EL MODELO LINEAL

Clase 6

Diagnósticos

Sergio Camiz

LIMA - Marzo-Mayo 2025

Asuntos de la clase 6

- Collinealidad
- La falta de ajuste del modelo lineal
- Independencia de las observaciones
- Homocedasticidad
- Testar los residuos
- Puntos aberrantes

Collinealidad

La collinealidad, o sea el exceso de correlación entre regresores, debería ser averiguado antes, debido a que este influye fuertemente sobre la varianza de los estimadores, en cuanto contribuye a esta como $(X'X)^{-1}$. Si hay dependencia lineal, $X'X$ es singular y no se puede invertir, pues $det(X'X) = 0$, pero también si hay collinealidad, o sea algunos regresores están muy correlacionados entre ellos, $(X'X)^{-1}$ puede ser inestable.

En primero lugar tiene que ver la dimensión verdadera del espacio. Esto se ve a través del numero de autovalores no cero de la matriz C de correlaciones entre regresores, que se corresponde a la dimensión del espacio que generan.

```
> eigen(cor(X))$values
```


Si hay q ceros, significa que q regresores dependen de los demás.

Si no hay, el *número de condición*, o sea la razón entre el autovalor de C más grande y el más chiquito $\kappa = \lambda_1/\lambda_n$ informa sobre la linealidad. Si $\kappa \geq 10$ se considera que si hay collinealidad, si $\kappa \geq 30$ esta es grave.

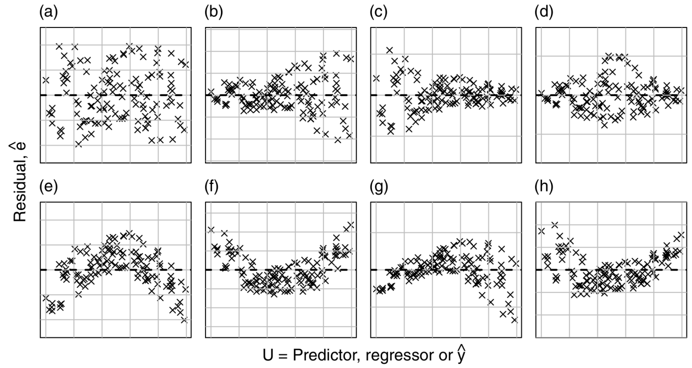
Para detectar cuales son los regresores que causan collinealidad, se hace recurso a la correlación múltiple entre ellos: los que tienen una correlación múltiple igual a -1 o 1 tienen que ser tirados, ya que corresponden a los autovalores cero. Igualmente, si por unos regresores esta se acerca de -1 o 1, estos causan collinealidad, ajuntan poca información, al contrario aumentan la varianza de los estimadores y se pueden borrar.

- Efectivamente, se resulta la relación entre la correlación múltiple $R_{m,i}$ entre cada regresor y los demás y $(c_{ii})_{i=1,\dots,p} = \text{diag}(C^{-1})$, dada para

$$R_{m,i} = R_{i;1,..\hat{i}...p} = \sqrt{\left(1 - \frac{1}{c_{ii}}\right)}$$

- A $c_{ii} = \frac{1}{1-R^2_{m,i}}$ se le llama también *factor de inflación de la varianza* (*VIF*), ya que el desvío estándar de los $\hat{\boldsymbol{\beta}}$ es dado para $\sigma_j = \sqrt{MS_e\, c_{jj}}$.
- Por tanto más grande la correlación múltiple, más grande *VIF* y más grande el desvío estándar del $\hat{\beta}$ correspondiente.
- A su inversa $1/VIF$ se le llama *tolerancia*, alta indicando baja correlación múltiple.

La falta de ajuste del modelo lineal



- En todo lo que precede, se hizo la hipótesis que la relación entre $\boldsymbol{X}$ y $\boldsymbol{y}$ era conocida como lineal o esta era en buena aproximación lineal.
- En realidad es importante de comprobar que la relación entre $\boldsymbol{X}$ y $\boldsymbol{y}$ sea lineal y se necesita por tanto averiguar lo que se llama la *falta de ajuste* del modelo lineal.
- Sabemos que por cada $\boldsymbol{x}_i$, el modelo estima el punto

$$\eta_{\boldsymbol{x}_i} \;=\; E(y_i|\boldsymbol{x}_i)$$

que corresponde al promedio de los $\boldsymbol{y}_i$ que se encuentran en correspondencia del valor $\boldsymbol{x}_i$.

- Si la relación que se trata como lineal no es tal, ningún modelo lineal va pasar para todos los promedios.
- En consecuencia MS_e, el estimador de la varianza disponible para σ^2, que depende del modelo empleado mediando los desvíos por respecto de puntos diferentes del promedio, va sobreestimar la varianza.
- Entonces se trata de estimar la varianza de los $\boldsymbol{y}_i$ de otra manera independiente y comparar las dos.
- Para una medida alternativa se necesita que a por lo menos uno de los $\boldsymbol{x}_i$ se corresponden por lo menos dos medidas y_{i1} y y_{i2}, aunque para una buena estimación claro que sería preferible conocer diversos valores y_{ik} por cada $\boldsymbol{x}_i$.

- Supongamos entonces haber elegido $m > 3$ vectores $\boldsymbol{x}_i, i = 1, 2, ..., m$ y por cada uno haber medido n_i valores $y_{ik}, k = 1, 2, ..., n_i$ con a lo menos uno $n_i > 1$.
- Los estimadores de mínimos cuadrados se pueden calcular como siempre, resultando los promedios $\eta_i = \bar{y}_i$ por cada i.
- Igualmente, la suma de los cuadrados de los residuos vale

$$SS_W = \sum_{i=1}^{m} \sum_{k=1}^{n_i} (y_{ik} - \eta_i)^2$$

- SS_W, suma de cuadrados *intra* solo informa sobre σ^2, pues es formada de sumas de desvíos al promedio en cada grupo.

- Como en cada grupo los y tienen una distribución independiente y idéntica, con promedio $\eta_{x_i} = E(y_{ik}|\boldsymbol{x}_i)$ y varianza $V(y_{ik}) = \sigma^2$, resulta

$$E(SS_W) = E\left(\sum_{i=1}^{m} \sum_{k=1}^{n_i} (y_{ik} - \bar{y}_i)^2\right) = \sigma^2 \sum_{i=1}^{m} (n_i - 1) = \sigma^2 (n - m)$$

- Por lo tanto $MS_W = SS_W/(n-m)$ es un estimador insesgado de σ^2, *que no depende del modelo* de regresión.
- Se resulta la tabla de análisis de varianza

Fuente	*Grados de libertad (DF)*	*Sumas de cuadrados (SS)*	*Cuadrados promedios (MS)*	*Esperanza de cuadrados promedios*	F	$p-value$
Between	m	SS_B	$MS_B = SS_B/m$	$\sigma^2 + \sum_k n_k \beta_k^2/m$	MS_B/MS_W	p
Within	$n-m$	SS_W	$MS_W = SS_W/(n-m)$	σ^2		
Total	n	SS_T				

con MS_W que solo informa sobre σ^2.

Supongamos entonces que el modelo no ajuste bien, o sea que

$$E(\boldsymbol{y}|\boldsymbol{X}) = \boldsymbol{\gamma} \neq \boldsymbol{\eta} = \boldsymbol{X\beta}.$$

Como $E(\boldsymbol{y}|\boldsymbol{X}) = \bar{\boldsymbol{y}}_X$, los promedios de las clases, esto significa que η no estima los promedios verdaderos y por lo tanto $SS_e > SS_W$, el residuo verdadero. Se resulta que $SS_B > SS_r$, se puede compartir en $SS_r + SS_F$, y también $SS_e = SS_W + SS_F$. Proyectando $\boldsymbol{\gamma}$ sobre $\boldsymbol{X}$ y se consigue su proyección

$$\boldsymbol{\gamma}^\circ = \mathscr{P}\boldsymbol{\gamma}$$

y el vector $\boldsymbol{\gamma} - \boldsymbol{\gamma}^\circ = \mathscr{E}\boldsymbol{\gamma}$ que se puede llamar *vector residual del modelo*. Este informa sobre el desvío entre la esperanza verdadera y la supuesta. Su cuadrado es

$$\Lambda^2 = (\boldsymbol{\gamma} - \boldsymbol{\gamma}^\circ)'(\boldsymbol{\gamma} - \boldsymbol{\gamma}^\circ) = SS_F$$

- Si $\boldsymbol{\gamma} = \boldsymbol{\eta}$, entonces $SS_F = \Lambda^2 = 0$.
- SS_B tiene m grados de libertad;
- SS_r tiene p grados de libertad
- SS_F tiene $m - p$ grados de libertad
- así que se resulta

$$E(SS_F) = (m - p)\sigma^2 + \Lambda^2$$

- Si la regresión no es lineal, entonces los y_{ik} informan sobre los valores *verdaderos*, mientras $\hat{\eta}$ solo informa sobre la linealidad.

- por esto SS_e además que informar sobre σ^2 tiene que informar también sobre el desvío a la linealidad de la función verdadera y sera mas grande que σ^2.
- Come se repitieron algunas medidas para los mismos $\boldsymbol{x}_i$ se está en condición de medir σ^2 y por tanto de comprobar su diferencia con SS_e.
- Para esto se hace una *análisis de varianza* sobre los datos, compuestos de m grupos de n_i observaciones bajo la asunción que las esperanzas de los y en cada grupo resultan de la recta de regresión

$$E(y_{ik}|\boldsymbol{x}_i) \;=\; \hat{\alpha} + \hat{\beta}\boldsymbol{x}_i \;,\;\; i \;=\; 1, 2, ..., m$$

si bien que se duda que esas sean diferentes.

- Análogamente al caso lineal se llega a la tabla de análisis de varianza

Fuente	*Grados de libertad* (DF)	*Sumas de cuadrados* (SS)	*Cuadrados promedios* (MS)	*Esperanza de los cuadrados promedios* $E(MS)$	F_F	*p-value*
Inter	p	$SS_r = \hat{\boldsymbol{\eta}}'\hat{\boldsymbol{\eta}} = \boldsymbol{y}'\mathscr{P}\boldsymbol{y}$	$MS_r = SS_r/p$	$\sigma^2 + \boldsymbol{\gamma}^{\circ\prime}\boldsymbol{\gamma}^{\circ}$		
Inter falta	$m-p$	$SS_F = \sum_k n_k(\bar{y}_k - \hat{\eta}_k)^2$	$MS_F = SS_F/(m-p)$	$\sigma^2 + \Lambda^2/(m-p)$	MS_M/MS_W	p
Intra	$n-m$	$SS_W = \sum_{k=1}^{m}\sum_{i=1}^{n_k}(y_{ik} - \bar{y}_k)^2$	$MS_W = SS_W/(n-m)$	σ^2		
Total	n	$SS_T = \boldsymbol{y}'\boldsymbol{y}$				

- Para testar el ajuste del modelo, se puede rechazar la hipótesis de linealidad a nivel de probabilidad π si

$$F_F = \frac{MS_F}{MS_W} > F_{m-p,n-m;\pi}$$

y aceptarla en caso contrario.

- Si la hipótesis de linealidad no es aceptada, *cualquier* relación no lineal puede ser comprobada.
- Sin observaciones repetidas, es necesario observar la distribución de los residuos sobre gráficos de dispersión en respecto a x y $\hat{\eta}$: si se distribuyen regularmente en una cinta alrededor de la recta horizontal $e = 0$, se pueden aceptar las hipótesis hechas, en particular homocedasticidad y linealidad.
- la función R `residualPlots(mod)` brinda gráficos de residuos con todos los regresores. Además proporciona un test de Tucker entre la variable respuesta y cada regresor: la falta de linealidad se encuentra por valores pequeños del $p-$valor $p < \alpha$).

Independencia de los residuos

La necesidad de las observaciones correspondientes a diferentes $\boldsymbol{x}_i$ de ser independientes deriva de la recuesta que la matriz de varianza/covarianza entre observaciones sea diagonal. Sin esto ni se puede suponer la varianza de los residuos ser constante.

SOLO si la proximidad entre observaciones en el archivo tiene sentido (p.e., en series de tiempo), se puede aplicar el test de Durbin y Watson:

$$d = \frac{\Sigma_{i=2}^{n}(e_i - e_{i-1})^2}{\Sigma_{i=1}^{n}\, e_i^2}$$

d varia entre 0 y 4, el 2 indicando independencia, 0 autocorrelación positiva y 4 autocorrelación negativa.

El test tiene como $H_0 : d = 2$, entonces si la probabilidad asociada al test es menor del umbral de significación elegido, se rechaza, aceptando que no hay independencia.

En R el test se consigue con el comando:

```
library(car)
durbinWatsonTest(mod)
```

Mediando valores consecutivos, esto test solo tiene sentido si la secuencia de datos tiene un sentido:

- series de tiempo, ya que observaciones contiguas son cercanas en el tiempo,
- datos de panel, donde los datos repetidos de un jurado son registrados en secuencia,
- datos espaciales, sitios cercanos siendo registrados vecinos.

Homocedasticidad

- Para identificar un modelo lineal con los mínimos cuadrados, se asume que la varianza $V(\varepsilon_i|x_i) = \sigma^2$ sea constante, o sea un presupuesto de *homocedasticidad*.
- Si al contrario esto no se cumple, el teorema de Gauss–Markov no se puede aplicar, así que los estimadores de mínimos cuadrados no son más los mejores estimadores insesgados y de varianza mínima.
- La heterocedasticidad no afecta la estimación de los $\boldsymbol{\beta}$ pero si afecta la estimación de la varianza y del error estándar de los estimadores.
- Por lo tanto se consigue una correcta estimación de los parámetros $\boldsymbol{\beta}$, pero no es posible hacer inferencia sobre el modelo mismo.

Supongamos de estimar el modelo $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Si el modelo cumple con el presupuesto de varianza constante, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ y su varianza estimada σ^2. Si no cumple, se puede imaginar que la varianza depende linealmente de los mismos regresores. El teste Breusch–Pagan (revisto para Cook y Weisberg) se la estima como

$$\hat{\boldsymbol{\varepsilon}}^2 = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\delta}$$

Se prueba que el múltiple del coeficiente de determinación nR^2 de este modelo resulta asintóticamente un χ^2_{p-1} bajo la hipótesis nula de homocedasticidad.

En R el test se realiza con los siguientes comandos:

```
library(lmtest)
ncvTest(bptest)

library(car)
ncvTest(mod)
```

El primero test no está corregido según Cook y Weisberg.

Si el $p-$valor que resulta es menor del umbral de significación fijado, se puede rechazar la hipótesis de homocedasticidad, y considerar que la varianza no es constante.

Testar los residuos

Supongamos que ya tenemos un modelo, en R

```
mod = lm( y ~ x1 + x2 +..., data=datos)
```

La primera observación gráfica es de representar en un gráfico los $\hat{\eta}_x$ con los y, que puede nos informar también sobre la falta de ajuste.

```
plot (mod$fitted.values, data$y)
abline(0,1)
```

La recta representa la coincidencia entre y_i y $\hat{\eta}_{x_i}$ y por tanto la distancia en vertical $e_i = y - \hat{\eta}_x$ representa el residuo.

Los residuos se representan por respecto a $\boldsymbol{y}$ y a todos los $\boldsymbol{x}_j$, mostrando también si la varianza es constante o no:

```
plot (data$y, mod$residuals)
abline(0,0)
plot (data$x_j, mod$residuals)
abline(0,0)
```

La función `residualPlots` del paquete R `car` brinda los gráficos de los residuos con una curva para ver si hay falta de ajuste, juntamente con el test de Tukey para la no-linealidad:

```
library(car)
residualPlots(mod)
```


Test por la normalidad de los residuos

Para testar la distribución normal de los residuos, se representan con un Q-Q plot, que compara el cuantíl de cada observación ordenada con los teóricos resultando de la distribución normal.

```
qq <- qqnorm(mod$residuals)      # dibuja el gráfico
qqline(mod$residuals)            # recta de normalidad
identify(qq$x,qq$y,lab=labels)   # nombres a placer
shapiro.test(mod$residuals)      # test de Shapiro-Wilk
```

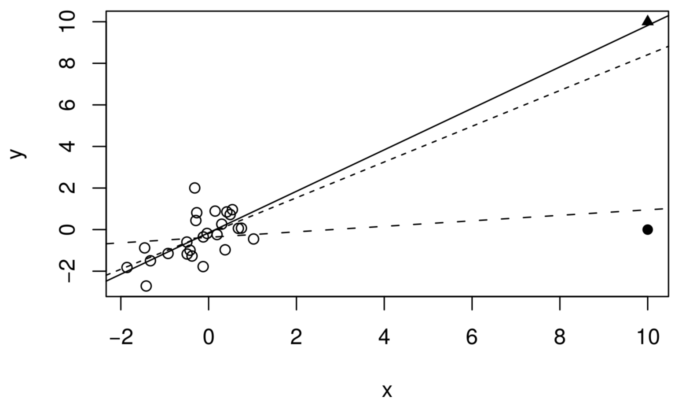
El test de Shapiro-Wilk se usa para contrastar la normalidad de un conjunto de datos. Se plantea como hipótesis nula que una muestra $x_1, \ldots, x_n$ proviene de una población normalmente distribuida.

El estadístico del test es: $W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}$ donde

- $x_{(i)}$ es el número que ocupa la i-ésima posición en la muestra;
- $\bar{x} = (x_1 + \ldots . + x_n)/n$ es la media de la muestra;
- las variables a_i se calculan $(a_1, \ldots, a_n) = \frac{\boldsymbol{m}' V^{-1}}{(\boldsymbol{m}' \boldsymbol{V}^{-1} \boldsymbol{V}^{-1} \boldsymbol{m})^{1/2}}$ donde $\boldsymbol{m} = (m_1, \ldots, m_n)'$, vector de los valores medios del estadístico ordenado, de variables aleatorias independientes e idénticamente distribuidas, muestradas de distribuciones normales;
- $\boldsymbol{V}$ es la matriz de covarianza de ese estadístico de orden.

La hipótesis nula de normalidad se rechazará si el p-valor asociado W es menor del umbral establecido.

Apalancamiento



Ya se ha encontrado la matriz sombrero $\mathscr{P} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}' = \boldsymbol{H}$ proyección sobre el espacio generado para $\boldsymbol{X}$, h_{ij} indicando la influencia de y_j sobre la predicción η_i. A h_{ii} se lo llama *apalancamiento* y representa cuanto la i-ésima observación influye sobre la regresión. La verdad es que esto solo depende de la distancia de la observación del baricentro de los x.

Como el promedio de los h_{ii} es p/n, tiene que mirar de cerca las observaciones con $h_{ii} \geq 2p/n$.

Efectivamente estas observaciones, si se encuentran lejos de las demás, pueden influenciar de manera relevante la regresión misma y hay que considerar si guardarlas o no para la estimación: ya se ha experimentado no ser automático.

La función `influence` de R brinda unos resultados útiles:

- *hat*: la diagonal de la matriz sombrero (hat matrix), o sea los apalancamientos;
- *coefficients*: la diferencia entre los $\hat{\beta}_j$ y los $\hat{\beta}_{(i),j}$ calculados tirando la i-ésima observación;
- *sigma*: el desvío estándar de los residuos, calculados tirando la i-ésima observación:

`halfnorm` evidencia los k apalancamientos más grandes.

```
library(faraway)
halfnorm(influence(mod)$hat,nlab=k,labs=labc)
```

seminormal (half-normal) siendo la distribución de los valores absolutos de una variable aleatoria que tiene una distribución normal.

Puntos aberrantes

Sabemos que el error depende de la unidad de medición y que la varianza de $\boldsymbol{e}$ es

$$V(\boldsymbol{e}) = \sigma^2(I_n - \mathscr{P}) = \sigma^2(I_n - H)$$

por tanto resulta ser $V(e_i) = \sigma^2(1 - h_{ii})$ la varianza de los residuos disminuyendo según el apalancamiento.

Se resultan los residuos *estandarizados*, con varianza 1:

$$se_i = \frac{e_i}{\sqrt{MSe(1 - h_{ii})}},$$

que permite de comparar mejor los residuos entre ellos. Se sugiere de considerar aberrantes unidades con residuo estandarizado > 3.

Como un punto aberrante atrae la recta, el resultante residuo podría no aparecer. Por esto es mejor emplear *residuos estimados* $\hat{\varepsilon}_i = y_i - \eta_{(i),i}$ el i-ésimo siendo estimado para la regresión sin la i-ésima observación.

Estandarizando cada uno con la varianza estimada para $MSe_{(i)}$ resultando de $\eta_{(i),i}$ se consiguen los residuos *studentizados*:

$$ste_i = \frac{\hat{\varepsilon}_i}{\sqrt{MSe_{(i)}(1-h_{ii})}} = se_i\sqrt{\frac{(n-p-1)}{(n-p-se_i^2)}},$$

la segunda igualdad permitiendo un cálculo más rápido.

De manera estandarizada estos residuos indican cuanto mal los y_i son estimados para una regresión que no los incluye.

Se sugiere de considerar aberrantes unidades con residuo estandarizado > 3.

Como se resulta que estos tienen una distribución t de Student, con $n-p-1$ grados de libertad, su ser influyentes se testa contra la hipótesis nula que no son influyentes.

En R el comando es

```
rstudent(mod)
```

Es conveniente en este caso también trazar el $Q-Q$-plot:

```
rs = rstudent(mod)
qqnorm(rs)
qqline(rs)
```


Indicando con $\hat{\boldsymbol{\beta}}_{(i)}$ la estimación de los $\boldsymbol{\beta}$ hecha tirando la i-ésima observación de la muestra, resulta

$$\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)} = \frac{(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i e_i}{1-h_{ii}}$$

que media los desvíos de los betas tirando del modelo la i-ésima observación.

Sumando los cuadrados y estandarizando, se resulta la *distancia de Cook*

$$D_i = \frac{\left(\hat{\eta}_i - \hat{\eta}_{(i),i}\right)^2}{pMS_e} = \frac{\left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)' X'X \left(\hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(i)}\right)}{pMS_e} = \frac{se_i^2}{p}\frac{h_{ii}}{(1-h_{ii})}$$

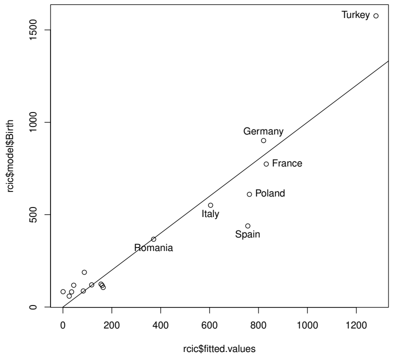
que se basa sobre el producto del cuadrado del residuo estandarizado y el apalancamiento.

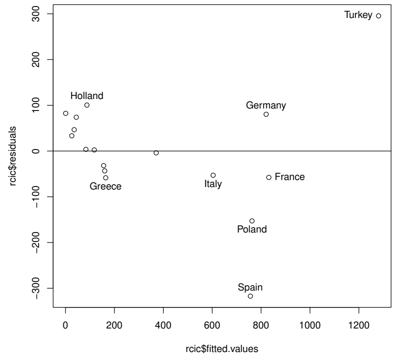

```
cooks.distance(mod)
```

¿Como detectar los puntos aberrantes?

- Una regla elemental es que posibles puntos aberrantes son los cuya distancia $D_i > 3\bar{D}$, la distancia de Cook promedio.
- Otra propuesta es de chequear los puntos con $D_i > \frac{4}{n}$ con n el número de observaciones.
- Otros autores sugieren que cualquier D_i "grande" sea examinado, pero sin definir que tal "grande"...
- Alternativa técnica es de utilizar como umbral el 50 percentil de la distribución F.

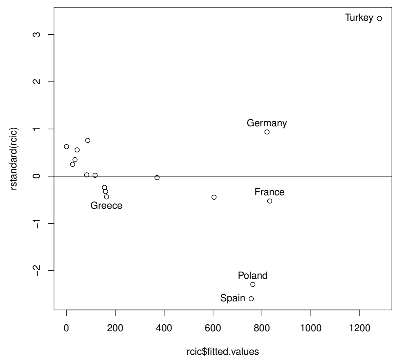
Ejemplo: Cigueñas





residuos estandarizados: Turquía resulta > 3

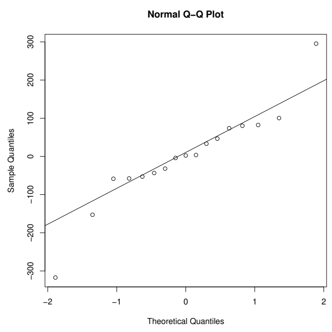
```
> rstandard(rcic)
 Albania     Austria     Belgium     Bulgaria    Denmark     France      Germany     Greece      Holland
0.62554539  0.02761022  0.55697132 -0.32521339  0.25186998 -0.52560926  0.93745406 -0.43803095  0.75816180
 Hungary     Italy       Poland      Portugal    Romania     Spain       Switzerland Turkey
-0.23848733 -0.44566094 -2.29284124  0.01821925 -0.02937552 -2.59608144  0.35120143  3.34060473
```



Test de homoscedasticidad y de normalidad

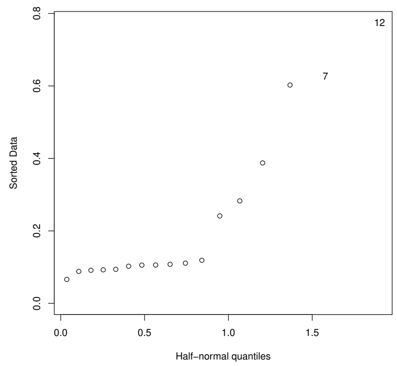
```
> ncvTest(rcic)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 13.69344, Df = 1, p = 0.00021521
no hay homoscedasticidad
```

```
> shapiro.test(rcic$residuals)
Shapiro-Wilk normality test
data:  rcic$residuals
W = 0.9188, p-value = 0.141
si hay normalidad
```



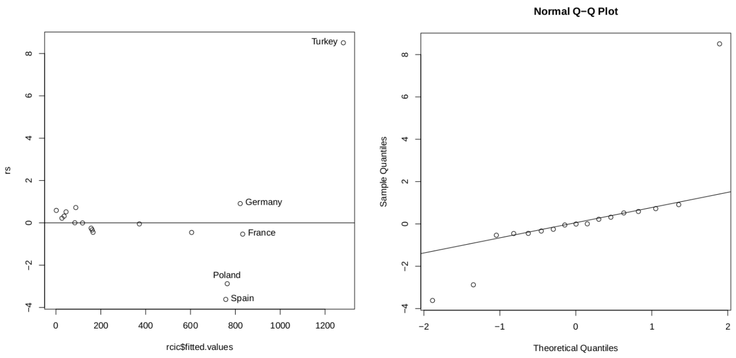
apalancamientos: mayores Polonia y Alemania

```
  influence(rcic)$hat
  Albania     Austria     Belgium    Bulgaria     Denmark      France     Germany      Greece     Holland
0.11900848  0.10250893  0.10580913  0.09398970  0.11107498  0.38753751  0.62722071  0.09265502  0.10791279
Hungary     Italy       Poland      Portugal    Romania     Spain       Switzerland Turkey
0.09140244  0.28282549  0.77437997  0.08823535  0.06612908  0.24127770  0.10554654  0.60248618
```



residuos studentizados: Turquía y España > 3

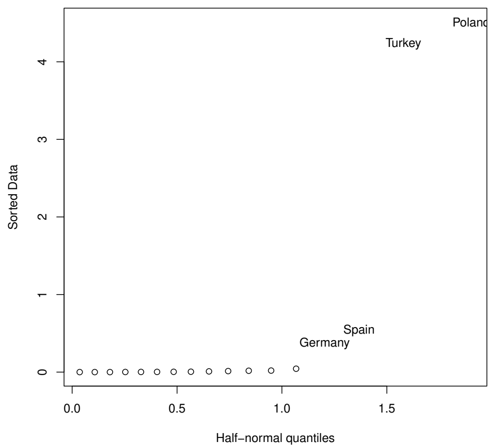
```
> rs = rstudent(rcic); rs
Albania     Austria     Belgium     Bulgaria    Denmark     France      Germany     Greece      Holland
0.61025931  0.02652782  0.54162208 -0.31373374  0.24258146 -0.51044191  0.93275625 -0.42398703  0.74507671
Hungary     Italy       Poland      Portugal    Romania     Spain       Switzerland Turkey
-0.22963410 -0.43148599 -2.85439038  0.01750471 -0.02822402 -3.59425795  0.33903562  8.53030031
```




```
> round(cooks.distance(rcic),3)        # Cook's distance
Albania     Austria     Belgium     Bulgaria    Denmark     France
0.013       0.000       0.009       0.003       0.002       0.044
Germany     Greece      Holland     Hungary     Italy       Poland
0.370       0.005       0.017       0.001       0.020       4.511
Portugal    Romania     Spain       Switzerland Turkey
0.000       0.000       0.536       0.004       4.228

> which(cooks.distance(rcic)>3)
Poland Turkey
12     17

> which(cooks.distance(rcic)>4/nc)
Germany  Poland   Spain   Turkey
7        12       15      17
```

El ploteo estándar de `lm`: `plot(mod)`