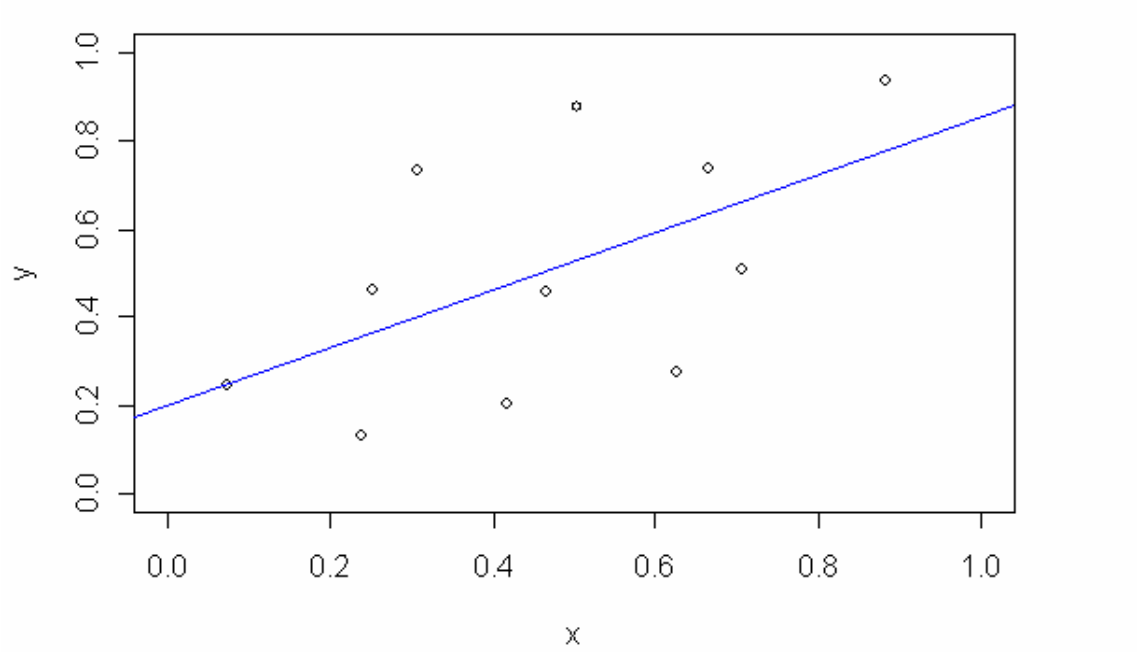
REGRESIÓN LINEAL SIMPLE

Dado un conjunto de pares de datos (x_i, y_i), $i = 1,\ldots,n.$, se han desarrollado varios métodos para ajustar una recta de la forma

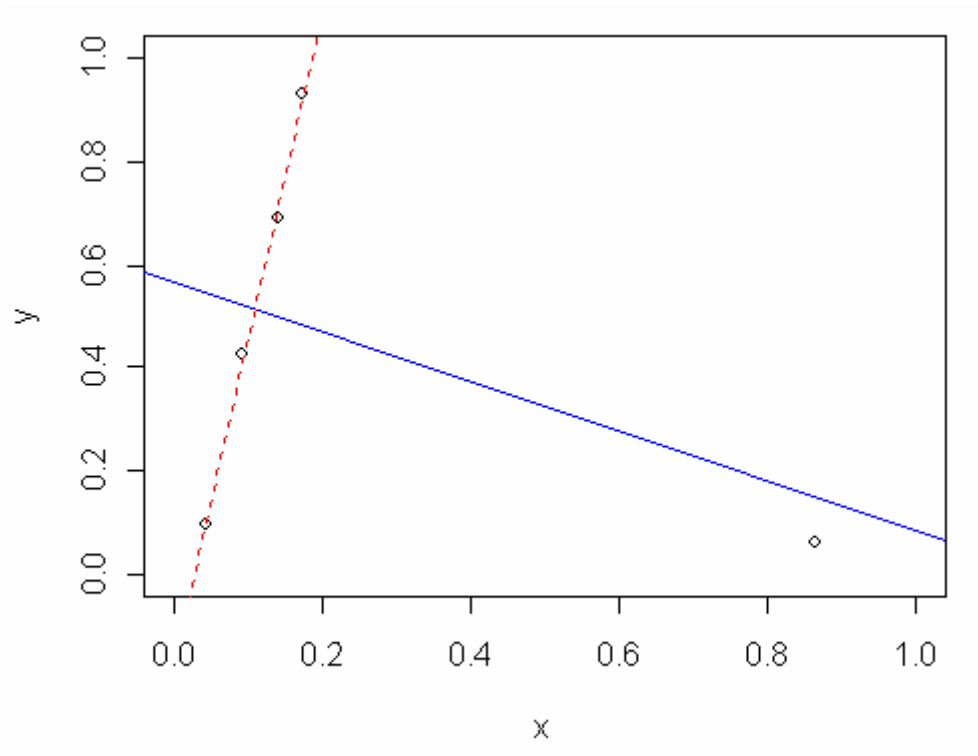
$$y = a + bx,$$

al diagrama de dispersión de dichos datos.



El más conocido y el más utilizado es el Método de Cuadrados Mínimos que requiere cálculos algebraicamente sencillos y desarrollo matemático directo.

Desafortunadamente la recta ajustada por el método de cuadrados mínimos no es resistente. Un sólo punto puede tomar control del ajuste, como lo muestra la figura siguiente, y llevar a una interpretación absolutamente errónea de la relación entre *x* e *y.*



La figura muestra el ajuste por cuadrados mínimos y un ajuste resistente.

Hemos obtenido el gráfico anterior en R mediante las siguientes instrucciones:

```
>
> library(rrcov)
>
plot(c(0,1),c(1,0),type="n",xlab="x",ylab="y
")
>    xy<-locator(type="p")  # marcamos los puntos
```

en el gráfico y sus coordenadas son guardadas en **xy.**

```
>  xy
$x
[1]    0.04403185    0.09177722    0.13952259
0.17294434 0.86525219

$y
```

```
[1]    0.09642654    0.42798173    0.68887762
0.92803218 0.06381455
```

Trazamos la recta de cuadrados mínimos

```
> abline(lm(y~x,xy),col=4)
```

Trazamos la recta resistente, que minimiza la suma de los cuadrados de la mitad de los residuos más pequeños.

```
> abline(ltsReg(xy$x,xy$y),lty=2,col=2)
```

En S-plus

Trazamos la recta de cuadrados mínimos igual que en R

```
> abline(lm(y~x,xy),col=4)
```

Pero trazamos la recta resistente, que también minimiza la suma de los cuadrados de la mitad de los residuos más pequeños, con la función **`ltsreg`** . No se utiliza niguna librería adicional.

```
> abline(ltsreg(xy$x,xy$y),lty=2,col=2)
```

Método de Cuadrados mínimos

Los coeficientes de la recta de cuadrados mínimos (CM) se eligen entre todos los posibles pares de valores aquellos que minimizan

$$S = \sum_{i=1}^{n} (y_i - a + b x_i)^2 \qquad (1)$$

Derivando (1) respecto de a y b

$$\partial S / \partial a = -2\sum_{i=1}^{n}(y_i - (a + bx_i))$$

$$\partial S / \partial b = -2\sum_{i=1}^{n}(y_i - (a + bx_i))\, x_i$$

e igualando a cero se obtienen los coeficientes de la recta estimada por CM:

$$b_{CM} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a_{CM} = \bar{y} - b_{CM}\,\bar{x}$$

Ninguna otra recta tendrá, para el mismo conjunto de datos, una suma de cuadrados de los residuos tan baja como la obtenida por CM. En este sentido, CM garantiza la solución que mejor ajusta a ese conjunto de datos.

Se puede obtener fácilmente, que la recta de CM pasa por el punto $(\bar{x}, \bar{y})$.

Rectas resistentes.

Recta resistente de tres grupos.

Para resumir el centro de un lote, el procedimiento resistente más simple es la mediana muestral. La técnica exploratoria que veremos para ajustar una recta (Tukey, 1970) deriva su resistencia de la mediana.

Formación de tres grupos.

Ordenamos los valores de *x* de manera que $x_1 \leq x_2 \leq \ldots \leq x_n$. Luego sobre la base de los datos ordenados dividimos los *n* datos (x_i, y_i) en tres grupos, el de la izquierda, el del centro y el de la derecha. Cuando no hay empates entre las x_i, la cantidad de puntos en los tres grupos depende del resto de dividir *n* por 3.

Ubicamos los puntos a los grupos de la siguiente manera:

Grupo	***n = 3k***	***n = 3k + 1***	***n = 3k + 2***
Izquierda	*k*	*k*	*k* + 1
Centro	*k*	*k* +1	*k*
Derecha	*k*	*k*	*k* + 1

Los empates entre las x_i pueden impedir que se logre esta ubicación exactamente porque no separamos los empates. Todos los datos con el mismo valor de *x* van al mismo grupo.

Puntos centrales de tres grupos (summary points).

Dentro de cada uno de los grupos formados, determinamos dos coordenadas de un punto central hallando primero la mediana de las *x's* y luego la mediana de las *y's.* Indicamos las coordenadas de los tres puntos centrales, ***I*** para izquierda, *C* para el centro y *D* para derecha:

$$(x_I, y_I), (x_C, y_C), (x_D, y_D).$$

La figura 1 muestra los puntos centrales en un ejemplo hipotético de nueve puntos. Como enfatiza la figura ninguno de los puntos centrales es necesariamente un

dato, pues las medianas de las *x's* y de las *y's* se calculan en forma separada.

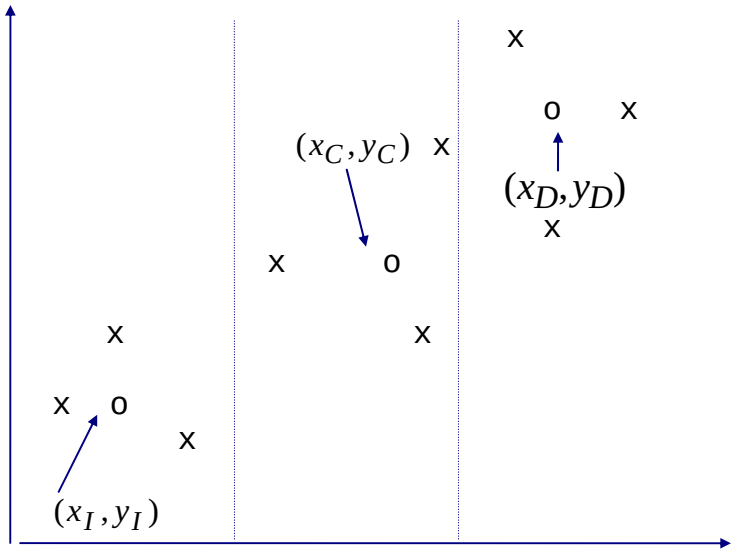


Figura 1: Los datos (x) y los puntos centrales (o) en un ejemplo hipotético.

Sin embargo todos los puntos centrales podrían ser datos, como ocurre frecuentemente cuando las x_i 's y las y_i 's tienen el mismo ordenamiento dentro de cada grupo.

Este método da un recta resistente. La mediana provee resistencia a valores salvajes en *x*, *y* o ambos, siempre y cuando el número de puntos en cada grupo no sea muy pequeño.

Pendiente y ordenada al origen (slope and intercept)

Vamos a utilizar los puntos centrales para obtener la pendiente y la ordenada al origen de una recta

$$y = a + bx$$

Es usual utilizar la notación de *y* "sombrero" en la expresión de la recta ajustada para recordar que es la fuente de los

valores ajustados tanto para valores *x*'s de los datos como en otros valores adecuados. Así tendremos:

$$\hat{y} = a + bx$$

La pendiente de la recta indica en cuantas unidades cambia *y* en respuesta a un cambio de una unidad en *x* y obtenemos esta información a partir de los puntos resumen izquierda y derecha:

$$b_0 = \frac{y_D - y_I}{x_D - x_I}.$$

De esta manera buscamos un balance entre

(a) la ventaja de medir el cambio de *y* sobre un intervalo amplio de *x* y
(b) la necesidad de tener suficientes datos en el grupo de la izquierda y el de la derecha para tener una resistencia adecuada.

Para cada punto central $(x_I, y_I), (x_C, y_C), (x_D, y_D)$, hay una recta que pasa por dicho punto y tiene como pendiente la pendiente ajustada b_0, el promedio de las ordenadas al origen de cada una de esas rectas es la ordenada al origen ajustada:

$$a_0 = \frac{1}{3}[(y_I - b_0 x_I) + (y_C - b_0 x_C) + (y_D - b_0 x_D)]$$

Nuevamente como los puntos centrales están basados en medianas a_0 es resistente.

Para comparar, consideremos la pendiente y la ordenada al origen de la recta ajustada por cuadrados mínimos

$$b_{CM} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

$$a_{CM} = \bar{y} - b_{CM}\bar{x}$$

La imposibilidad de cualquier tipo de resistencia en los estimadores de cuadrados mínimos es evidente en la forma en que todos los datos entran en el cálculo de los coeficientes.

Pendiente y valor central

Ajustar un recta en términos de la pendiente y la ordenada al origen es convencional pero generalmente artificial. La ordenada al origen da un valor de *y* cuando *x* = 0, que puede estar determinado muy imprecisamente y no tener importancia cuando los valores de *x* caen lejos del cero.

Generalmente es más útil realizar el ajuste en base a la pendiente y la ordenada correspondiente a un *x* que puede ser $x = \bar{x}$ ó $x = \text{mediana}(x_i)$ ó $x = x_C$, llamada **valor central** Si por conveniencia trabajamos en $x = x_C$, entonces la recta inicial es:

$$\hat{y} = a + b_0(x - x_C) \qquad (1)$$

Igual que antes consideramos las tres rectas, de la forma (1) con pendiente b_0 que pasan por los tres puntos centrales respectivamente. Promediamos los valores a obtenidos para cada una de ellas:

para (x_I, y_I) , $a_I = y_I - b_0(x_I - x_C)$,

para (x_C, y_C) , $a_C = y_C$,
para (x_D, y_D) , $a_D = y_D - b_0(x_D - x_C)$.

Por lo tanto la recta ajustada será

$$\hat{y} = a_0^* + b_0(x - x_C)$$

con b_0 igual que antes y la ordenada en el valor central a_0^*, también llamado nivel, está dada por

$$\begin{aligned} a_0^* &= \frac{1}{3}[a_I + a_C + a_D] \\ &= \frac{1}{3}\{[y_I - b_0(x_I - x_C)] + y_C + [y_D - b_0(x_D - x_C)]\}. \end{aligned}$$

Residuos.

Una vez que se han obtenido la pendiente y el nivel para la recta ajustada, el paso inmediato siguiente es calcular el residuo para cada dato:

$$r_i = y_i - [a^* + b(x_i - x_C)]$$

Los residuos son la base de varios gráficos que permiten revelar una gran variedad de aspectos y patrones del ajuste.

Pero, en este caso, solamente necesitamos enfatizar una propiedad general de un conjunto de residuos, tanto en *y* versus *x* como en situaciones más complejas:

Sustituir los residuos en vez de los valores *y* originales (i.e. utilizar (x_i, r_i) en vez de (x_i, y_i), $i = 1, \ldots, n$) y repetir el ajuste lleva a un ajuste cero.

Para la recta esto significa que utilizar (x_i, r_i) como datos lleva a pendiente cero y nivel cero. En otras palabras, los residuos ya no contienen más información lineal para resumir, tal como ocurre al utilizar el método de cuadrados mínimos.

Una característica importante de los procedimientos resistentes es que frecuentemente requieren iteración. La recta resistente de 3 grupos es un primer ejemplo. Si los residuos de la recta con pendiente b_0 y nivel a_0^* no tienen pendiente cero y nivel cero, ajustamos una recta a ellos. La nueva pendiente y el nuevo nivel serán sustancialmente menores (en magnitud) que b_0 y a_0^*. Por esta razón pensamos a b_0 como un valor inicial para la pendiente y a_0^* como un valor inicial para el nivel (de allí el subíndice 0).

Iteración

Generalmente esperamos que b_0 y a_0^* necesiten alguna corrección. Ajustar una recta a los residuos de la recta inicial da las correcciones δ_1 y γ_1 para la pendiente y el nivel respectivamente. Específicamente, utilizamos los residuos iniciales

$$r_i^{(0)} = y_i - [a_0^* + b_0(x_i - x_C)], \qquad i = 1, \cdots, n,$$

en lugar de y_i para repetir la mayoría de los pasos anteriores del proceso de ajuste. Las x_i no han cambiado,

de manera que los grupos y las medianas de las *x*'s no varían durante el proceso iterativo.

La pendiente y el nivel corregidos son $b_0 + \delta_1$ y $a_0^* + \gamma_1$ y los nuevos residuos son

$$r_i^{(1)} = r_i^{(0)} - [\gamma_1 + \delta_1(x_i - x_C)], \qquad i = 1, \cdots, n.$$

Podríamos ahora intentar otra iteración. en general no sabemos si tenemos un conjunto adecuado de residuos hasta que verificamos que tienen un ajuste cero. En la práctica, continuamos las iteraciones hasta que la corrección a la pendiente es suficientemente pequeña (a lo sumo 1% ó 0.01%) del tamaño de b_0. Cada iteración agrega las correcciones de la pendiente y el nivel a los valores anteriormente modificados:

$$b_1 = b_0 + \delta_1, \qquad b_2 = b_1 + \delta_2, \ldots$$

y

$$a_1^* = a_0^* + \gamma_1, \qquad a_2^* = a_1^* + \gamma_2, \ldots$$

Las iteraciones son generalmente lo suficientemente pocas de manera de no llevar demasiado tiempo, la resistencia justifica el esfuerzo.

Para algunos conjuntos de datos las correcciones de la pendiente decrecen demasiado lentamente o después de algunos pasos pueden dejar de decrecer y en cambio oscilar entre dos valores con la misma magnitud y signo opuesto. Después del ejemplo veremos una variación del procedimiento que elimina este problema y disminuye drásticamente la cantidad de iteraciones.

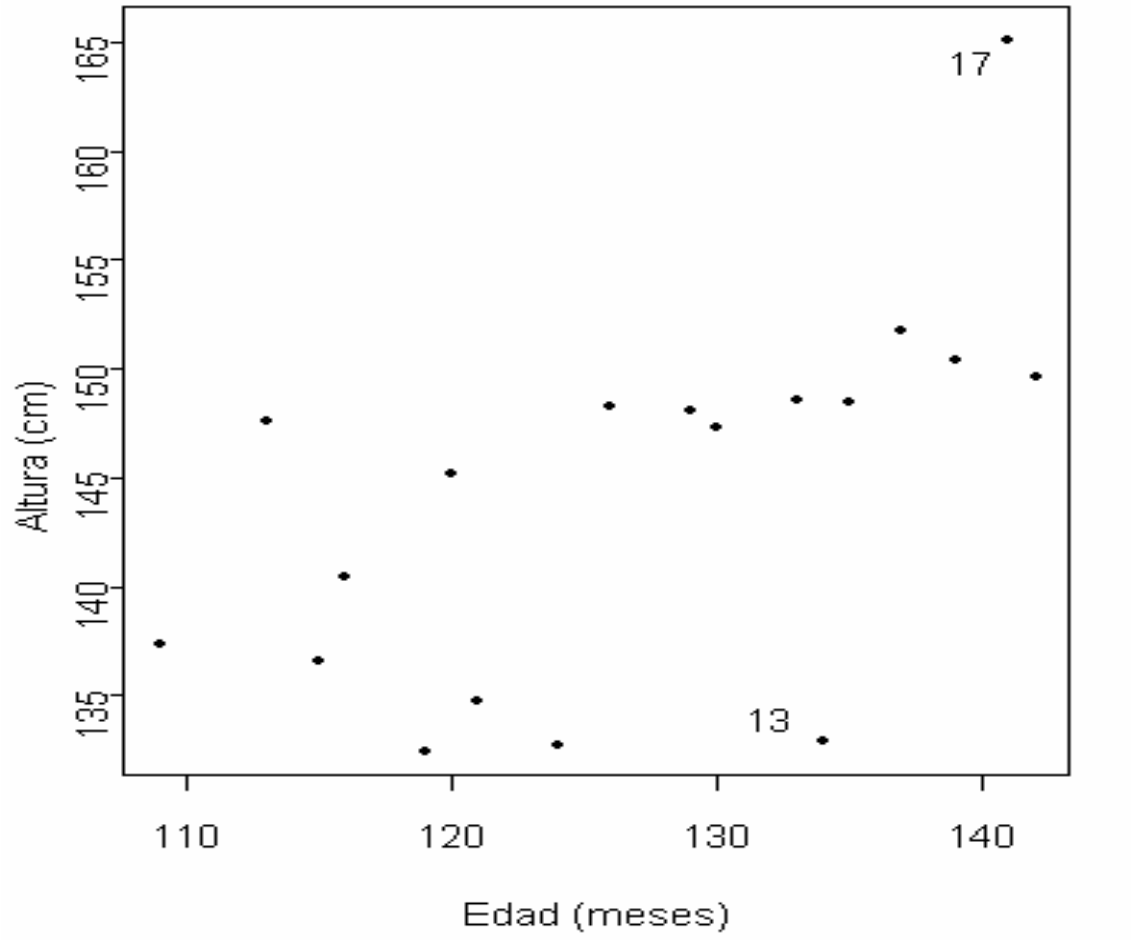
EJEMPLO: Como un ejemplo de una discusión de 1953, Greenberg dio las edades y las alturas de dos muestras de

niños -una muestra de una escuela privada de ciudad y otra muestra de una escuela pública rural. reproducimos los datos correspondientes a 18 niños de la escuela privada en la tabla 1 y graficamos la altura contra la edad en la figura 1.

Tabla1: Edad y altura de niños en una escuela privada.

Niño	Edad (meses)	Altura (cm)
1	109	137.6
2	113	147.8
3	115	136.8
4	116	140.7
5	119	132.7
6	120	145.4
7	121	135.0
8	124	133.0
9	126	148.5
10	129	148.3
11	130	147.5
12	133	148.8
13	134	133.2
14	135	148.7
15	137	152.0
16	139	150.6
17	141	165.3
18	142	149.9

Aunque los datos no siguen claramente una recta tampoco presentan un patrón notablemente curvo. De manera que una recta ajustada debería servir para resumir como aumenta la altura (y) con la edad (x) en ese grupo de niños.



Desde el SPLUS leemos los datos de un archivo texto sin la fila de nombres y luego asignamos nombres a las variables

```
> altedad <-
read.table("c:\\d\\master\\altedad.txt")
> names(altedad)<-c("altura","edad")
> plot(altedad$edad,altedad$altura,
xlab="Edad (meses)",ylab="Altura (cm) ")
> identify(altedad$edad,altedad$altura)
[1] 17 13
```

Los puntos correspondientes a los niños 13 y 17 parecen sobresalir así que los seguiremos a medida que desarrollamos el proceso.

Como los datos no incluyen empates, los dividimos en tres grupos con 6 puntos cada uno. Los tres puntos centrales son:

$$(x_I, y_I) = (115.5, 139.15)$$
$$(x_C, y_C) = (127.5, 147.9)$$
$$(x_D, y_D) = (138, 150.25).$$

```
> sapply(as.list(altedad[1:6,]),median)
   edad altura
  115.5 139.15
>
sapply(as.list(altedad[7:12,]),median)
   edad altura
  127.5  147.9
>
sapply(as.list(altedad[13:18,]),median)
  edad altura
   138 150.25
```

Por lo tanto la pendiente inicial es

$$b_0 = \frac{y_D - y_I}{x_D - x_I} = \frac{150.25 - 139.15}{138 - 115.5} = 0.4933.$$

y el nivel inicial es

$$a_0^* = \frac{1}{3}(145.07 + 147.90 + 145.07) = 146.01$$

La tabla 2 muestra los datos separados por grupo y los residuos de esta recta inicial.

Tabla 2. Altura y edad de niños - los tres grupos y los residuos iniciales

Niño	Edad (x)	Altura (y)	y - [146.01 + 0.4933(x- 127.5)]
1	109	137.6	0.716
2	113	147.8	8.943
3	115	136.8	-3.044
4	116	140.7	0.363
5	119	132.7	-9.117
6	120	145.4	3.090
7	121	135.0	-7.804
8	124	133.0	-11.28
9	126	148.5	3.230
10	129	148.3	1.550
11	130	147.5	0.257
12	133	148.8	0.077
13	134	133.2	-16.02
14	135	148.7	-1.010
15	137	152.0	1.304
16	139	150.6	-1.083
17	141	165.3	12.63
18	142	149.9	-3.263

```
> altedad$altura -(146.01+
                  0.4933*(altedad$edad-127.5))
 [1]   0.71605   8.94285  -3.04375   0.36295
 [5]  -9.11695   3.08975  -7.80355 -11.28345
 [9]   3.22995   1.55005   0.25675   0.07685
[13] -16.01645  -1.00975   1.30365  -1.08295
[17]  12.63045  -3.26285
```

Calculemos las correcciones a la pendiente y al nivel:

$$\delta_1 = \frac{-1.045 - 0.545}{138 - 115.5} = -0.0707$$

$$\gamma_1 = \frac{1}{3}(-0.30 + 0.07 - 0.30) = -0.14.$$

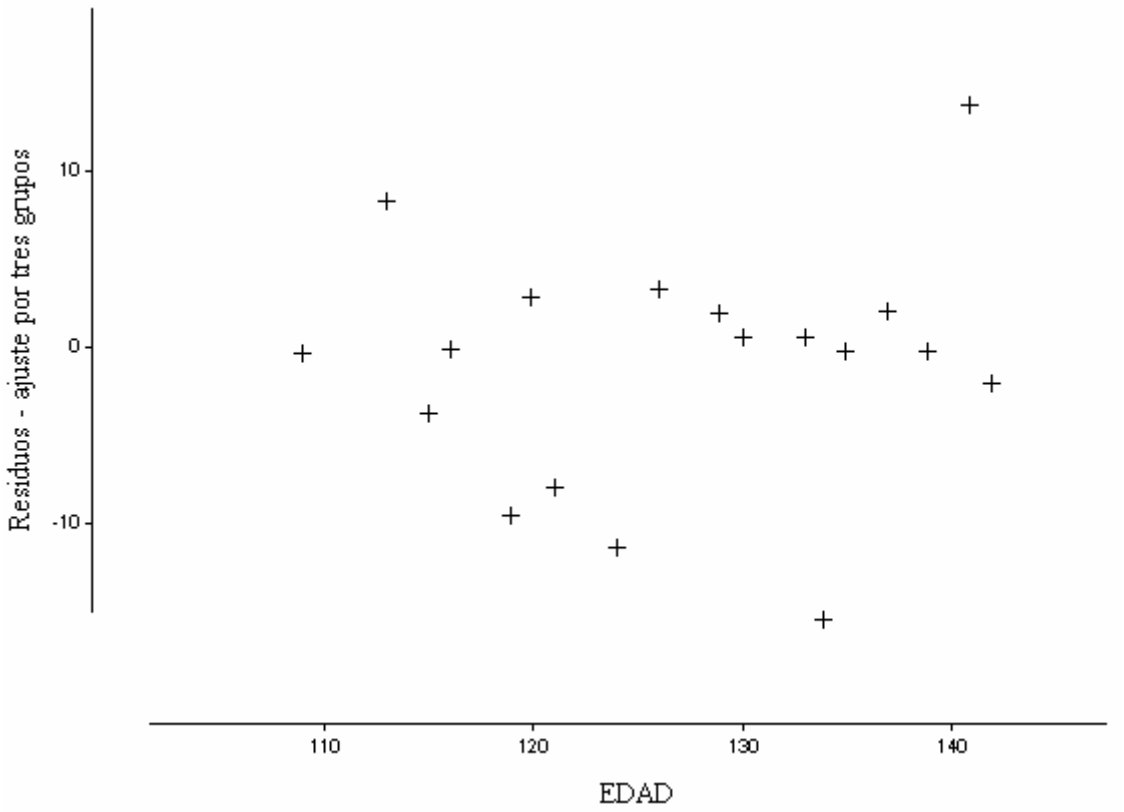
δ_1 es sustancialmente menor que b_0 pero aún no es despreciable. Dos iteraciones más nos llevan a una

situación en que el proceso puede parar: La corrección más reciente lleva e un cambio menor del 1% en la pendiente.

La recta resultante es

$$\hat{y} = 145.86 + 0.4285(x - 127.5)$$

Figura 3. Residuos de altura vs. edad después de un ajuste iterado por tres grupos

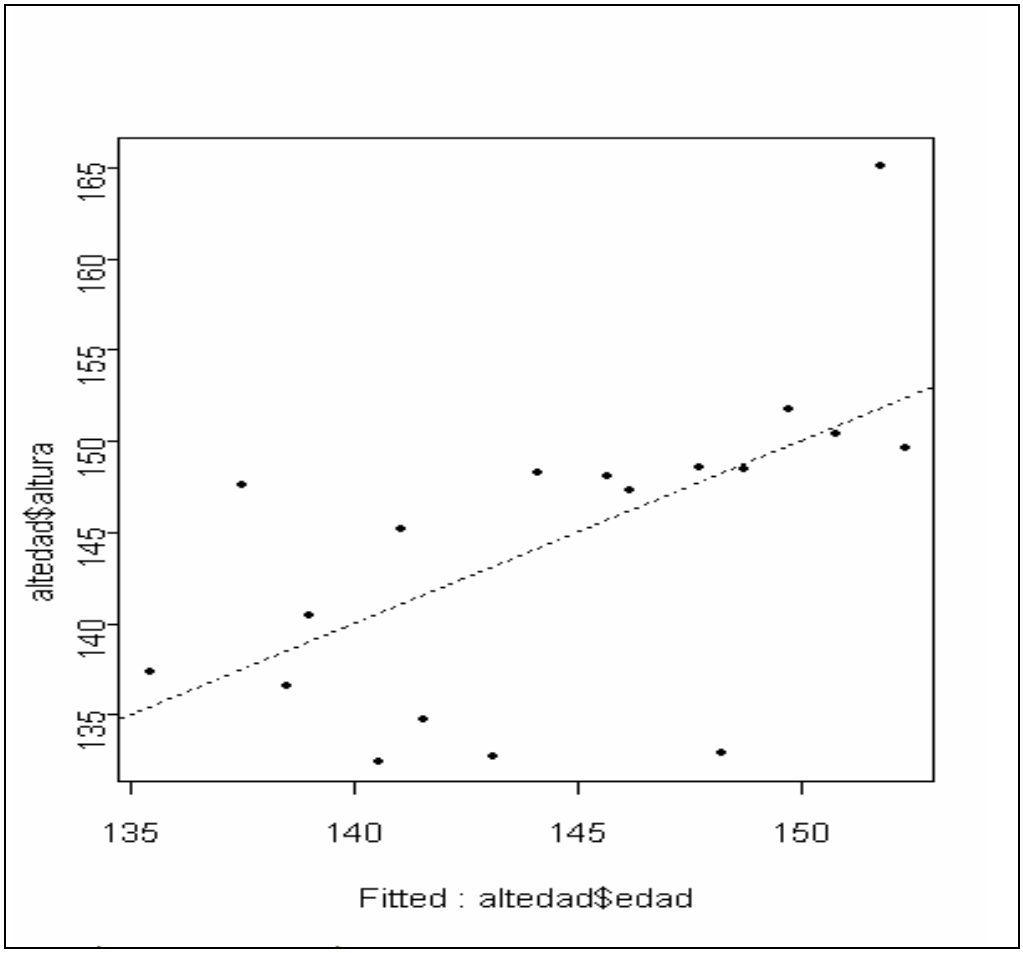


Se destacan un punto bajo y otro alto que corresponden a los niños 13 y 17. Estos residuos son lo suficientemente extremos, si los analizáramos mediante un boxplot están fuera de los valores adyacentes.

También parecen bajos los residuos de tres niños con edades alrededor de 120 meses. Si tuviéramos más información podríamos intentar aprender porqué estos niños son bajos o altos de acuerdo a su edad. La distinción entre varones y mujeres podría ayudar.

Los puntos inusuales casi no tuvieron efecto en la recta que resume al conjunto de los datos. Una recta de cuadrados mínimos enfrenta más riesgo de distorsión debido a esos puntos.

Veamos que ocurre con el ajuste por cuadrados mínimos.




```
>   lm(altedad$altura~altedad$edad)
Call:
lm(formula = altedad$altura ~ altedad$edad)

Coefficients:
 (Intercept) altedad$edad
    79.69623    0.5112868

Degrees of freedom: 18 total; 16 residual
Residual standard error: 7.028649
```

La recta de regresión por cuadrados mínimos es

$$\hat{y} = 79.7 + 0.511x$$

ó

$$\hat{y} = 144.9 + 0.511(x - 127.5),$$

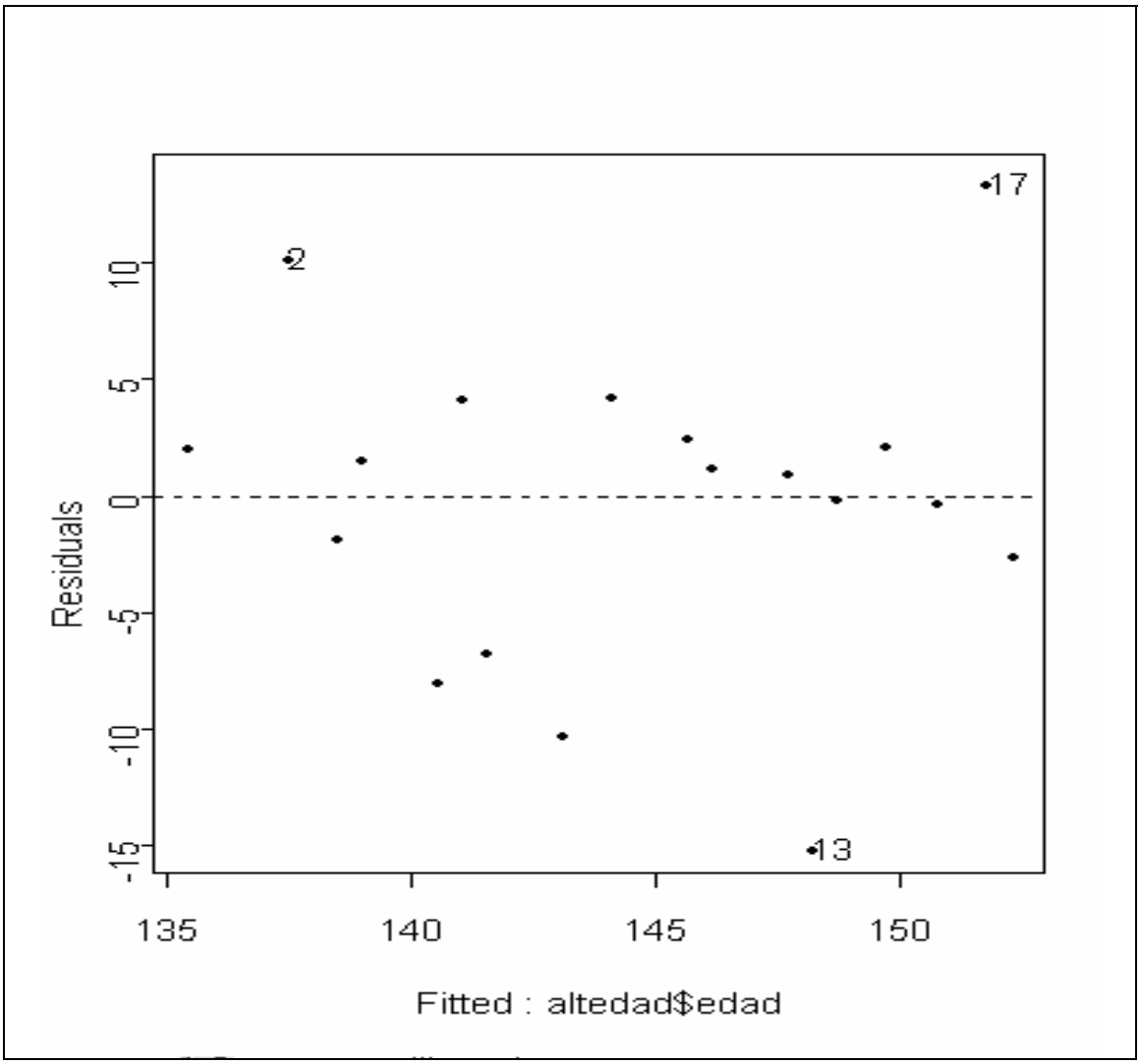


Figura 4. Residuos del ajuste de cuadrados mínimos

sugiriendo que los valores *y* de los niños 5, 7, 8 y 17 pueden haber ayudado a torcer la recta.

Si el valor del niño 13 no hubiese sido tan bajo la recta habría sido más empinada aún.

Los residuos con igual varianza se obtienen con **`type="pearson`** "en la función **`residuals`**:

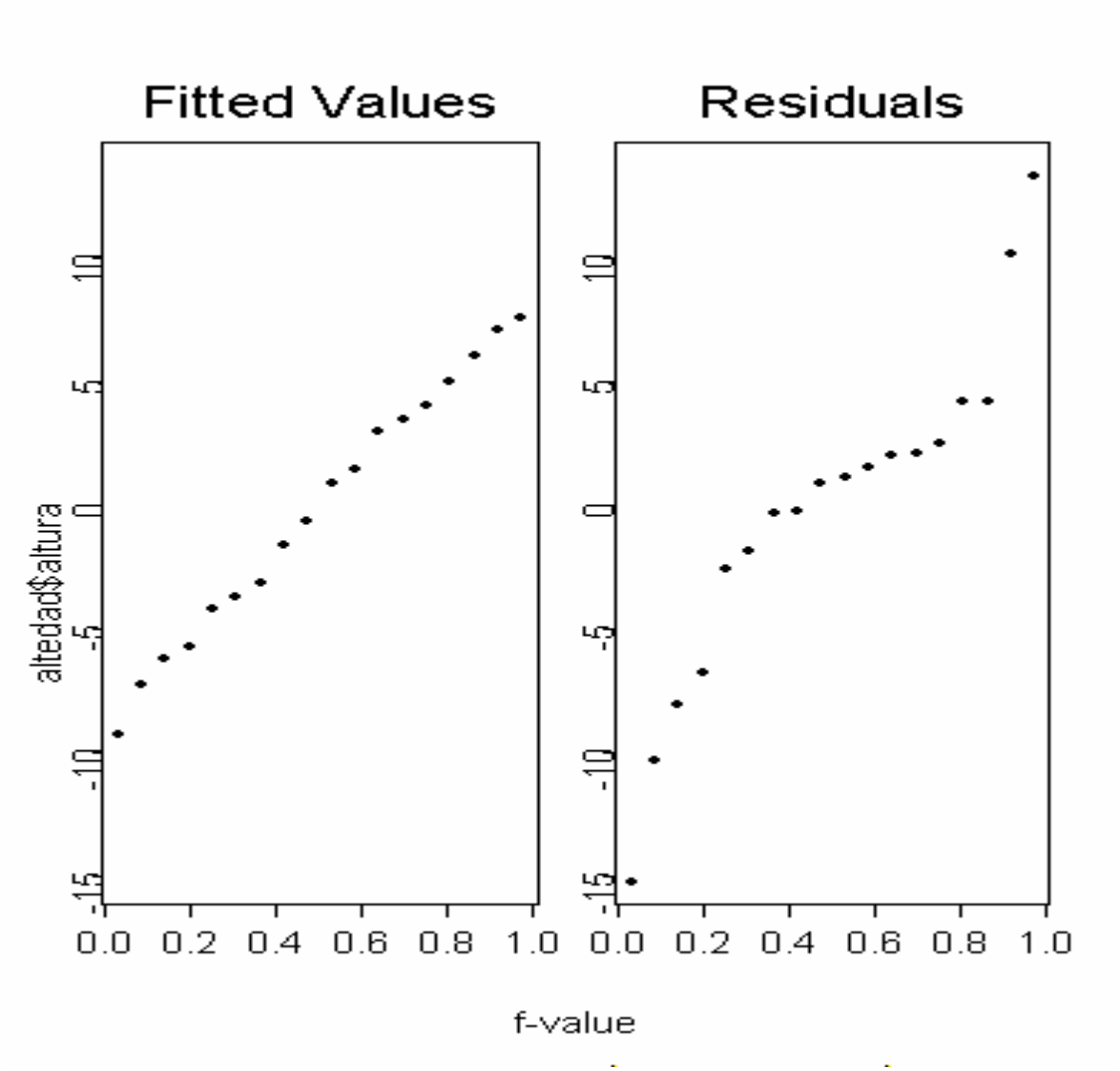
```
> a<-lm(altedad$altura~altedad$edad)

> residuals(a,type="pearson")

       1        2         3        4         5        6
2.173504 10.32836 -1.694217 1.694496 -7.839364 4.349349
       7        8        9       10       11       12
-6.561938 -10.0958 4.381628 2.647767 1.336481 1.10262
      13          14       15         16       17
-15.00867 -0.01995356 2.257473 -0.1651009 13.51233
      18
-2.398961
```

Coeficientes y medidas de ajuste de la recta de cuadrados mínimos

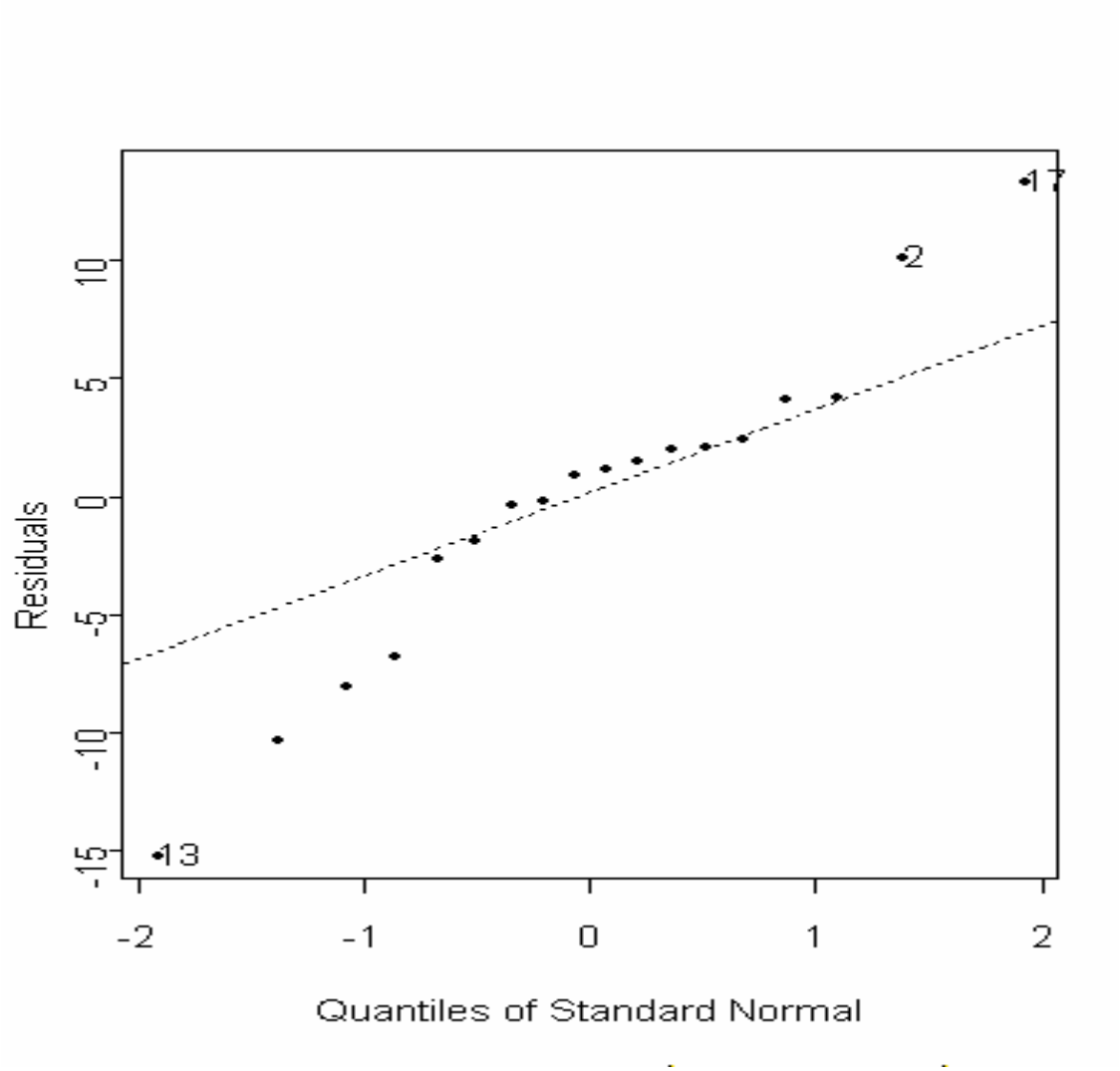
```
> summary(lm(altedad$altura~altedad$edad)

Call: lm(formula = altedad$altura ~
altedad$edad)
Residuals:
    Min     1Q Median   3Q   Max
 -15.01 -2.223   1.22 2.55 13.51

Coefficients:
              Value Std. Error t value Pr(>|t|)
 (Intercept) 79.6962 21.2511     3.7502  0.0017
altedad$edad  0.5113  0.1670     3.0608  0.0075

Residual standard error: 7.029 on 16 degrees
of freedom
Multiple R-Squared: 0.3693
F-statistic: 9.369 on 1 and 16 degrees of
freedom, the p-value is 0.007468
```

La figura 4 muestra los residuos correspondientes al ajuste de la recta de cuadrados mínimos en función de la edad. A pesar de ser muy similar a la figura 3, muestra una leve tendencia hacia abajo. Es decir que los residuos de cuadrados mínimos podrían estar más cercanos a la horizontal si les extrajéramos una recta con una leve pendiente negativa.

En este ejemplo, la variabilidad de los residuos llama más la atención que la diferencia entre las pendientes de ambos métodos. La tabla 3 muestra que el desvío estándar de los residuos es 7.03 y el error estándar de la pendiente es 0.167, alrededor del doble de la diferencia entre las pendientes.



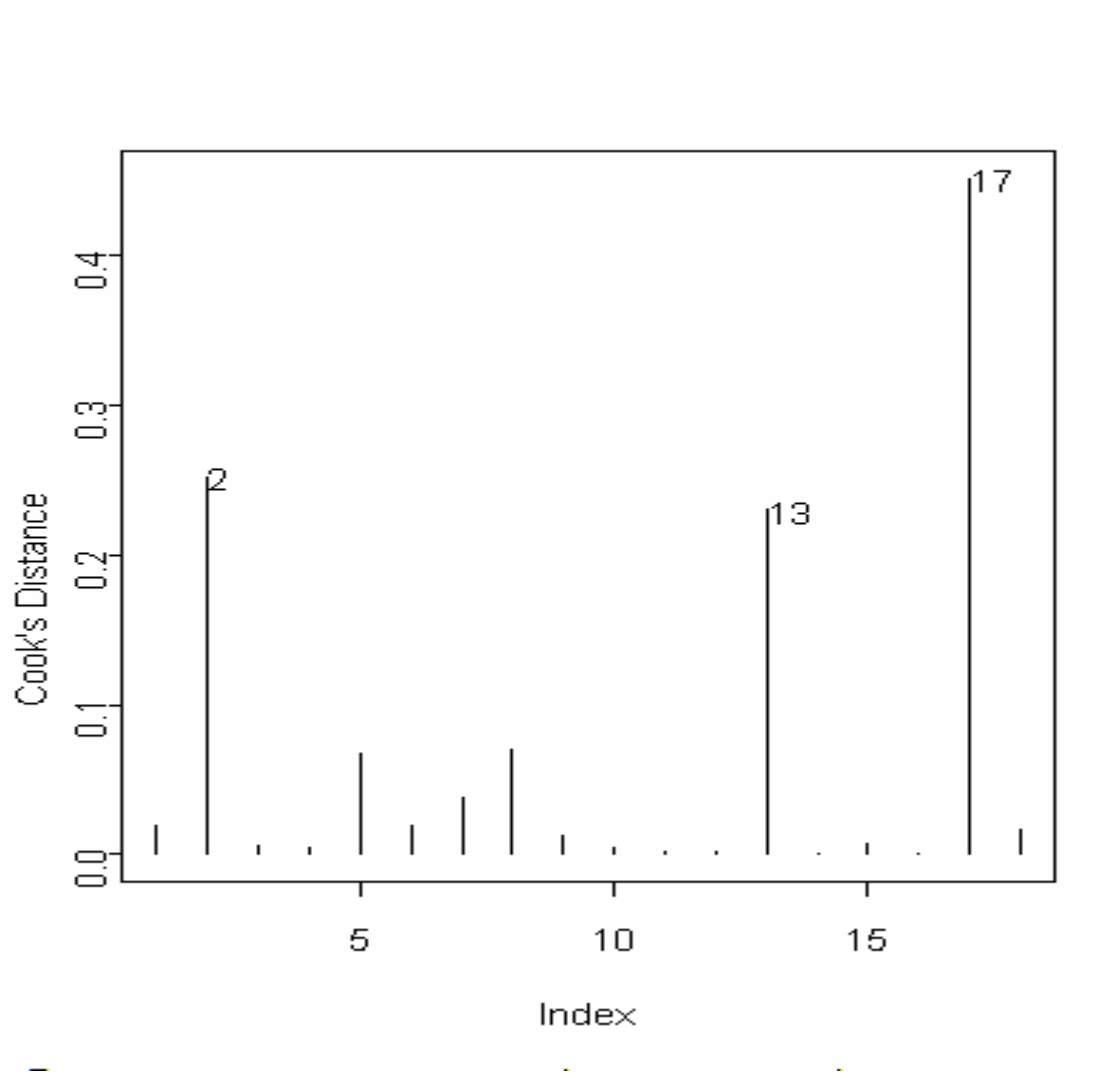
Vemos la gran variabilidad de los residuos comparada con la variabilidad de la variable que queremos explicar (altura). En un buen ajuste se espera que la variabilidad de los residuos sea menor que la variabilidad observada de la variable respuesta.



En el gráfico cuantil-cuantil de los residuos se destacan los puntos 13 y 17.

Vemos, cualitativamente como algunos datos afectan al método de cuadrados mínimos más que al método resistente. Sería muy fácil construir un ejemplo que enfatice la no resistencia del método de cuadrados mínimos. Sin embargo es útil ver que cuando el comportamiento de los puntos es razonable las dos rectas son similares.

El siguiente gráfico permite detectar puntos influyentes.



Recta por medianas repetidas.

El procedimiento, propuesto por Siegel (1982), consiste en estimar la pendiente de la recta en dos etapas. En la primera etapa tomamos la mediana de las pendientes de las *n-1* rectas que pasan por un punto dado (x_i, y_i), $i = 1,...,n;$ y en la segunda etapa se toma la mediana de estas n pendientes. Es decir:

si definimos

$$b_{ij} = \frac{y_i - y_j}{x_i - x_j}$$

entonces la mediana de las pendientes de todas las rectas que pasan por (x_i, y_i) y algún otro punto será

$$\underset{j \neq i}{\text{med}}\{b_{ij}\}$$

luego a través de los puntos tendremos

$$b_{RM} = \underset{i}{\text{med}}\{\underset{j \neq i}{\text{med}}\{b_{ij}\}\}$$

Para ajustar la ordenada al origen se calcula

$$a_i = y_i - b_{RM} x_i$$

y luego

$$a_{RM} = \underset{i}{\text{med}}\{a_i\}.$$

Siegel muestra que el método tiene punto de ruptura cercano a 1/2.

Veamos una derivación heurística para el caso $n = 2k$. En este caso el punto de ruptura exacto es $(k\text{-}1)/n$. Para ver esto supongamos que k -1 de los datos son “salvajes” y que los restantes k+1 son “buenos”. Definimos

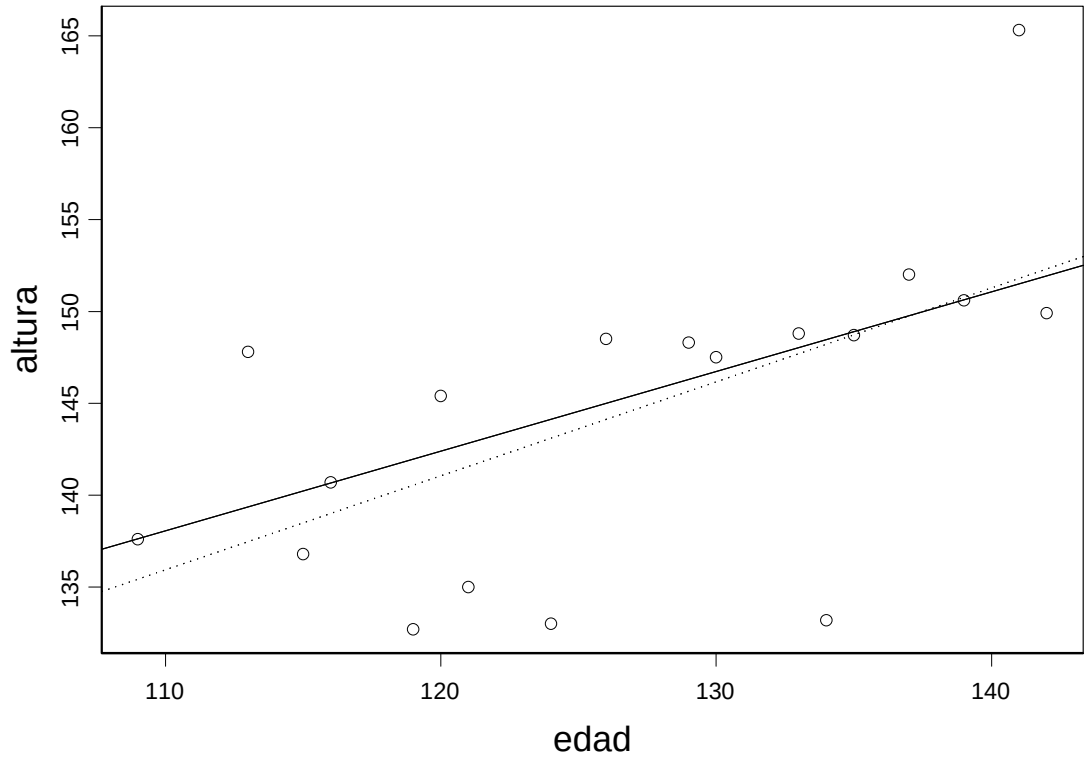
$$b_i = \underset{j \neq i}{\text{med}}\{b_{ij}\}$$

Si $i0$ indica un punto “bueno”, $b_{i\,0}$ está determinado por los restantes k puntos “buenos” y no por los k -1 puntos “salvajes”. Por otro lado b_i para un punto “salvaje” debe ser “salvaje”. Exactamente k + 1 de los b_i son “buenos” y estas estimaciones de la pendiente determinan b_{RM} .

Cualquier número mayor de puntos salvajes” causaría la ruptura de b_{RM}. Como la ordenada al origen involucra únicamente una mediana simple, tolera $k\text{-}1$ puntos salvajes entre $2k$. Tanto la pendiente como la ordenada al origen tienen el mismo punto de ruptura.

EJEMPLO (cont) **Cálculo de la pendiente y ordenada al origen por el método de medianas repetidas**

Las funciones "`repmedians`" escrita con ciclos y "`repmedianas`" que utiliza cálculos matriciales, dan el mismo resultado. Están descriptas en la práctica 8.




```
> attach(altedad)

> unlist(repmedianas(edad, altura))
                                  #cálculo matricial
 ord.origen pendiente
       90.4 0.4333333

> lm(altura ~ edad)[[1]] # Cuadrados Mínimos
 (Intercept)      edad
    79.69623 0.5112868

> plot(edad, altura)
> abline(unlist(repmedianas(edad, altura)))
> abline(lm(altura ~ edad), lty = 2)
```