

### 3. Diagnóstico en Regresión

Las técnicas del diagnóstico en regresión se abocan a validar que los supuestos realizados por el modelo sean apropiados para los datos con los que se cuenta. Son realizadas a posteriori del ajuste (aunque filosóficamente se deberían realizar antes) y están basadas en general en los residuos (o versiones apropiadamente escaladas) de ellos. Constan principalmente de técnicas gráficas, aunque también en la exhibición de algunas medidas de bondad de ajuste. Si el modelo propuesto, una vez ajustado a los datos, no proporciona residuos que parezcan razonables, entonces comenzamos a dudar de que algún aspecto del modelo (o todos) sea apropiado para nuestros datos. Un tema relacionado es asegurarse que la estimación realizada no sea tremadamente dependiente de un sólo dato (o un pequeño subconjunto de datos) en el sentido en que si no se contara con dicho dato las conclusiones del estudio serían completamente diferentes. La identificación de estos puntos influyentes forma parte relevante del diagnóstico (y de esta sección).

#### 3.1. Medidas de diagnóstico

##### 3.1.1. Leverage de una observación

El valor predicho de un dato puede escribirse como combinación lineal de las observaciones

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i = \sum_{k=1}^n h_{ik} Y_k \quad (33)$$

donde

$$h_{ik} = \frac{1}{n} + \frac{(X_i - \bar{X})(X_k - \bar{X})}{S_{XX}}$$

y como caso particular tenemos que

$$h_{ii} = \frac{1}{n} + \frac{(X_i - \bar{X})^2}{S_{XX}}. \quad (34)$$

Recordemos que hemos llamado  $S_{XX}$  a la cantidad

$$S_{XX} = \sum_{k=1}^n (X_k - \bar{X})^2.$$

Vale que

$$\sum_{k=1}^n h_{ik} = 1, \quad \sum_{i=1}^n h_{ik} = 1 \quad (35)$$

$$\begin{aligned} \sum_{i=1}^n h_{ii} &= 2 \\ \frac{1}{n} \leq h_{ii} &\leq \frac{1}{s} \leq 1. \end{aligned} \quad (36)$$

donde  $s$  es la cantidad de observaciones con predictor igual a  $X_i$  en la muestra. La cantidad  $h_{ii}$  se denomina *leverage del dato i-ésimo*. Es una medida que resume cuán lejos cae el valor de  $X_i$  de la media muestral de las  $X$ . Mide, de alguna manera, cuánto es el aporte de la observación i-ésima a la varianza muestral de las  $X$  (que es  $\frac{s_{xx}}{n-1}$ ). La traducción de leverage al castellano es usualmente palanca, o influencia. Observemos que es un concepto que no depende del valor  $Y_i$  observado.

### 3.1.2. Residuos

Dijimos en la Sección 2.8 que los residuos son cantidades observables, que representan de alguna manera el correlato empírico de los errores. Para verificar los supuestos del modelo lineal, suelen usarse métodos gráficos que involucran a los residuos. El modelo lineal

$$Y = \beta_0 + \beta_1 X + \varepsilon$$

supone que los errores  $\varepsilon$  tienen media poblacional cero y varianza constante (que denominamos  $\sigma^2$ ), y que son independientes para distintas observaciones. Sin embargo, ya hemos visto que no ocurre lo mismo con los residuos. Vimos que los residuos no son independientes. Además, puede probarse que

$$\begin{aligned} E(e_i) &= 0 \\ Var(e_i) &= \sigma^2(1 - h_{ii}) \end{aligned} \quad (37)$$

donde  $h_{ii} = \frac{1}{n} + \frac{(x_i - \bar{X})^2}{s_{xx}}$ , es el leverage de la observación i-ésima. En consecuencia la varianza del residuo de un dato depende del valor de la covariable, y los residuos de distintos casos tienen diferentes varianzas. De la ecuación (37) vemos que cuánto mayor sea  $h_{ii}$ , menor será la varianza del  $e_i$ : mientras más cercano a uno sea  $h_{ii}$  más cercana a cero será la varianza del residuo de la observación i-ésima. Esto quiere decir que para observaciones con gran  $h_{ii}$ ,  $\hat{Y}_i$  tenderá a estar cerca del valor observado  $Y_i$ , sin importar cuánto sea el valor  $Y_i$  observado. En el caso extremo e hipotético en que  $h_{ii} = 1$ , la recta ajustada sería forzada a pasar por el valor observado  $(X_i, Y_i)$ .

### 3.1.3. Residuos estandarizados

Para hacer más comparables a los residuos entre sí, podemos dividir a cada uno de ellos por un estimador de su desvío estándar, obteniendo lo que se denominan *residuos estandarizados*:

$$rest_i = \frac{e_i}{\sqrt{\hat{\sigma}^2(1-h_{ii})}}. \quad (38)$$

Recordemos que el estimador de  $\sigma^2$  bajo el modelo de regresión está dado por

$$\hat{\sigma}^2 = \frac{SSRes}{n-2}$$

Puede probarse que los residuos estandarizados tienen media poblacional cero (igual que los residuos), y varianza poblacional igual a uno, es decir

$$\begin{aligned} E(rest_i) &= 0 \\ Var(rest_i) &= 1, \quad \text{para todo } i. \end{aligned}$$

### 3.1.4. Los residuos cuando el modelo es correcto

Para chequear que los supuestos del modelo lineal son apropiados para un conjunto de datos, suelen hacerse una serie de gráficos. El más importante es el scatter plot de residuos versus la covariante. Esto se conoce como gráfico de residuos (o residual plot). En el caso de regresión lineal simple, los valores ajustados o predichos  $\hat{Y}_i$  representan un cambio de escala lineal respecto de los valores  $X_i$  ya que  $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ . Luego, es equivalente al gráfico recién descrito el scatter plot de residuos versus los valores ajustados. ¿Cómo debe lucir este gráfico si el modelo es correcto?

1. Puede probarse que  $E(e | X_1, \dots, X_n) = 0$ . Esto quiere decir que el scatter plot de los residuos versus las  $X$  debe estar centrado alrededor del cero (de la recta horizontal de altura cero).
2. Mencionamos que cuando el modelo es correcto,  $Var(e_i | X_1, \dots, X_n) = \sigma^2(1-h_{ii})$ . Luego el gráfico de residuos versus la covariante debería mostrar menor variabilidad para los valores de  $X$  más alejados de la media muestral (serán los que tengan mayor leverage  $h_{ii}$ ). Por este motivo, suele ser más frecuente graficar los residuos estandarizados versus la covariante. En ese caso, deberíamos ver la misma variabilidad para los distintos valores de la covariante.
3. Los residuos de distintas observaciones están correlacionados entre sí, pero esta correlación no es muy importante, no será visible en los gráficos de residuos.

En resumen, si el modelo es correcto, el gráfico de los residuos versus predichos o versus la covariable debería lucir como una nube de puntos sin estructura, ubicada alrededor del eje horizontal.

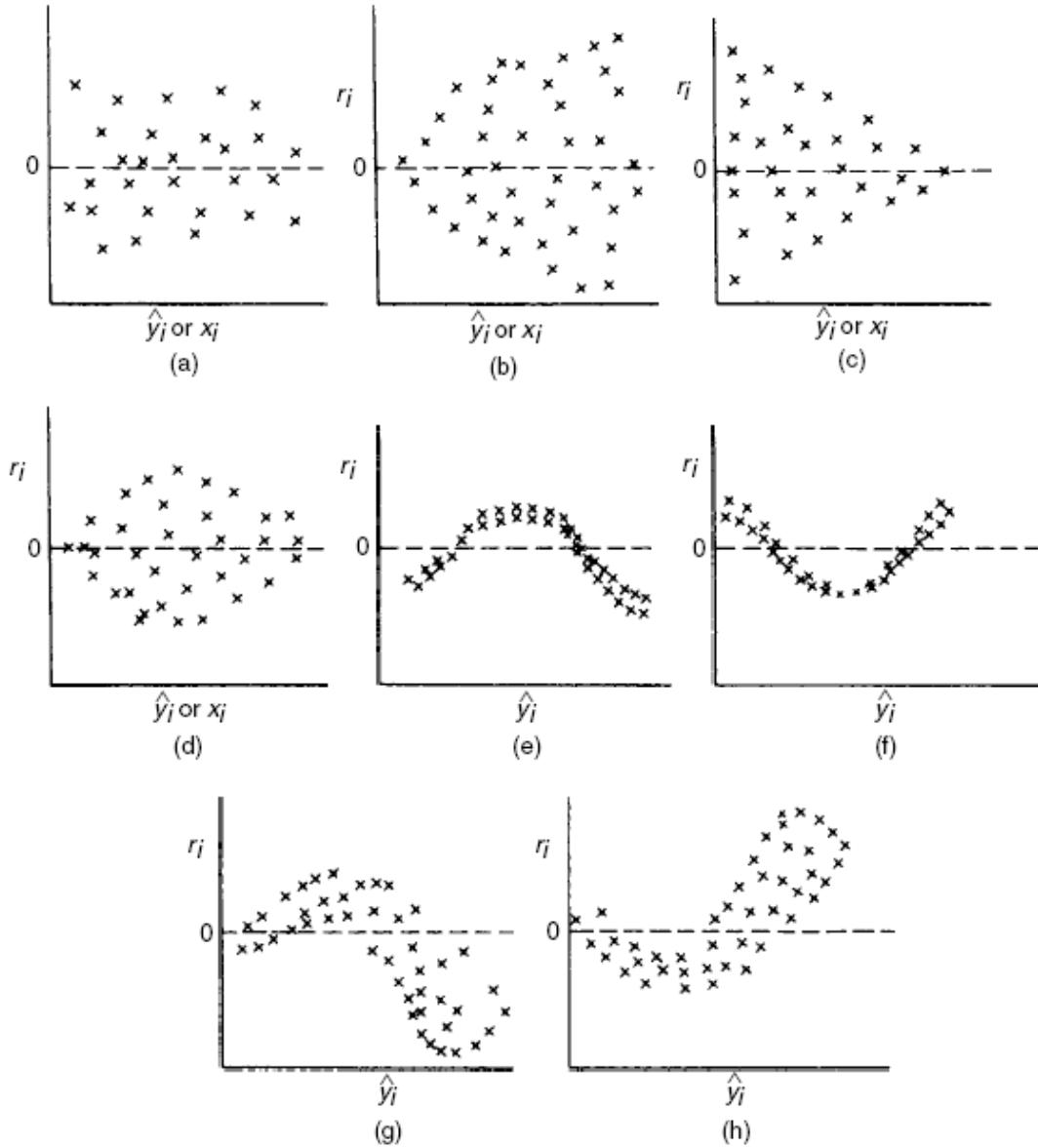
### 3.1.5. Los residuos cuando el modelo es incorrecto

En el caso en el que el modelo es incorrecto, el gráfico de residuos (o de residuos estandarizados) versus la variable predictora (o versus los valores predichos) suele tener algún tipo de estructura. En la Figura 25 se ven varios de estos posibles scatter plots (algo idealizados, claro).

El primero de ellos es una nube de puntos sin estructura que indica que no hay problemas con el modelo ajustado. De las Figuras 25(b) a 25(d) inferiríamos que el supuesto de homogeneidad de varianzas no se satisface: la varianza depende de la variable graficada en el eje horizontal. Las Figuras 25(e) a 25(h) son indicadoras de que se viola el supuesto de linealidad de la esperanza condicional, lo cual nos lleva a pensar que el vínculo entre el valor esperado de la variable respuesta  $Y$  y la covariable se ve mejor modelado por una función más complicada que la lineal (lo que genéricamente suele denominarse una curva). Las dos últimas figuras, las 25(g) y 25(h) sugieren la presencia simultánea de curvatura y varianza no constante.

En la práctica, los gráficos de residuos no son tan claros como estos... Es útil recordar que aún cuando todos los datos satisficieran todos los supuestos, la variabilidad muestral podría hacer que el gráfico tuviera pequeños apartamientos de la imagen ideal.

Figura 25: Gráficos de residuos: (a) nube de datos sin estructura, (b) varianza que crece con  $X$  (forma de megáfono abierto a la derecha), (c) varianza que decrece con  $X$  (forma de megáfono abierto a la izquierda), (d) varianza que depende de la covariable, (e)-(f) no linealidad, (g)-(h) combinación de no linealidad y función de varianza no constante. Fuente: Weisberg [2005], pág. 172.



### 3.1.6. Los residuos en el ejemplo

La Figura 26 muestra el gráfico de residuos en el ejemplo de los 100 bebés de bajo peso. Por ejemplo, el primer dato observado ( $i = 1$ ) corresponde a un bebé de 29 semanas de gestación cuyo perímetro cefálico fue de 27 cm. El valor predicho para este caso es

$$\hat{Y}_1 = 3,9143 + 0,7801 \cdot 29 = 26,537$$

y el residuo asociado a esa observación es

$$e_1 = Y_1 - \hat{Y}_1 = 27 - 26,537 = 0,463,$$

como ya habíamos calculado. Luego, el punto  $(26,537, 0,463)$  será incluido en el gráfico, que es un scatter plot de los puntos  $(\hat{Y}_i, e_i)$  para las 100 observaciones de la muestra.

En él vemos que hay un residuo en particular que es un poco más grande que el resto: este punto corresponde a la observación 31, que corresponde a un bebé cuya edad gestacional es de 31 semanas y cuyo perímetro cefálico es de 35 centímetros. De acuerdo al modelo, el valor predicho para su perímetro cefálico sería

$$\hat{Y}_{31} = 3,9143 + 0,7801 \cdot 31 = 28,097$$

un valor mucho menor que el observado, por lo tanto el residuo resulta grande

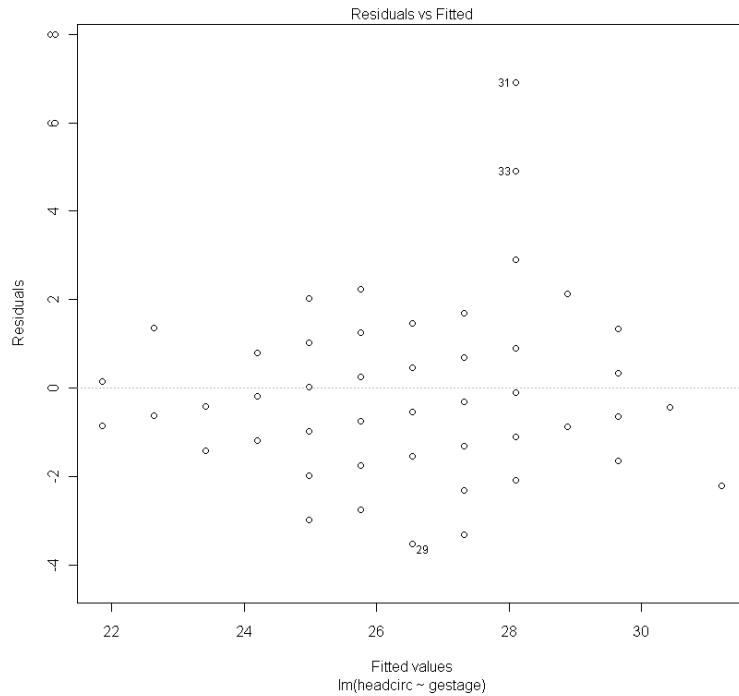
$$e_{31} = Y_{31} - \hat{Y}_{31} = 35 - 28,097 = 6,903.$$

Podemos probar sacar este punto de la muestra, volver a realizar el ajuste y luego comparar los dos modelos para medir el efecto del punto en la estimación de los coeficientes de la recta. No lo haremos aquí puesto que en las secciones subsiguientes propondremos otros modelos que ajustarán mejor a nuestros datos. En cuanto al gráfico de residuos, este no muestra evidencia de que el supuesto de homoscedasticidad sea violado, o que haya algún tipo de curvatura en el vínculo entre los residuos y los predichos, indicando que el modelo ajusta bien a los datos.

### 3.1.7. ¿Cómo detectar (y resolver) la curvatura?

Para ayudarnos a decidir si un gráfico de residuos corresponde (o no) a una nube de puntos es posible hacer un test de curvatura. El más difundido es el test de no aditividad de Tuckey, que no describiremos aquí. Sin embargo, sí diremos que un remedio posible al problema de la curvatura consiste en transformar alguna de las variables  $X$  o  $Y$  (o ambas), y luego proponer un modelo lineal para las variables transformadas. Hay técnicas que ayudan a decidir qué transformaciones de los datos puede ser interesante investigar. Las transformaciones de Box-Cox son las más difundidas de estas técnicas, ver Kutner, Nachtsheim, Neter, y Li [2005].

Figura 26: Gráfico de residuos versus valores ajustados para el ajuste lineal de perímetro cefálico en función de la edad gestacional, en el caso de los 100 bebés de bajo peso.



Otra posibilidad consiste en proponer modelos más complejos que contemplen un vínculo más general entre  $X$  e  $Y$ , por ejemplo

$$E(Y | X) = \beta_0 + \beta_1 X + \beta_2 X^2.$$

Es posible estudiar estos modelos como un caso particular de los modelos de regresión lineal, pero con dos covariables ( $X$  y  $X^2$ ), lo cual nos lleva a tratarlos dentro de los modelos de regresión múltiple, que presentaremos más adelante.

### 3.1.8. ¿Qué hacer si la varianza no es constante?

En un gráfico de residuos, una función de la varianza no constante puede indicar que el supuesto de varianza constante es falso. Hay por lo menos cuatro remedios básicos en este caso, que describiremos siguiendo a Weisberg [2005], Sección 8.3. El primero es el uso de una *transformación estabilizadora de la varianza* para transformar a las  $Y$ , ya que el reemplazo de  $Y$  por  $Y_{\text{transformada}}$  puede inducir

varianza constante en la escala transformada. Una segunda opción es encontrar los pesos que podrían ser utilizados en los *mínimos cuadrados ponderados*. El método de mínimos cuadrados ponderados o pesados es una técnica estadística que ataca una versión más general del problema de regresión que hemos descrito hasta ahora. Lo presentamos a continuación, en su caso más simple. Seguimos trabajando bajo el supuesto de linealidad de la esperanza

$$E(Y | X = x_i) = \beta_0 + \beta_1 x_i,$$

pero ahora relajamos el supuesto de que la función de varianza  $Var(Y | X)$  sea la misma para todos los valores de  $X$ . Supongamos que podemos asumir que

$$Var(Y | X = x_i) = Var(\varepsilon_i) = \frac{\sigma^2}{w_i}$$

donde  $w_1, \dots, w_n$  son números positivos conocidos. La función de varianza todavía queda caracterizada por un único parámetro desconocido  $\sigma^2$ , pero las varianzas pueden ser distintas para distintos valores de  $X$ . Esto nos lleva al método de mínimos cuadrados pesados o ponderados (en inglés *weighted least squares*, o wls) en vez del método usual de mínimos cuadrados (*ordinary least squares*, ols) para obtener estimadores. En este caso, se buscan los valores de los parámetros que minimizan la función

$$g_{wls}(a, b) = \sum_{i=1}^n w_i (Y_i - (a + bX_i))^2.$$

Existen expresiones explícitas para los parámetros estimados con este método, y los softwares más difundidos realizan el ajuste. En las aplicaciones, por supuesto, se agrega la complejidad extra de elegir los pesos  $w_i$  que en general no vienen con los datos. Muchas veces se usan pesos empíricos, que se deducen de algunos supuestos teóricos que se tengan sobre las variables, por ejemplo. Si hubiera replicaciones, es decir varias mediciones de la variable respuesta realizadas para el mismo valor de la covariable, podría estimarse la varianza dentro de cada grupo y conseguirse de este modo pesos aproximados. También es posible usar *modelos de mínimos cuadrados generalizados*, en los que se estiman simultáneamente los parámetros del modelo y los pesos, que exceden por mucho estas notas (consultar por ejemplo Pinheiro y Bates [2000], Sección 5.1.2).

La tercera posibilidad es no hacer nada. Los estimadores de los parámetros, ajustados considerando una función de varianza incorrecta o mal especificada, son de todos modos insesgados, aunque ineficientes. Los tests e intervalos de confianza calculados con la función de varianza errada serán inexactos, pero se puede recurrir a métodos de bootstrapping para obtener resultados más precisos.

La última opción es usar modelos de regresión que contemplan la posibilidad de una función de varianza no constante que dependa de la media. Estos modelos se denominan *modelos lineales generalizados*, de los cuales por ejemplo, los modelos de regresión logística forman parte. Puede consultarse el texto clásico McCullagh y Nelder [1989] y también el libro de Weisberg [2005], Sección 8.3 y Sección 12.

### 3.1.9. ¿Cómo validamos la independencia?

Si las observaciones con las que contamos fueron producto de haber tomado una muestra aleatoria de sujetos de alguna población, entonces en principio, tendremos observaciones independientes. Algunas situaciones en las que este supuesto puede fallar se describen a continuación.

Los estudios en los cuales los datos se recolectan secuencialmente pueden dar lugar a observaciones que no resulten independientes. Lo mismo puede suceder en las determinaciones de laboratorio hechas secuencialmente en el tiempo, ya que pueden mostrar un cierto patrón, dependiendo de cómo funcionan los equipos, los observadores, etc. El modo de detección de estas situaciones suele ser graficar los residuos versus la secuencia temporal en la que fueron relevados.

Si los datos fueron obtenidos por dos observadores distintos A y B, podríamos esperar que las observaciones de un observador tiendan a parecerse más entre ellas. La manera de detectar que esto sucede es graficar las  $Y$  versus las  $X$  identificando los puntos de cada grupo. En ocasiones, la variabilidad debida a la regresión puede ser explicada por la pertenencia al grupo. Tampoco serán independientes las observaciones si varias de ellas fueron realizadas sobre los mismos sujetos (o animales). Si este fuera el caso, puede considerarse un modelo de regresión múltiple donde el operador (o el sujeto) entre como covariante. Nos ocuparemos de discutir esto más adelante, ya que los modelos correctos para este tipo de situaciones son los modelos de ANOVA con efectos aleatorios, o los modelos de efectos mixtos, que exceden el contenido de estas notas. Ver para ello, el libro de Pinheiro y Bates [2000].

### 3.1.10. ¿Cómo validamos la normalidad?

El supuesto de normalidad de los errores juega un rol menor en el análisis de regresión. Es necesario para realizar inferencias en el caso de muestras pequeñas, aunque los métodos de bootstrap (o resampling) pueden usarse si este supuesto no está presente. El problema con las muestras pequeñas es que chequear el supuesto de normalidad a través de los residuos cuando no hay muchas observaciones es muy difícil. Los gráficos cuantil cuantil de los residuos (qq-plots) y los tests de normalidad realizados sobre ellos pueden ayudar en esta tarea. Hay varios tests posibles que ayudan a descartar la normalidad, entre ellos el test de Shapiro-

Wilks (que está esencialmente basado en el cuadrado de la correlación entre las observaciones ordenadas y sus valores esperados bajo normalidad), o el test de Kolmogorov-Smirnov, que están implementados en los paquetes.

En la práctica los supuestos de normalidad y homoscedasticidad nunca se cumplen exactamente. Sin embargo, mientras más cerca estén nuestros datos de los supuestos del modelo lineal, más apropiados serán los tests e intervalos de confianza que construyamos.

Para muestras grandes el supuesto de distribución normal no es crucial. Una versión extendida del Teorema Central del Límite garantiza que el estimador de mínimos cuadrados de la pendiente tiene distribución de muestreo aproximadamente normal cuando  $n$  es grande.

## 3.2. Outliers y observaciones influyentes

### 3.2.1. Outliers

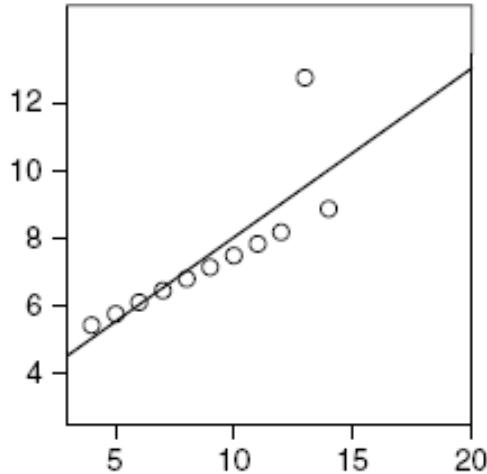
En algunos problemas, la respuesta observada para algunos pocos casos puede parecer no seguir el modelo que sí ajusta bien a la gran mayoría de los datos. Un ejemplo (de datos ficticios) puede verse en el scatter plot de la Figura 27. Los datos de este ejemplo sugieren que el modelo lineal puede ser correcto para la mayoría de los datos, pero uno de los casos está muy alejado de lo que el modelo ajustado le prescribe. Diremos que este dato alejado es un outlier. Observemos que el concepto de outlier (o sea, dato atípico) es un concepto relativo al modelo específico en consideración. Si se modifica la forma del modelo propuesto a los datos, la condición de ser outlier de un caso individual puede modificarse. O sea, un **outlier** es un caso que no sigue el mismo modelo que el resto de los datos. La identificación de estos casos puede ser útil. ¿Por qué? Porque el método de cuadrados mínimos es muy sensible a observaciones alejadas del resto de los datos. De hecho, las observaciones que caigan lejos de la tendencia del resto de los datos pueden modificar sustancialmente la estimación.

### 3.2.2. Un test para encontrar outliers

Si sospechamos que la observación  $i$ -ésima es un outlier podemos proceder del siguiente modo. Este procedimiento es clásico dentro de la regresión y corresponde a muchos otros procedimientos en estadística que son genéricamente conocidos como “*leave one out procedures*”.

1. Eliminamos esa observación de la muestra, de modo que ahora tenemos una muestra con  $n - 1$  casos.

Figura 27: Datos hipotéticos que muestran el desajuste de una observación al modelo ajustado.



2. Usando el conjunto de datos reducidos volvemos a estimar los parámetros, obteniendo  $\hat{\beta}_{0(i)}, \hat{\beta}_{1(i)}$  y  $\hat{\sigma}_{(i)}^2$  donde el subíndice  $(i)$  está escrito para recordarnos que los parámetros fueron estimados sin usar la  $i$ -ésima observación.
3. Para el caso omitido, calculamos el valor ajustado  $\hat{Y}_{i(i)} = \hat{\beta}_{0(i)} + \hat{\beta}_{1(i)}X_i$ . Como el caso  $i$ -ésimo no fue usado en la estimación de los parámetros,  $Y_i$  y  $\hat{Y}_{i(i)}$  son independientes. La varianza de  $Y_i - \hat{Y}_{i(i)}$  puede calcularse y se estima usando  $\hat{\sigma}_{(i)}^2$ .
4. Escribamos

$$t_i = \frac{Y_i - \hat{Y}_{i(i)}}{\sqrt{\widehat{Var}(Y_i - \hat{Y}_{i(i}))}},$$

la versión estandarizada del estadístico en consideración. Si la observación  $i$ -ésima sigue el modelo, entonces la esperanza de  $Y_i - \hat{Y}_{i(i)}$  debería ser cero. Si no lo sigue, será un valor no nulo. Luego, si llamamos  $\delta$  a la esperanza poblacional de esa resta,  $\delta = E(Y_i - \hat{Y}_{i(i)})$ , y asumimos normalidad de los errores, puede probarse que la distribución de  $t_i$  bajo la hipótesis  $H_0 : \delta = 0$  es una  $t$  de Student con  $n - 3$  grados de libertad,  $t_i \sim t_{n-3}$  (recordar que hemos excluido una observación para el cálculo del error estándar que figura en el denominador, por eso tenemos un grado de libertad menos que con

los anteriores tests), y rechazar cuando este valor sea demasiado grande o demasiado pequeño.

Hay una fórmula computacionalmente sencilla para expresar a  $t_i$  sin necesidad de reajustar el modelo lineal con un dato menos, ya que es fácil escribir al desvío estándar estimado sin la observación i-ésima ( $\hat{\sigma}_{(i)}$ ) en términos del leverage de la observación i-ésima ( $h_{ii}$ ) y el desvío estándar estimado con toda la muestra ( $\hat{\sigma}$ ). Es la siguiente

$$t_i = \frac{e_i}{\hat{\sigma}_{(i)}\sqrt{1-h_{ii}}} = rest_i \sqrt{\frac{n-3}{n-2-rest_i}} \quad (39)$$

donde el residuo estandarizado  $rest_i$  lo definimos en la ecuación (38). Esta cantidad se denomina el **residuo estudentizado** i-ésimo. La ecuación (39) nos dice que los residuos estudentizados y los residuos estandarizados llevan la misma información, ya que pueden ser calculados uno en función de otro. Vemos entonces que para calcular los residuos estudentizados no es necesario descartar el caso i-ésimo y volver a ajustar la regresión (cosa que tampoco nos preocuparía mucho ya que es la computadora la que realiza este trabajo).

Para completar el test, nos queda únicamente decidir contra qué valor comparar el  $t_i$  para decidir si la i-ésima observación es o no un outlier. Si el investigador sospecha de antemano a realizar el ajuste que la observación i-ésima es un outlier lo justo sería comparar el valor absoluto de  $t_i$  con el percentil  $1 - \frac{\alpha}{2}$  de la  $t$  de student con  $n - 3$  grados de libertad. Pero es rara la ocasión en la que se sospecha de un dato antes de hacer el análisis. Si en cambio el analista hace el ajuste, luego computa los residuos estudentizados, y, a raíz de lo obtenido sospecha de aquella observación con mayor valor absoluto de  $t_i$ , entonces en el fondo está realizando  $n$  tests de significatividad, uno para cada observación. Para tener controlada la probabilidad de cometer un error de tipo I en alguno de los  $n$  tests (es decir, decidir falsamente que una observación que en realidad no es outlier sea declarada como tal), puede usarse un procedimiento conservativo conocido como método de Bonferroni para comparaciones múltiples. Este procedimiento propone rechazar  $H_0 : \text{ninguna de las } n \text{ observaciones es un outlier}$ , versus  $H_1 : \text{hay al menos un outlier}$ , cuando alguno de los  $|t_i|$  es mayor que el percentil  $1 - \frac{\alpha}{2n}$  de la  $t_{n-3}$ . Por ejemplo, si  $n = 20$  (pensamos en una muestra con 20 observaciones) y nivel simultáneo 0,05, entonces en vez de comparar con el percentil 0,975 de una  $t_{17}$  que es 2,11, la comparación correcta es con el percentil  $1 - \frac{\alpha}{2n} = 1 - \frac{0,05}{2 \cdot 20} = 0,99875$  de una  $t_{17}$  que es 3,543.

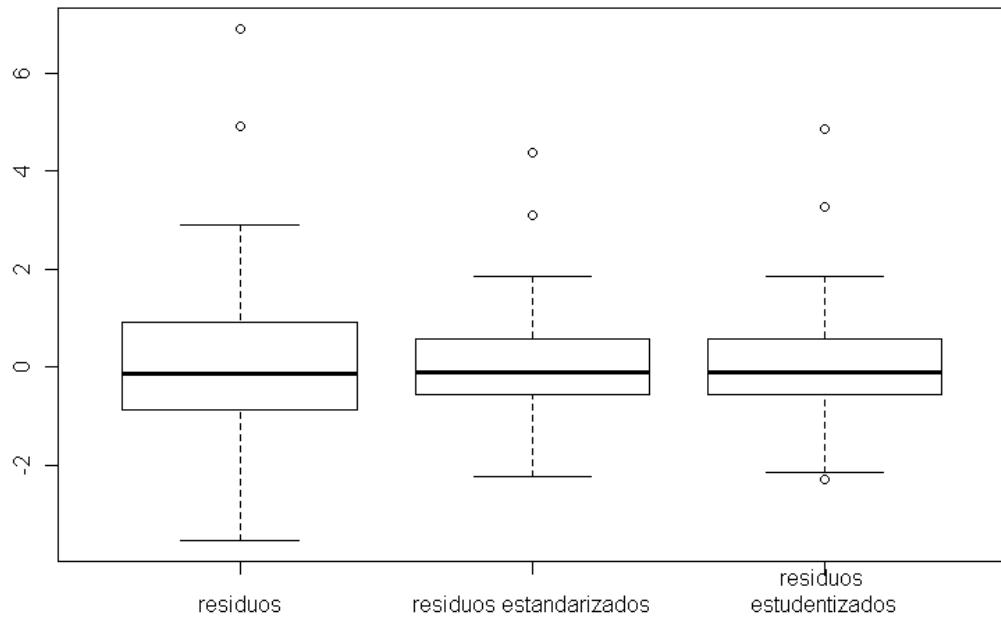
Apliquemos este test al ejemplo de los bebés de bajo peso.

**Ejemplo 3.1** En el caso de los 100 bebés, para detectar outliers a nivel 0,05 debemos computar el residuo estudentizado para cada caso, y compararlo con el percentil

$$1 - \frac{\alpha}{2n} = 1 - \frac{0,05}{2 \cdot 100} = 0,99975$$

de una  $t_{97}$ , que resulta ser 3,602. El único residuo estudentizado cuyo valor absoluto sobrepasa este punto de corte es el correspondiente a la observación 31, que es 4,857. En la Figura 28 pueden verse los boxplots de los residuos, los residuos estandarizados y los residuos estudentizados para el ajuste de perímetrocefálico en función de la edad gestacional.

Figura 28: Los boxplots de los residuos, los residuos estandarizados y los residuos estudentizados para el ajuste de perímetrocefálico en función de la edad gestacional en el ejemplo.



Este test ubica un outlier, pero no nos dice qué hacer con él. Cuando detectamos un outlier, sobre todo si es severo, es importante investigarlo. Puede tratarse de un dato mal registrado, o que fue mal transcripto a la base de datos. En tal caso podremos eliminar el outlier (o corregirlo) y analizar los casos restantes. Pero si el dato es correcto, quizás sea diferente de las otras observaciones y encontrar las causas de este fenómeno puede llegar a ser la parte más interesante del análisis. Todo esto depende del contexto del problema que uno esté estudiando. Si el dato es

correcto y no hay razones para excluirlo del análisis entonces la estimación de los parámetros debería hacerse con un método robusto, que a diferencia de mínimos cuadrados, no es tan sensible a observaciones alejadas de los demás datos.

Antes de terminar, correspondería hacer un alerta. Los residuos estudentizados son una herramienta más robusta que los residuos estandarizados para evaluar si una observación tiene un residuo inusualmente grande. Éste método para detectar outliers parece una estrategia muy apropiada. Y lo es... siempre que en la muestra haya a lo sumo un outlier. Pero, como todos los procedimientos de *leave one out*, puede conducirnos a conclusiones erradas si en la muestra hubiera más de un dato atípico, pues en tal caso, al calcular el residuo estudentizado de la observación  $i$ -ésima, la otra (u otras) observaciones atípicas aún presentes en la muestra podrían tergiversar el ajuste del modelo  $\hat{Y}_{i(i)}$  o la estimación del desvío estándar  $\hat{\sigma}_{(i)}$ , alterando la distribución de los residuos estudentizados resultantes. Por eso, una estrategia todavía mejor para detectar la presencia de outliers que el estudio de los residuos estudentizados, es comparar el ajuste obtenido por cuadrados mínimos con el ajuste al modelo lineal que proporciona un método robusto, como describiremos en la Sección 3.2.4.

### 3.2.3. Observaciones influyentes

Estudiar la influencia de las observaciones es, de alguna manera, estudiar los cambios en el análisis cuando se omiten uno o más datos (siempre una pequeña porción de los datos disponibles). La idea es descubrir los efectos o la influencia que tiene cada caso en particular comparando el ajuste obtenido con toda la muestra con el ajuste obtenido sin ese caso particular (o sin esos pocos casos particulares). Una observación se denomina **influente** si al excluirla de nuestro conjunto de datos la recta de regresión estimada cambia notablemente. Ejemplificaremos los conceptos en forma gráfica.

En la Figura 29 se observan scatter plots de cuatro conjuntos de 18 datos cada uno. En el gráfico (1), el conjunto de datos no presenta ni puntos influyentes ni outliers, ya que todas las observaciones siguen el mismo patrón. En los restantes tres gráficos se conservaron 17 de las observaciones del gráfico (1) y se intercambió una de ellas por los puntos que aparecen indicados como A, B y C en los respectivos scatter plots, y que son puntos atípicos en algún sentido, es decir, puntos que no siguen el patrón general de los datos. No todos los casos atípicos tendrán una fuerte influencia en el ajuste de la recta de regresión.

En la Figura 29 (2), entre las observaciones figura una que rotulamos con la letra A. El caso A puede no ser muy influyente, ya que hay muchos otros datos en la muestra con valores similares de  $X$  que evitarán que la función de regresión se desplace demasiado lejos siguiendo al caso A. Por otro lado, los casos B y C ejercerán una influencia muy grande en el ajuste, ya que como vimos en las Sec-

ciones 3.1.1 y 3.1.2 el leverage de ambas será bastante alto. Mientras mayor sea el leverage de la observación, menor será la variabilidad del residuo, esto quiere decir que para observaciones con gran leverage, el valor predicho tendrá que estar cerca del valor observado. Por eso se dice que tienen un alto grado de apalancamiento, o que cada uno de ellos es un punto de alta palanca. Luego la recta ajustada se verá matemáticamente obligada a acercarse a dichas observaciones, alejándose para ello, de los demás datos.

En la Figura 29 (3) aparece una observación indicada con B. Esta observación será influyente en el ajuste, pero como sigue el patrón lineal de los datos (o sea, sigue la estructura de esperanza condicional de  $Y$  cuando  $X$  es conocida que tienen el resto de los datos) no hará que la recta estimada cuando el punto está en la muestra varíe mucho respecto de la recta estimada en la situación en la que no está, pero reforzará (quizá artificialmente) la fuerza del ajuste observado: reforzará la significatividad de los tests que se hagan sobre los parámetros.

La Figura 29 (4) presenta la observación C. Esta observación será muy influyente en el ajuste, arrastrando a la recta estimada a acercarse a ella. Como no sigue la misma estructura de esperanza condicional que el resto de las observaciones, la recta ajustada en este caso diferirá mucho de la que se ajusta a los datos de la Figura 29 (1). Sin embargo, si una vez realizado el ajuste intentamos identificar este punto mirando las observaciones de mayores residuos (o residuos estandarizados) es posible que no la detectemos (dependerá de cuán extrema sea) ya que al arrastrar la recta hacia ella, tendrá un residuo mucho menor que el que tendría si usáramos la recta que ajusta a los datos del gráfico (1).

Constatemos que lo afirmado antes es cierto, buscando la recta que mejor ajusta a cada conjunto de datos, por mínimos cuadrados. A continuación figuran las salidas del R a los cuatro ajustes, y en la Figura 30 nuevamente los scatter plots de los cuatro conjuntos de datos, con las rectas ajustadas superpuestas.

#### Ajuste de mínimos cuadrados a los datos de la Figura 29 (1)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	6.4063	2.0364	3.146	0.00625
xx	2.3987	0.3038	7.895	6.58e-07

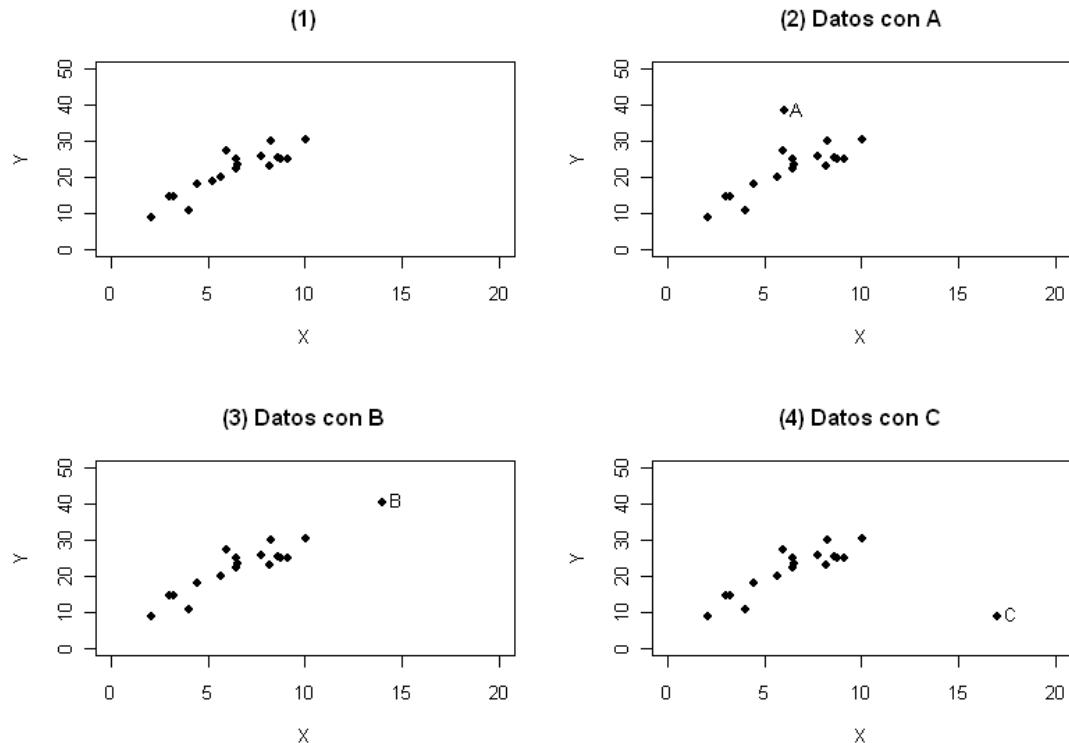
Residual standard error: 2.899 on 16 degrees of freedom

Multiple R-squared: 0.7957, Adjusted R-squared: 0.783

F-statistic: 62.33 on 1 and 16 DF, p-value: 6.579e-07

```
> confint(lm(yy~xx))
```

Figura 29: Scatter plot de 4 conjuntos de datos (hay 18 observaciones en cada uno): El gráfico (1) no presenta ni puntos influyentes ni outliers, (2) entre las observaciones figura la indicada con A, que es un outlier, no muy influyente, (3) en este gráfico figura la observación B, influyente pero no outlier, (4) este gráfico muestra la observación C, simultáneamente influyente y atípica.



	2.5 %	97.5 %
(Intercept)	2.089294	10.723305
xx	1.754633	3.042795

---

Ajuste de mínimos cuadrados a los datos de la Figura 29 (2)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.8387	3.6856	2.127	0.049338
xx	2.3281	0.5469	4.257	0.000602

Residual standard error: 5.184 on 16 degrees of freedom  
 Multiple R-squared: 0.5311, Adjusted R-squared: 0.5018  
 F-statistic: 18.12 on 1 and 16 DF, p-value: 0.000602

```
> confint(lm(yy~xx))
      2.5 %    97.5 %
(Intercept) 0.02561661 15.651834
xx          1.16881319  3.487375
```

---

Ajuste de mínimos cuadrados a los datos de la Figura 29 (3)

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	6.2614	1.7778	3.522	0.00283
xx	2.4242	0.2412	10.049	2.57e-08

Residual standard error: 2.9 on 16 degrees of freedom  
 Multiple R-squared: 0.8632, Adjusted R-squared: 0.8547  
 F-statistic: 101 on 1 and 16 DF, p-value: 2.566e-08

```
> confint(lm(yy~xx))
      2.5 %    97.5 %
(Intercept) 2.492573 10.03017
xx          1.912797  2.93559
```

---

Ajuste de mínimos cuadrados a los datos de la Figura 29 (4)

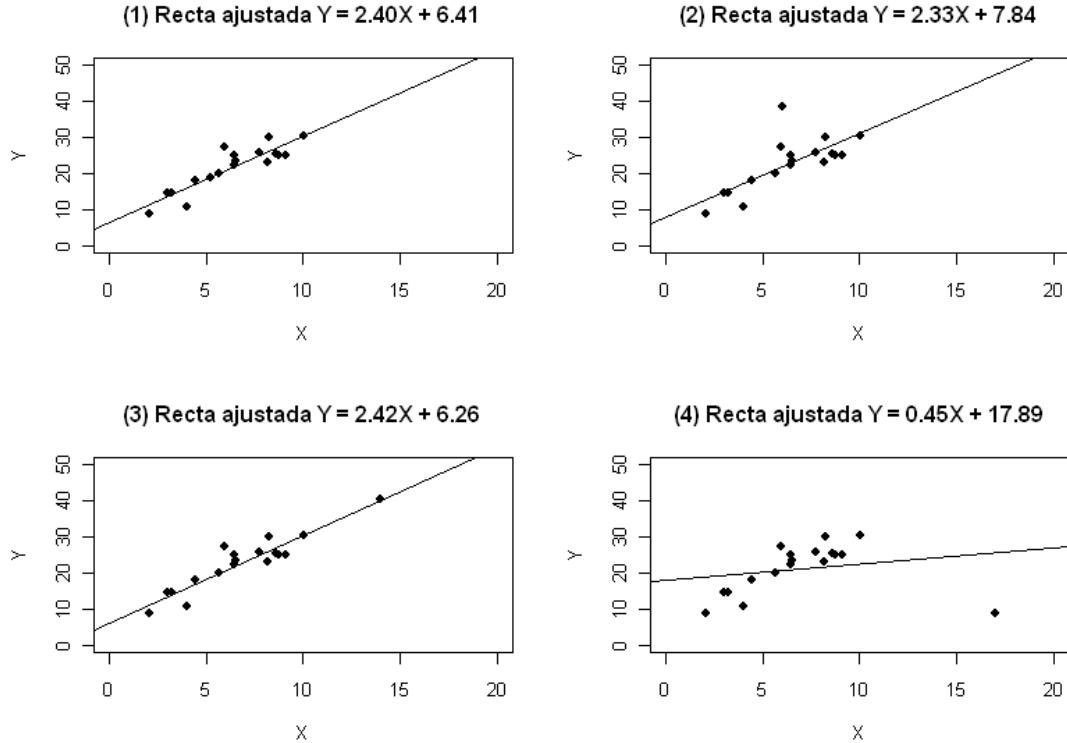
Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	17.8872	3.8042	4.702	0.00024
xx	0.4471	0.4933	0.906	0.37823

Residual standard error: 6.91 on 16 degrees of freedom  
 Multiple R-squared: 0.04883, Adjusted R-squared: -0.01062  
 F-statistic: 0.8214 on 1 and 16 DF, p-value: 0.3782

```
> confint(lm(yy~xx))
      2.5 %    97.5 %
(Intercept) 9.8226420 25.951836
xx          -0.5986414  1.492747
```

Figura 30: Nuevamente los scatter plots de los 4 conjunto de datos, esta vez con las rectas ajustadas.




---

Una vez realizado el ajuste vemos que se verifica lo anticipado. Las pendientes de las rectas estimadas en los 3 primeros gráficos no difieren demasiado entre sí, en el gráfico (2) la ordenada al origen es mayor ya que la observación A está ubicada muy por encima de los datos. La recta estimada en (3) pasa casi exactamente por el dato B y la significatividad del test para la pendiente aumenta en este caso, comparada con la del gráfico (1). Además también se incrementa el R cuadrado, que pasa de 0,79 en (1) a 0,86 en (3). En el gráfico (4) vemos que la recta ajustada difiere completamente de la recta estimada para el conjunto (1), de hecho la pendiente que era significativa para los datos del gráfico (1) deja de serlo en este caso. Vemos que la observación C arrastró la recta hacia ella. La observación C es la que más tergiversó las conclusiones del ajuste lineal.

Un comentario más que habría que hacer con respecto a la influencia es que

en este caso hemos presentado un ejemplo muy sencillo donde para cada conjunto de datos hay un sólo dato sospechoso. En las situaciones prácticas, cuando hay más de un dato anómalo en un conjunto de datos, esta presencia simultánea puede enmascararse: la técnica de sacar las observaciones de a una muchas veces no logra detectar los problemas. En regresión simple nos salva un poco el hecho de que podemos graficar muy bien los datos. No será esta la situación en regresión múltiple, por lo que se vuelve importante tener medidas cuantitativas que permitan medir el grado de influencia (al menos potencial) que tiene cada dato en un conjunto de datos.

### 3.2.4. Alternativa: comparación con un ajuste robusto

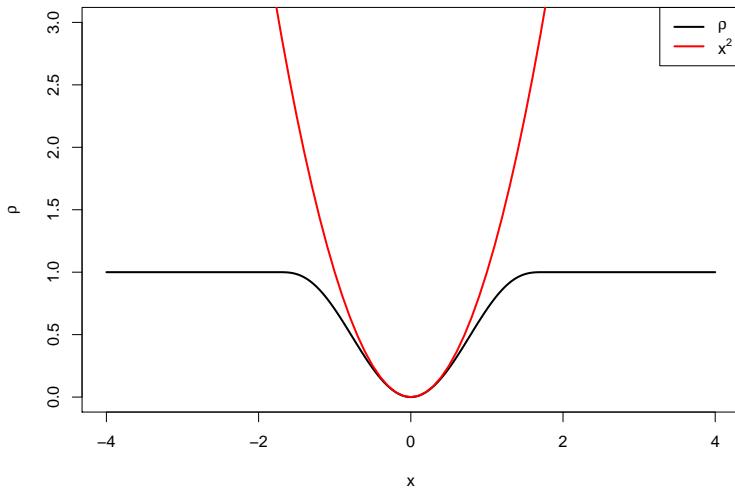
Como hemos dicho antes, el método de cuadrados mínimos como estrategia para encontrar estimadores de los parámetros del modelo lineal, resulta ser muy sensible a observaciones alejadas del resto de los datos. También vimos a través de un ejemplo sencillo, cómo una sola observación apartada del resto puede modificar sustancialmente la estimación realizada. Los métodos de estimación que son más resistentes a la influencia de observaciones apartadas del resto se denominan *métodos robustos de estimación*. De hecho, puede probarse matemáticamente, que la influencia que una sola observación atípica puede tener sobre los estimadores de mínimos cuadrados es (potencialmente) ilimitada. Este déficit no está en el modelo lineal, sino en el método de estimación elegido para ajustarlo: el método de mínimos cuadrados que propone como función para medir el desajuste (se denomina *función de pérdida*) a la suma del cuadrado de los residuos. Si en vez de tomar esa función de pérdida, eligiéramos otra, podríamos subsanar este déficit que tienen los estimadores de mínimos cuadrados, de volverse ilimitadamente sensibles a una observación atípica. ¿Qué forma debería tener la función de pérdida propuesta para que el estimador resultara robusto? Hay varias alternativas. Por un lado, la función debiera ser insensible a observaciones extremadamente grandes (o residuos enormes), y por lo tanto crecer mucho menos que el cuadrado cuando la miramos suficientemente lejos del cero. Por otro lado, si los datos de la muestra siguieran el modelo de regresión con errores normales, querriámos que el estimador que el método robusto calcula se parezca al de mínimos cuadrados, (esta propiedad en estadística se denomina *alta eficiencia*) por lo que la función de pérdida debiera parecerse al cuadrado para valores muy cercanos a cero. Es por esto que en vez de usarse el cuadrado se suele utilizar una función de pérdida del estilo (comparar con la función exhibida en (8))

$$g(a, b) = \sum_{i=1}^n \rho \left( \frac{Y_i - (a + bX_i)}{s_n} \right) \quad (40)$$

donde  $\rho : \mathbb{R} \rightarrow \mathbb{R}$  es una función acotada, creciente y simétrica alrededor del cero, y  $s_n$  es un estimador de escala que juega el papel de  $\sigma$  en el modelo clásico de regresión.

Una posibilidad es ajustar una recta usando un procedimiento de ajuste robusto, por ejemplo un MM-estimador de regresión, propuesto por Yohai [1987]. En R, esto está programado dentro de la rutina `lmrob` en el paquete `robustbase` de R. La estimación se hace en tres etapas, se propone un estimador inicial de los parámetros, a partir de él se estima a  $s_n$  y finalmente se obtienen los estimadores de los parámetros a partir de ellos, minimizando la función objetivo (40). La Figura 31 muestra una posible selección de la función  $\rho$ .

Figura 31: Ejemplo de una función  $\rho$  en la familia biquadrada (en negro) comparada con la cuadrática (en rojo).



Existen muchas otras propuestas de estimadores robustos para regresión, por ejemplo LMS (*least median of squares*), LTS (*least trimmed squares*),  $\tau$ -estimadores de regresión, y casi todas están implementadas en R. Entre ellas, otra buena opción para tener un ajuste robusto altamente robusto y eficiente implementado en R, que no comentaremos aquí, es la rutina `lmRob` del paquete `robust`. Una fuente completa para consultarlas es el libro de Maronna, Martin, y Yohai [2006]. La dificultad con los métodos robustos de ajuste radica en dos cuestiones. En primer lugar ya no es tan fácil (y en algunos casos no es posible) exhibir una fórmula cerrada que los compute, ni encontrar los mínimos absolutos de la función objetivo,

es decir los estimadores. Y por otro, no es sencillo dar la distribución de dichos estimadores, que nos permitirá calcular los p-valores para medir la significatividad de los tests. Sin embargo, utilizando algoritmos diseñados para hallar óptimos de funciones (el IRWLS, por ejemplo) y métodos de remuestreo apropiados, que explotan la capacidad de cómputo de las computadoras actuales, sí pueden obtenerse tests e intervalos de confianza, como veremos en las salidas que presentamos a continuación.

La salida del ajuste con `lmrob` a los datos de la Figura 29 (4) aparece en la Tabla 16. En ella vemos que los valores de la pendiente y ordenada al origen estimados resultan ser muy parecidos a los que se obtienen al ajustar por el método de mínimos cuadrados a los datos (1), que no están contaminados con outliers. Vemos que en lo que a la estimación de los parámetros del modelo lineal se refiere, el método robusto prácticamente ignora a la observación C que estaba distorsionando el ajuste clásico. Y que esto lo hace automáticamente, sin que tengamos que informarle que se trata de una observación potencialmente problemática. Vemos también que el ajuste robusto da p-valores e intervalos de confianza muy parecidos a los que proporciona el ajuste clásico; lo mismo sucede con el cálculo de  $R^2$ .

A posteriori del ajuste robusto, analizando los residuos identificamos rápidamente a la observación C como outlier, ya que el ajuste robusto no arrastra a la recta estimada y la magnitud del residuo refleja la distancia entre el valor observado y el predicho por el modelo, como puede verse en la Figura 32 donde aparecen los boxplots de los residuos y residuos estudentizados del ajuste por mínimos cuadrados, y los residuos del ajuste robusto. En los residuos del ajuste por mínimos cuadrados, no identificamos ninguna observación como outlier. El boxplot de residuos estudentizados muestra la presencia de una observación atípica. Sin embargo, vemos en el boxplot de los residuos del ajuste robusto que el outlier aparece mucho más extremo.

Tabla 16: Ajuste robusto dado por la función `lmrob` del paquete `robustbase`, a los datos de la Figura 29 (4)

```
> library(robustbase)
> summary(lmrob(yy~xx))
\--> method = "MM"

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  6.5147     1.8381   3.544   0.0027 ***
xx          2.3721     0.2676   8.866 1.43e-07 ***

---
Robust residual standard error: 3.149
```

```
Multiple R-squared:  0.7874,           Adjusted R-squared:  0.7741
Convergence in 8 IRWLS iterations
```

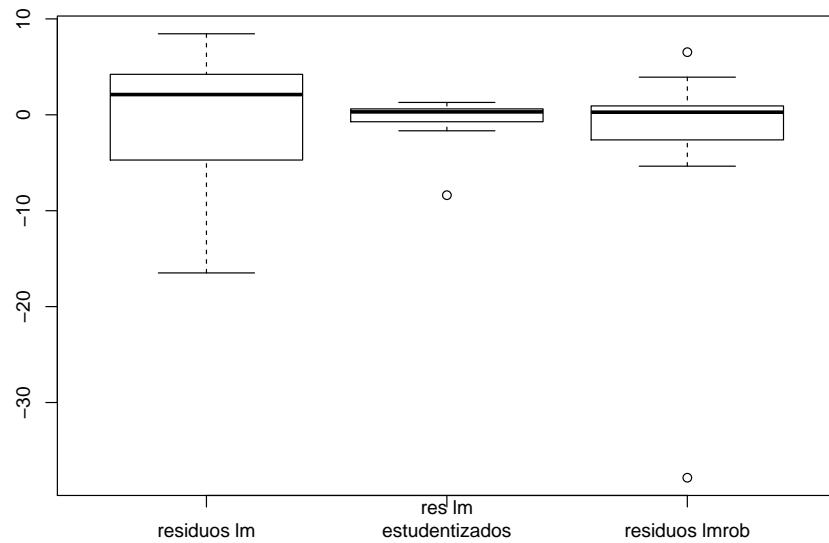
Robustness weights:

```
observation 18 is an outlier with |weight| = 0 (< 0.0056);
3 weights are ~= 1. The remaining 14 ones are summarized as
Min. 1st Qu. Median Mean 3rd Qu. Max.
0.6463 0.9142 0.9431 0.9167 0.9897 0.9959
```

```
> confint(lmrob(yy~xx))
      2.5 %    97.5 %
(Intercept) 2.618108 10.411366
xx          1.804946  2.939334
```

```
> boxplot(residuals(lm(yy~xx)),studres(lm(yy~xx)),
  residuals(lmrob(yy~xx)),names=c("residuos lm","res lm
estudentizados","residuos lmrob"))
```

Figura 32: Boxplot de los residuos para los datos de la Figura 29 (4), a la izquierda los residuos del ajuste por regresión lineal (`lm`), en el centro los residuos estudiantizados del ajuste lineal (`lm`) y a la derecha los residuos del ajuste robusto propuesto (`lmrob`). El ajuste de `lm` arrastra la recta hacia el dato C enmascarando la presencia del outlier. El ajuste del `lmrob`, al no dejarse influenciar por una observación atípica permite identificar un outlier severo al estudiar los residuos.



Una propiedad interesante que tienen los MM estimadores de regresión, es que como parte del ajuste, se computan pesos para las observaciones. Si uno corre el ajuste de *mínimos cuadrados ponderados*, como describimos en la Sección con esos pesos en las observaciones, obtenemos los mismos estimadores robustos, como puede verse comparando las Tablas 16 y Tabla 17. En esta última tabla puede observarse que el peso que el ajuste robusto otorga a cada observación es prácticamente el mismo y casi uno, excepto para la última observación (la C) que recibe peso cero, es decir no interviene en el ajuste. Esta posibilidad de detectar que la observación es atípica de manera automática es muy útil, y lo será aún más cuando en vez de trabajar con una sola variable explicativa, lo hagamos con muchas, y el scatterplot se vuelva una herramienta incompleta.

Aún cuando uno esté interesado solamente en la recta de mínimos cuadrados, de todas formas conviene hacer un ajuste robusto a los datos. Si se observara una fuerte diferencia entre las conclusiones del método clásico (el ajuste de mínimos cuadrados) y el robusto, ésto sólo es señal de que existen observaciones influentes y outliers entre los datos considerados. El estudio de los residuos del ajuste robusto permitirá la identificación de observaciones atípicas.

Tabla 17: Ajuste de mínimos cuadrados pesados a los datos de la Figura 29 (4), con los pesos calculados por el `lmrob`.

```
> ajusro<-(lmrob(yy~xx))
> robpesos<-ajusro$rweights
```

```

> robpesos
      1       2       3       4       5       6
0.9921401 0.9231724 0.9959484 0.9133037 0.9738549 0.9381117
      7       8       9      10      11      12
0.9935094 0.9992255 0.9170598 0.6463435 0.9825457 0.7538439
     13      14      15      16      17      18
0.9921665 0.9994115 0.9481115 0.9996075 0.8634395 0.0000000

> summary(lm(yy~xx,weights=robpesos))

Weighted Residuals:
    Min      1Q  Median      3Q      Max
-4.6502 -2.1647  0.2717  0.9219  5.2523

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.5147     1.9697   3.307  0.00479 **
xx          2.3721     0.2896   8.191 6.44e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.673 on 15 degrees of freedom
Multiple R-squared:  0.8173,    Adjusted R-squared:  0.8051
F-statistic: 67.09 on 1 and 15 DF,  p-value: 6.435e-07

```

### 3.2.5. ¿Cómo medir la influencia de una observación?

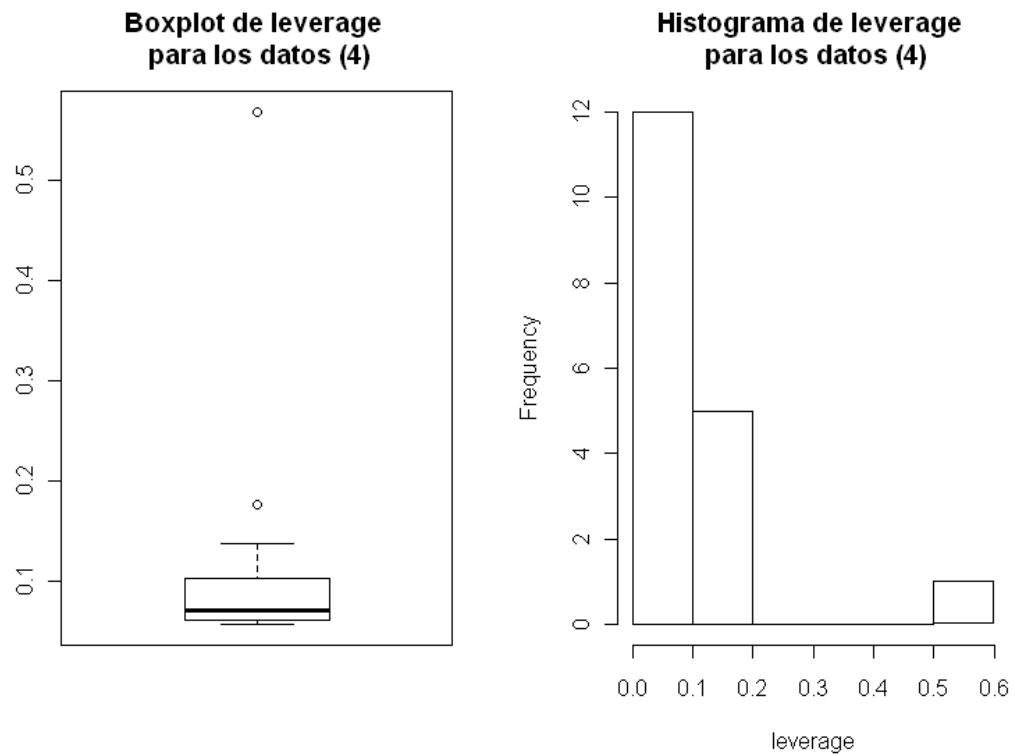
Tenemos dos medidas de influencia: el leverage y las distancias de Cook. El leverage lo definimos en la Sección 3.1.1. Pero cabe preguntarse cuan grande debe ser el leverage de una observación para declararla influyente. Se han sugerido diferentes criterios:

- Teniendo en cuenta que  $\sum_{i=1}^n h_{ii} = 2$  y por lo tanto el promedio  $\bar{h} = \frac{2}{n}$ , un criterio es considerar potencialmente influyentes las observaciones con  $h_{ii} > \frac{4}{n}$ .
- Otro criterio es declarar potencialmente influyentes a aquellas observaciones cuyos leverages  $h_{ii}$  cumplen  $h_{ii} > 0,5$  y evaluar o inspeccionar además los casos en que  $0,2 < h_{ii} \leq 0,5$ .
- Otro criterio es mirar la distribución de los  $h_{ii}$  en la muestra, en especial si

existen saltos en los valores de leverage de las observaciones. El modo más simple de hacerlo es a través de un box-plot o un histograma.

En la Figura 33 se exhiben el boxplot y el histograma de los leverage calculados para los datos de la Figura 29 (4). Hay un único dato con un leverage alto. Observemos que si hicieramos lo mismo para los datos (3) obtendríamos algo muy parecido, ya que el leverage sólo depende de los valores de la covariable (y no de la variable respuesta). En ese sentido es una medida de influencia potencial de los datos. Los leverages para los restantes conjuntos de datos pueden verse en la Figura 34.

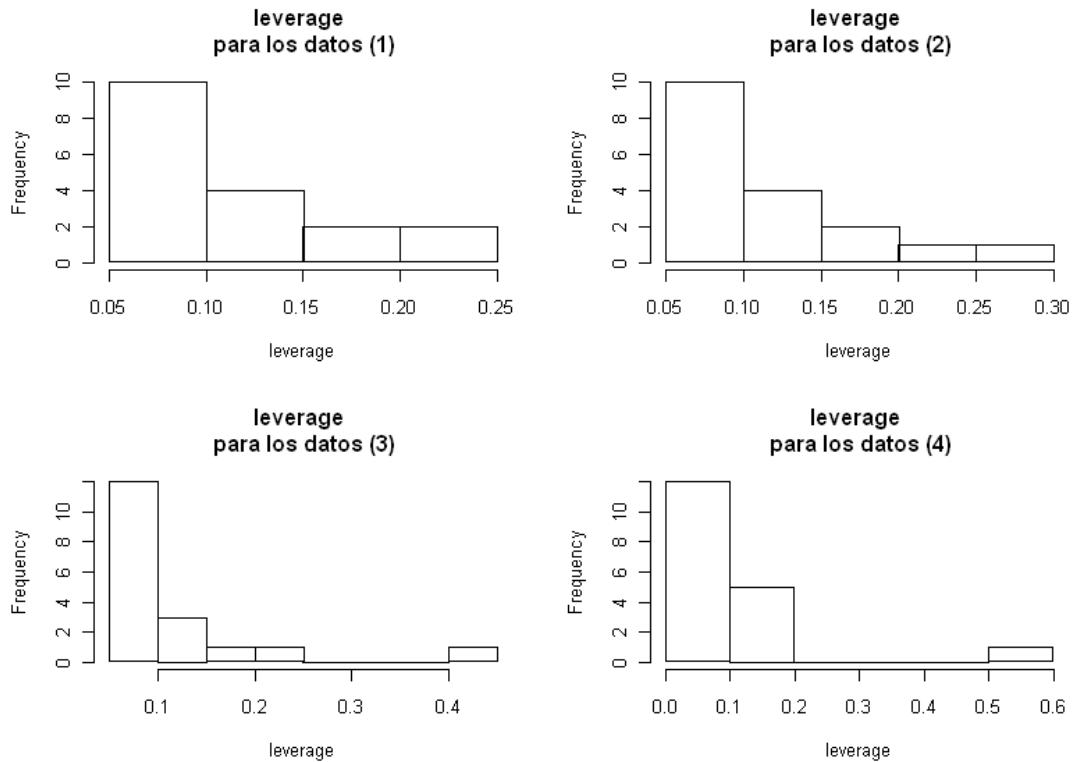
Figura 33: Boxplot e histograma para los leverage de los datos (4) graficados en la Figura 29.



La influencia de una observación depende de dos factores:

1. Cuán lejos cae el valor de  $Y$  de la tendencia general en la muestra para ese valor de  $X$ .

Figura 34: Histogramas de los leverage para los cuatro conjuntos de datos graficados en la Figura 29.



## 2. Cuán lejos se encuentra el valor de la variable explicativa de su media.

El leverage sólo recaba información de la situación descripta en 2). Una medida que toma en cuenta ambas facetas de una observación es la *Distancia de Cook*, definida por

$$D_i = \frac{(\widehat{Y}_{(i)i} - \widehat{Y}_i)^2}{2\widehat{\sigma}^2},$$

donde  $\widehat{Y}_{(i)i}$  corresponde al valor predicho para la  $i$ -ésima observación si se usaron las  $n-1$  restantes observaciones para hacer el ajuste, como lo habíamos definido en la Sección 3.2.2 y  $\widehat{Y}_i$  es el valor predicho para la  $i$ -ésima observación en el modelo ajustado con las  $n$  observaciones. Como en el caso de los residuos estudentizados, no es necesario recalcular el ajuste por mínimos cuadrados para calcular los  $D_i$ ,

ya que otra expresión para ellos es la siguiente

$$D_i = \frac{1}{2} (rest_i)^2 \frac{h_{ii}}{1 - h_{ii}}.$$

La distancia de Cook se compara con los percentiles de la distribución  $F$  de Fisher con  $2$  y  $n - 2$  grados de libertad en el numerador y denominador, respectivamente ( $2$  porque estamos estimando dos parámetros beta). El criterio para decidir si una observación es influyente es el siguiente:

- Si  $D_i <$  percentil  $0,20$  de la distribución  $F_{2,n-2}$  entonces la observación no es influyente.
- Si  $D_i >$  percentil  $0,50$  de la distribución  $F_{2,n-2}$  entonces la observación es muy influyente y requerirá tomar alguna medida.
- Si  $D_i$  se encuentra entre el percentil  $0,20$  y el percentil  $0,50$  de la distribución  $F_{2,n-2}$  se sugiere mirar además otros estadísticos.

Volviendo a los datos de la Figura 29, el percentil  $0,20$  de la distribución  $F_{2,16}$  es  $0,226$  y el percentil  $0,50$  de la distribución  $F_{2,16}$  es  $0,724$ . Los histogramas de las distancias de Cook calculadas en este caso están en la Figura 35. Vemos que sólo en el caso de los datos (4) aparece una observación (la C) cuya distancia de Cook supera al percentil  $0,50$  de la distribución de Fisher, indicando que hay una observación muy influyente.

Existen otras medidas de influencia. Los DFfits y los DFbetas son medidas bastante estudiadas. Una referencia para leer sobre ellos es el libro de Kutner et al. [2005]. Los gráficos de variables agregadas (en el caso de regresión múltiple) pueden servir también para identificar observaciones influyentes, pueden verse en Weisberg [2005] secciones 3.1 y 9.2.4 o Kutner et al. [2005] sección 10.

### 3.2.6. Instrucciones de R para diagnóstico

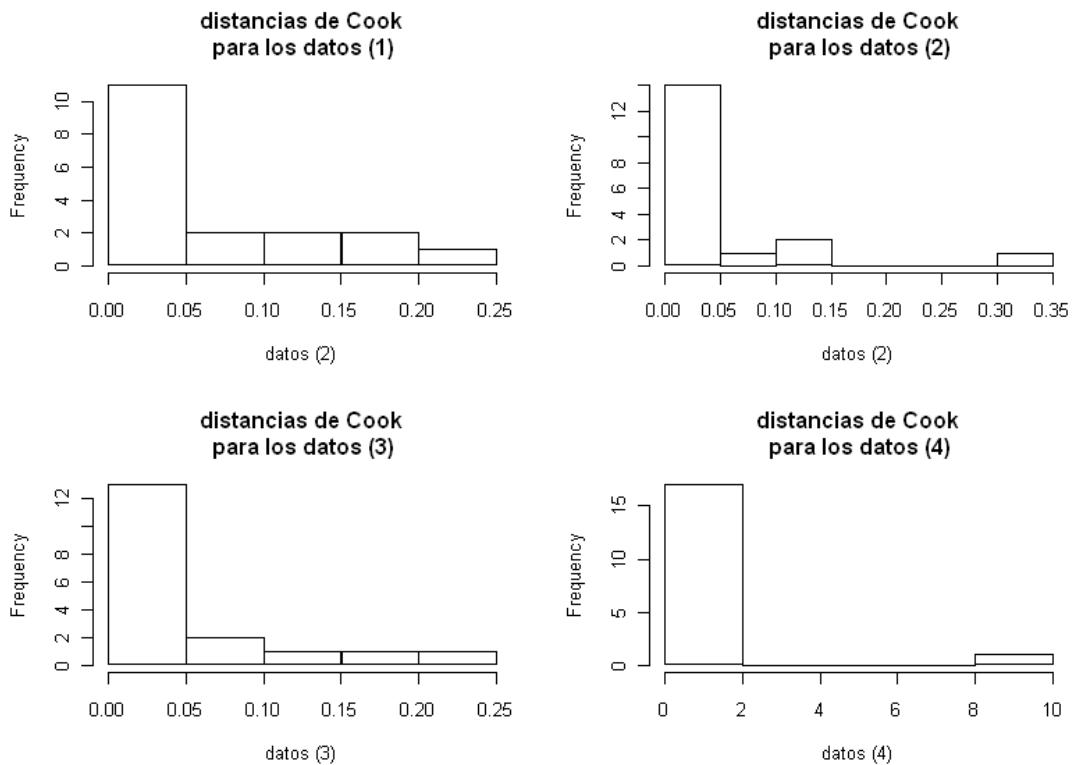
```
ajuste <- lm(yy ~ xx)

#residuos
rr1<-residuals(ajuste)

#residuos estandarizados
rr2<-rstandard(ajuste)

#residuos estudentizados
rr3<-rstudent(ajuste)
```

Figura 35: Histogramas de las distancias de Cook para los datos de la Figura 29



```
# predichos o valores ajustados
ff1<-fitted(ajuste)

# grafico de residuos estudentizados vs predichos
plot(ff1,rr3,xlab="predichos",ylab="residuos")
abline(0,0)
title("Residuos estudiantizados vs predichos")

#####
# test para encontrar outliers
ene<-length(rr3)
corte<-qt(1-0.05/(2*ene),df=ene-3)
(rr3 > corte)
sum(rr3 > corte)
```

```

#cuales son?
(1:ene)[rr3 > corte]

#p-valor, sin corregir
2*(1-pt(abs(rr3[3]),df=ene-3))
#####
#####

#leverage
hatvalues(ajuste)
hist(hatvalues(ajuste))

#para ver cuales son mayores que 0.2
lev <- hatvalues(ajuste)

#un criterio
(1:ene)[lev > 0.2]

#un criterio mas exigente
(1:ene)[lev > 4/ene]

#distancias de cook
dcook <- cooks.distance(ajuste)
hist(dcook)

#punto de corte
corted <- qf(0.5,2,ene-2)
(1:nn)[dcook > corted]

# qq plot de los residuos estandarizados, con la normal
qqnorm(rr2)
qqline(rr2) #le agrega una recta que pasa por cuartil 1 y 3

plot(ajuste) #hace varios graficos, entre ellos el QQplot de los
#residuos estandarizados

library(car)
outlierTest(ajuste) # da pval de Bonferoni para obs mas extremas
qqPlot(ajuste, main="QQ Plot") #qq plot de los resid studentizados,
#con la t de student apropiada

```

```
#####
# ajuste robusto

library(robustbase)
ajusterob <- lmrob(yy ~ xx)
summary(ajusterob)

#residuos
resrob <- residuals(ajusterob)

boxplot(resrob,residuals(ajuste))

#pesos
hist(ajusterob$rweights)
boxplot(ajusterob$rweights)

plot(ajusterob)      # hace varios graficos
```

### 3.3. Ejercicios

Estos ejercicios se resuelven con el archivo `script_diagnostico.R`

**Ejercicio 3.1** *Madres e hijas II.* Archivo de datos `heights.txt` del paquete `alr3`. Continuando con el ejercicio 2.6, en el que proponemos ajustar el modelo lineal simple para explicar la altura de la hija, `Dheight`, a partir de la altura de la madre, llamada `Mheight`, como la variable predictora.

- (a) Hacer gráficos para evaluar la adecuación del modelo lineal para explicar los datos.
- (b) Compare el ajuste clásico con el ajuste robusto propuesto.
- (c) Concluya respecto de la adecuación del modelo lineal en este caso.

**Ejercicio 3.2** *Medidas del cuerpo V.* Base de datos `bdims` del paquete `openintro`.

- (a) Realice gráficos de que le permitan evaluar los ajustes realizados en los ejercicios 2.1 y 2.2 con esta base de datos, tanto para explicar el peso por el contorno de cintura como el ajuste para explicar el peso por la altura. ¿Lo conforman estos modelos ajustados?

- (b) Compare el ajuste clásico del modelo lineal con el ajuste robusto. ¿Cambian mucho los modelos ajustados? ¿Qué indica esto? No se desanime, este ejercicio sigue en el capítulo próximo.

**Ejercicio 3.3** Mamíferos, Parte V. Base de datos *mammals* del paquete *openintro*.

- (a) En el ejercicio 1.7 observamos que el scatter plot del peso del cerebro de un mamífero (*BrainWt*) en función de su peso corporal (*BodyWt*) no se podía describir como una pelota de rugby más o menos achatada. Supongamos que no hubiéramos hecho el gráfico de dispersión, e intentemos ajustar un modelo lineal a los datos. Ajuste el modelo lineal simple que explica *BrainWt* en función de *BodyWt*. Luego realice el gráfico de residuos versus valores predichos. El gráfico de residuos estandarizados versus valores predichos. El de residuos estudentizados versus valores predichos. ¿Difieren mucho entre sí?
- (b) Use el test de outliers basado en los residuos estudiantizados. Indique cuáles son las observaciones candidatas a outliers.
- (c) Calcule los leverages. Identifique las observaciones candidatas a más influyentes según este criterio. Calcule las distancias de Cook, vea cuáles son las observaciones influyentes.
- (d) Compare con el ajuste robusto.
- (e) Finalmente, para el modelo de regresión propuesto en el ejercicio 2.9 para vincular los logaritmos en base 10 de ambas variables, haga un gráfico de residuos versus valores predichos, y algunos otros gráficos de diagnóstico. ¿Le parece que este modelo ajusta mejor a los datos?

**Ejercicio 3.4** Hacer un ajuste robusto a los datos de perímetro cefálico y edad gestacional. Comparar con el ajuste clásico. Identificar la presencia de outliers. ¿Son muy influyentes en el ajuste? Recordar que de todos modos este no es el último modelo que probaremos sobre estos datos.

**Ejercicio 3.5** Resuelva el ejercicio domiciliario que figura en el Apéndice A.

**Ejercicio 3.6** Resuelva el Taller 2 que figura en el Apéndice A.