseogado.

| Sesgo teorico | Sesgo Jacknife |
|---|---|
| 9.9900e-04 | 0.0015 |

Tabla 7: Comparacion Sesgo Teorico vs Jacknige