Enfoque estadístico del aprendizaje

Robustez en Regresión Lineal

María Eugenia Szretter Noste

octubre de 2019

Robustez en Regresión Lineal

Basado en el libro

Ricardo A. Maronna, R. Douglas Martin, Victor J. Yohai& Matías Salibián-Barrera (2019) *Robust Statistics: Theory and Methods (with R)*, 2nd Edition, John Wiley Sons.

Al final de cada capítulo, el libro trae una sección *Recommendations and software* que indica procedimientos recomendados por los autores y las funciones de R que los implementan.

Outliers y observaciones influyentes

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 dato (o pequeño conjunto de datos) muy alejados de lo que el modelo le prescribe se denomina outlier. Observemos que el concepto de outlier (o dato atípico) es un concepto relativo al modelo específico en consideración.

Problema: Puede modificar completamente el ajuste, distorsionando las conclusiones. 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.

Outliers y observaciones influyentes: ejemplo *de juguete*

Veámoslo a través de un ejemplo. "Fabricamos" (generamos) 11 observaciones que siguen el modelo lineal simple:

$$Y_i = 6 + X_i + \varepsilon_i$$

con $1 \leq i \leq n = 11$. ¿Cuánto valen β_0 yβ_1 en este modelo? Los valores observados son

X_i	Y_i	X_i	Y_i	X_i	Y_i	X_i	Y_i
4	9.40	7	13.13	10	15.30	13	19.23
5	11.27	8	13.88	11	16.89	14	21.03
6	11.97	9	14.50	12	18.14		

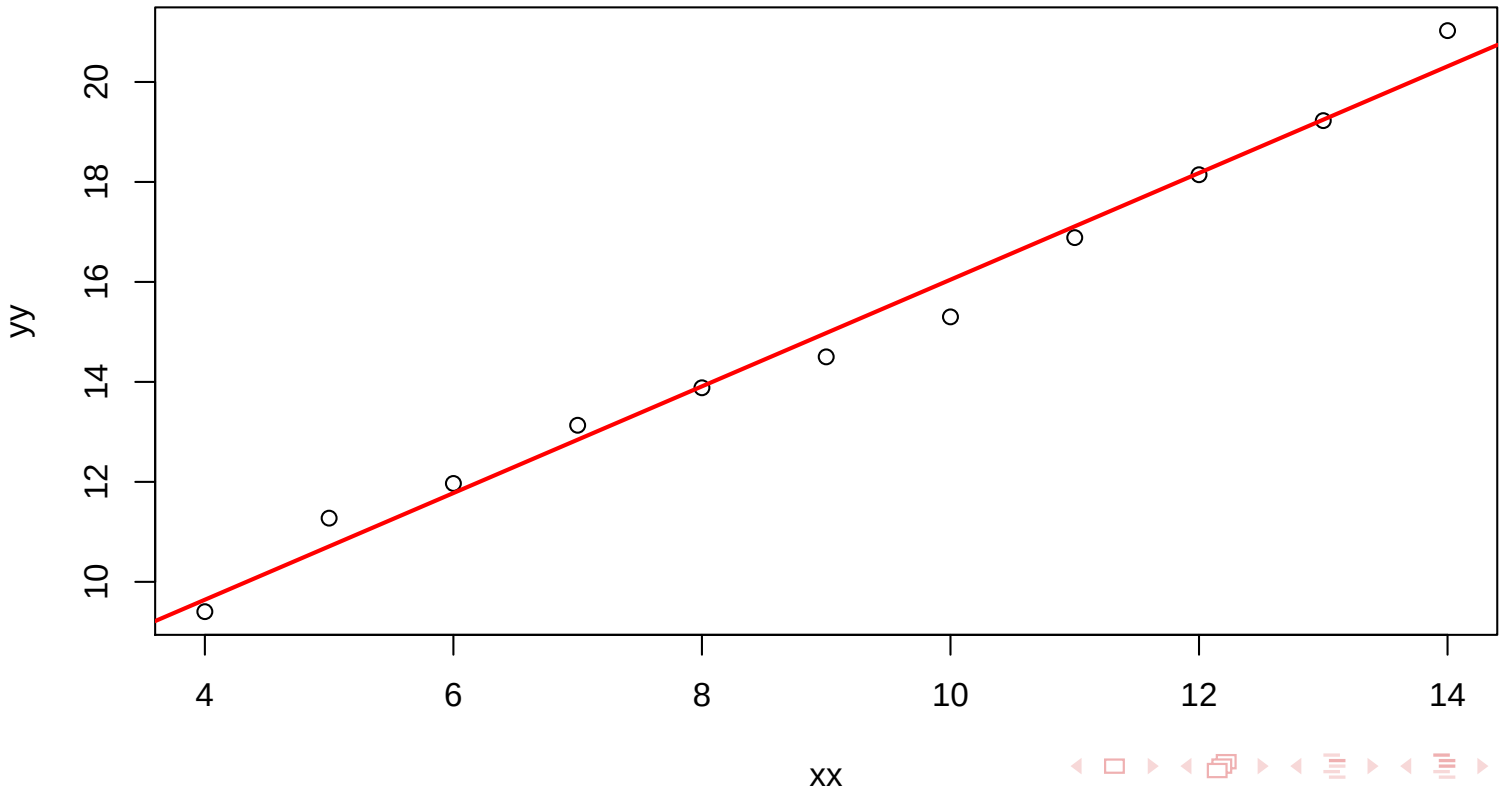
Ajuste OLS (*ordinary least squares*) en R

```
> set.seed(2345)
> xx<-4:14
> yy<- xx + 6 + rnorm(11,sd=0.5)
> ajuste1<-lm(yy ~ xx)
> summary(ajuste1)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.37349    0.41130   13.06 3.72e-07 ***
xx           1.06711    0.04312   24.75 1.38e-09 ***
---

Residual standard error: 0.4522 on 9 degrees of freedom
Multiple R-squared:  0.9855,Adjusted R-squared:  0.9839
F-statistic: 612.5 on 1 and 9 DF,  p-value: 1.377e-09
```

Scatter plot con recta ajustada por OLS

Figura 1: Datos originales con la recta de mínimos cuadrados (OLS) ajustada. Se observa una relación positiva entre las variables ($\widehat{\beta}_1 > 0$), que resulta significativa (su p–valor es menor a 0.05).



Datos contaminados

¿Qué pasa con el ajuste si cambiamos (por error) la última observación por el par $(37\,, 0{,}1)$?

¿Cómo cambiará la recta ajustada por mínimos cuadrados?

Ajuste OLS a los datos contaminados

```
> #contaminamos el ultimo dato
> yy[11]<-0.1
> xx[11]<-37
> ajuste2<-lm(yy ~ xx)
> summary(ajuste2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  16.9346     2.1487   7.881 2.49e-05 ***
xx           -0.3480     0.1528  -2.277   0.0488 *
---

Residual standard error: 4.379 on 9 degrees of freedom
Multiple R-squared:  0.3655,Adjusted R-squared:  0.295
F-statistic: 5.184 on 1 and 9 DF,  p-value: 0.04881
```

Ajuste OLS a los datos contaminados: scatterplot

Figura 2: Datos con un outlier. Recta de mínimos cuadrados (OLS) ajustada, ahora la relación que se observa entre las variables es negativa, ($\widehat{\beta}_1 < 0$), que también resulta significativa (su p–valor es menor a 0.05).

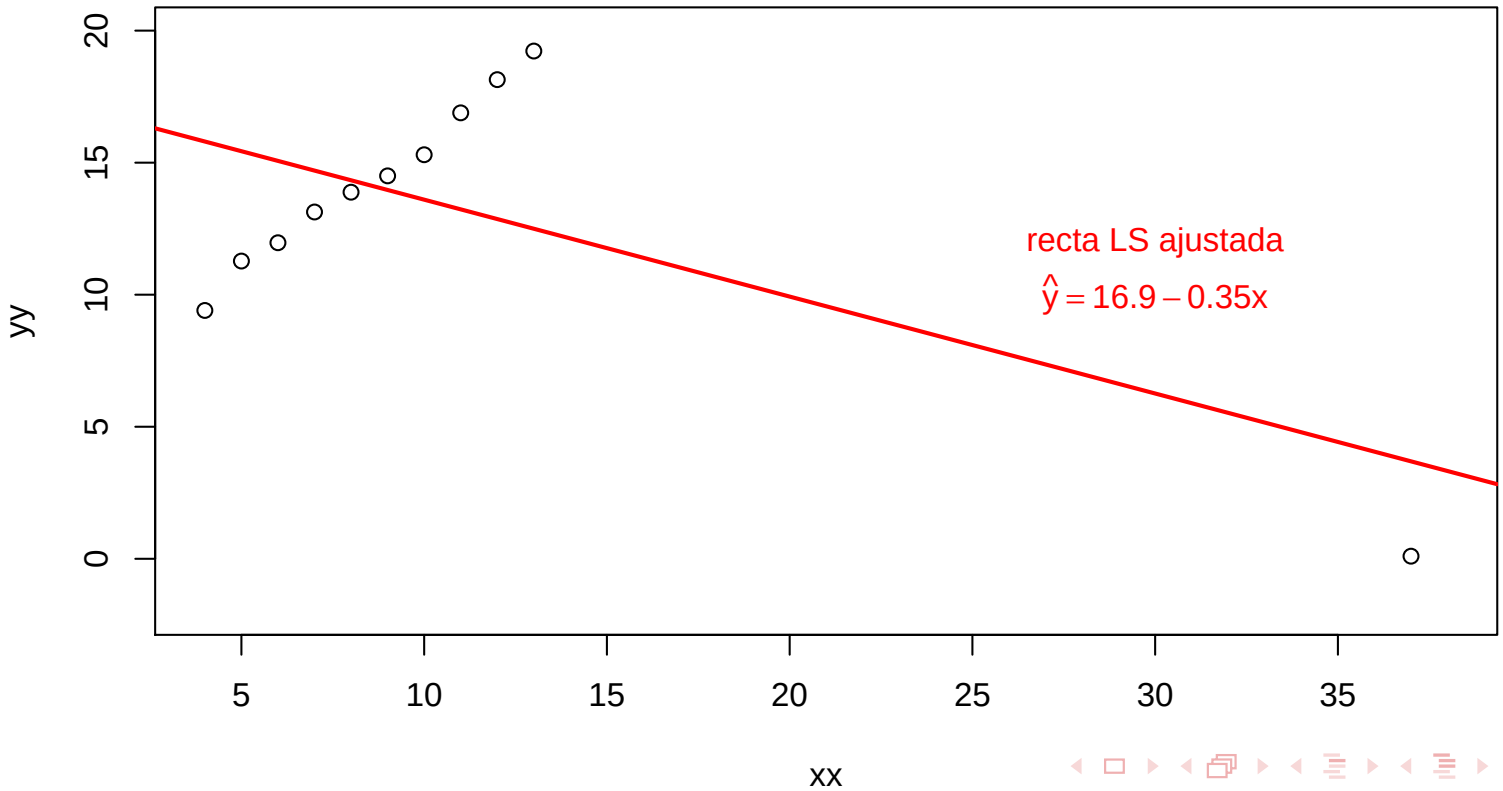
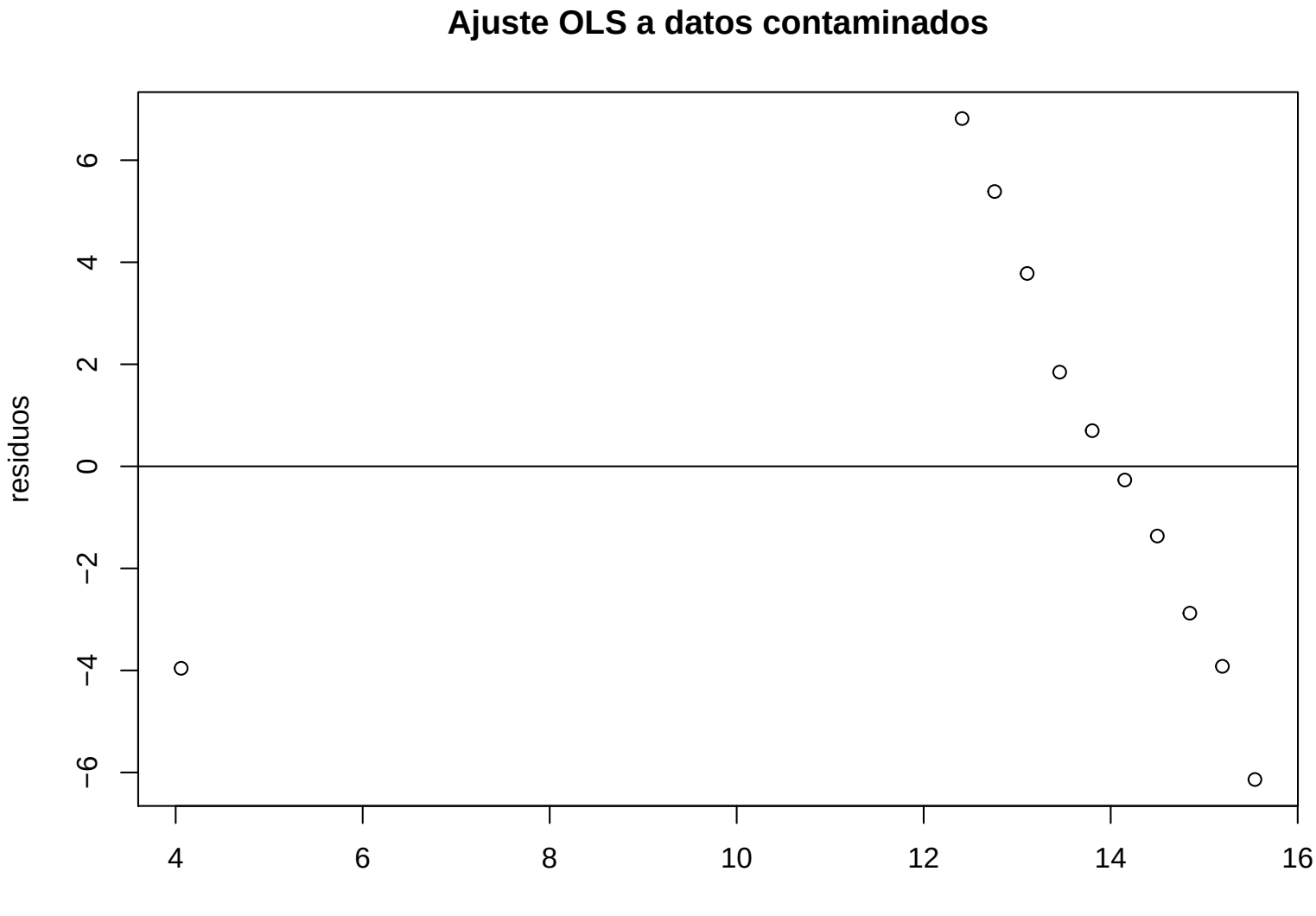


Gráfico de residuos vs predichos

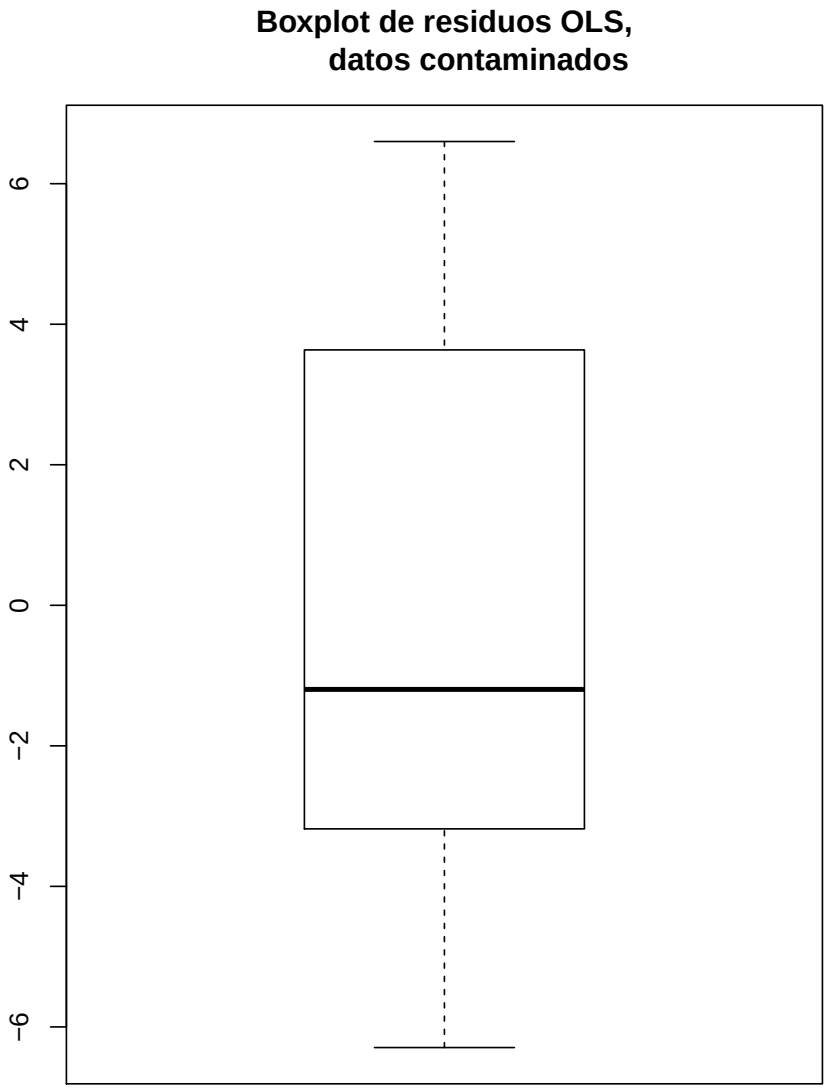
¿Nos daríamos cuenta del problema del ajuste OLS viendo el residual plot?
¿Identificaríamos al outlier como la observación asociada al mayor residuo?



¡No!

Boxplot de los residuos

¿Nos daríamos cuenta del problema del ajuste OLS estudiando los residuos?



Outliers y observaciones influyentes: solución, un ajuste robusto

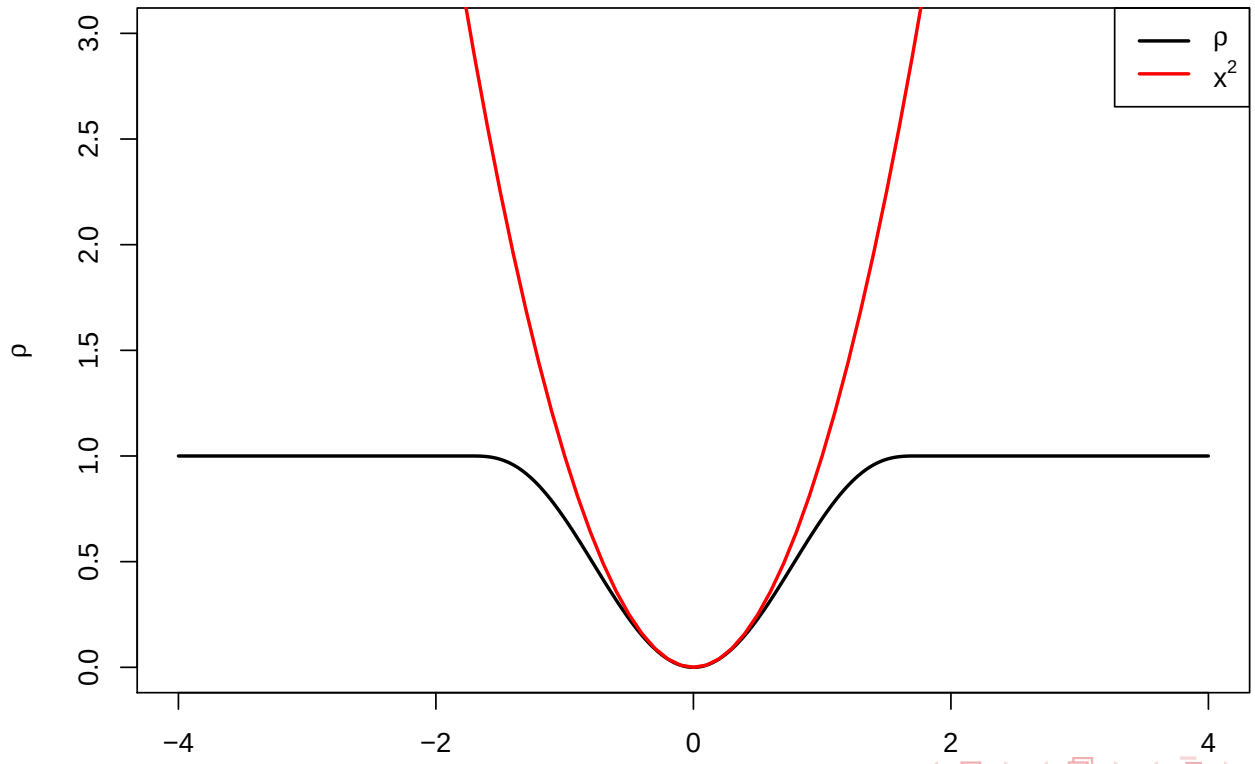
El método de cuadrados mínimos como estrategia para encontrar estimadores de los parámetros del modelo lineal, resulta ser muy sensible a observaciones alejadas del resto de los datos. Es por esto que en vez de usarse el cuadrado se puede utilizar una función de pérdida del del estilo

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

donde ρ tiene una forma del estilo dado en el gráfico que sigue.

Outliers y observaciones influyentes: solución, un ajuste robusto

Figura 3: Ejemplo de una función ρ en la familia bicuadrada (en negro) comparada con la cuadrática (en rojo).



M-estimadores robustos de regresión

El enfoque para controlar la influencia de observaciones atípicas que pueden estar tanto en la variable respuesta como en las covariables consiste en definir al M-estimador $\widehat{\boldsymbol{\beta}}$ como aquellos valores que satisfacen

$$\min_{\beta_0,\beta_1} \sum_{i=1}^{n} \rho\left(\frac{Y_i - (\beta_0 + \beta_1 \cdot X_i)}{\widehat{\sigma}}\right) = \sum_{i=1}^{n} \rho\left(\frac{Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot X_i\right)}{\widehat{\sigma}}\right) \qquad (2)$$

donde ρ es una función que se comporta como el cuadrado en una vecindad del cero, pero que es acotada, y $\widehat{\sigma}$ es una escala previamente calculada. La escala deberá cumplir ciertos requisitos. Si ρ se tomara como la cuadrátrica, $\rho(t) = t^2$ obtendríamos los LS.

Si ρ tiene derivada ψ, (2) es equivalente a hallar $\widehat{\boldsymbol{\beta}}$ que resuelva

$$\sum_{i=1}^{n} \psi\left(\frac{Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot X_i\right)}{\widehat{\sigma}}\right) x_i = 0 \text{ y } \sum_{i=1}^{n} \psi\left(\frac{Y_i - \left(\widehat{\beta}_0 + \widehat{\beta}_1 \cdot X_i\right)}{\widehat{\sigma}}\right) = 0$$

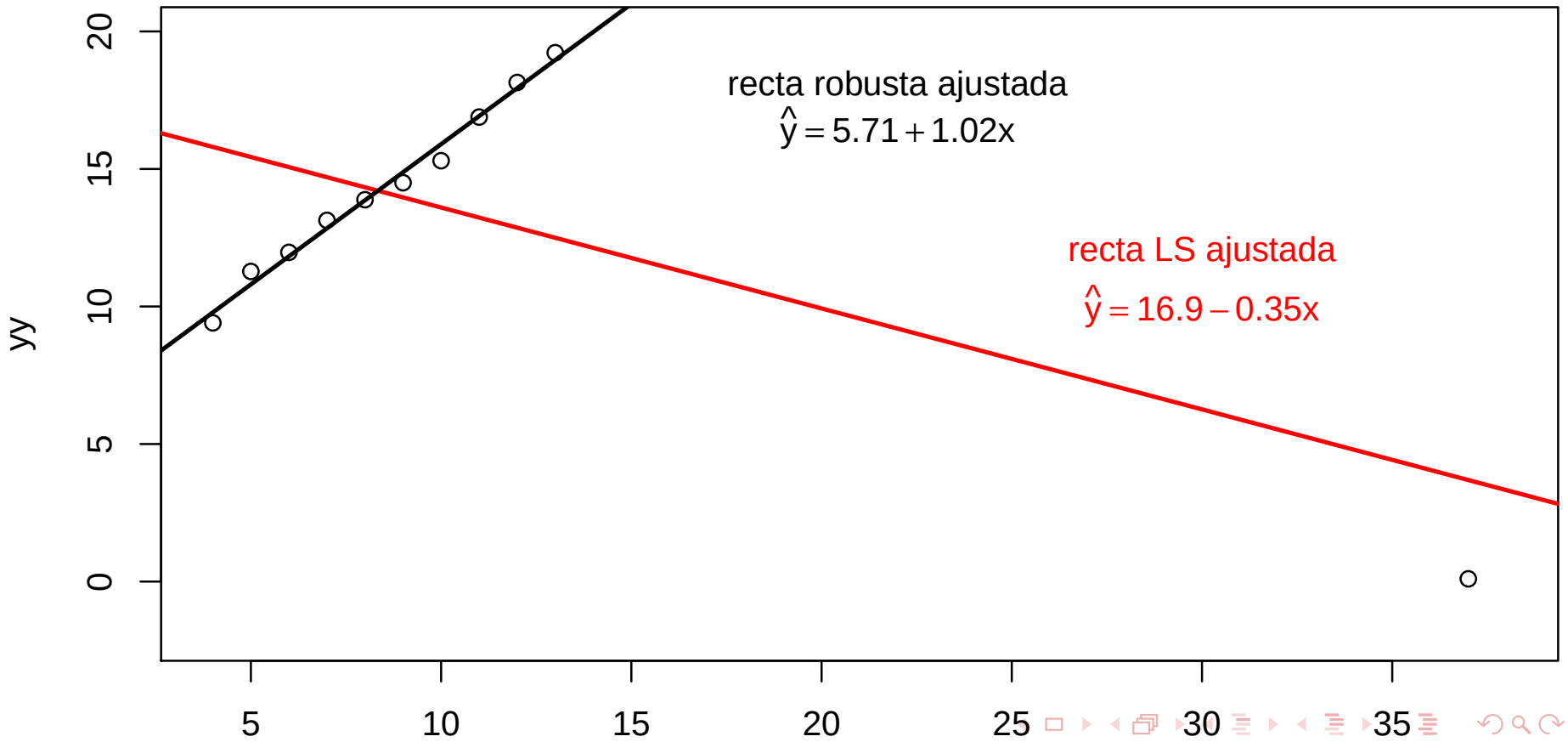
Estimador recomendado y rutina de R

El estimador resultante se denomina MM-estimador de regresión, y fue propuesto por Yohai en 1987.

En el libro de Maronna, Martin, Salibian-Barrera, Yohai recomiendan usar el MM-estimador de regresión (con función ρ que se denomina óptima y estimador inicial Peña-Yohai) que está implementado en la rutina `lmrobdetMM` (en la librería `RobStatTM` del R). Y también (con otros estimadores iniciales u otros algoritmos para hallar la raíz) en las rutinas `lmrob` (de la librería `robustbase`) y `lmRob` (de la librería `robust`).

Ajuste robusto a los datos contaminados

Mismos datos con un outlier. Ajuste de mínimos cuadrados y ajuste robusto. Rutina `lmrobdetMM`, del paquete o librería `RobStatTM`



Instrucciones de R para el ajuste robusto

Comparemos con el ajuste robusto

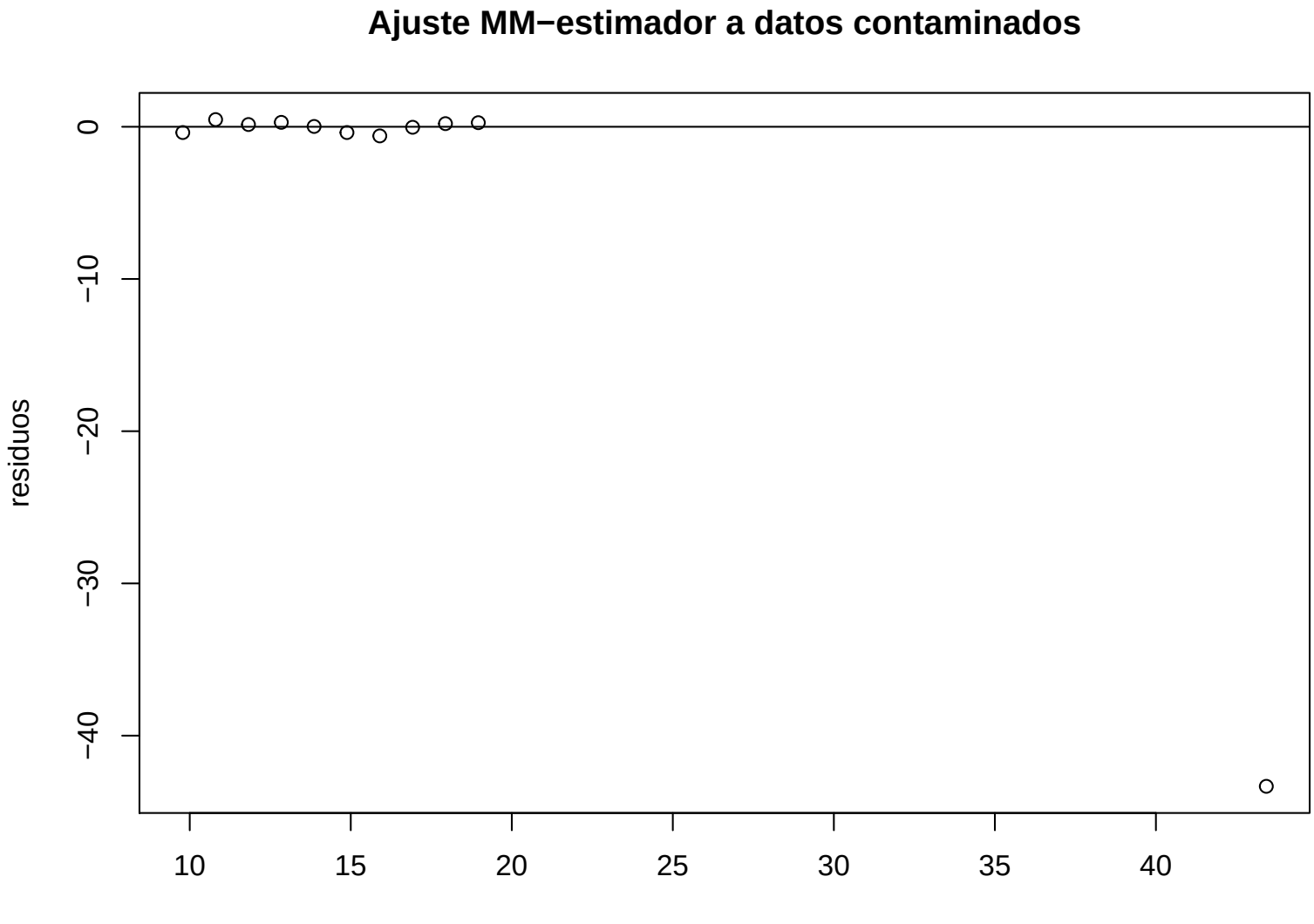
```
> library(RobStatTM)
> ajusterob<-lmrobdetMM(yy ~ xx)
> summary(ajusterob)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.70609    0.36186   15.77 7.30e-08 ***
xx           1.01952    0.04036   25.26 1.15e-09 ***
---

Robust residual standard error: 0.5229
Multiple R-squared:  0.9551,Adjusted R-squared:  0.9501
Convergence in 4 IRWLS iterations
```

Gráfico de residuos vs predichos, MM-estimador

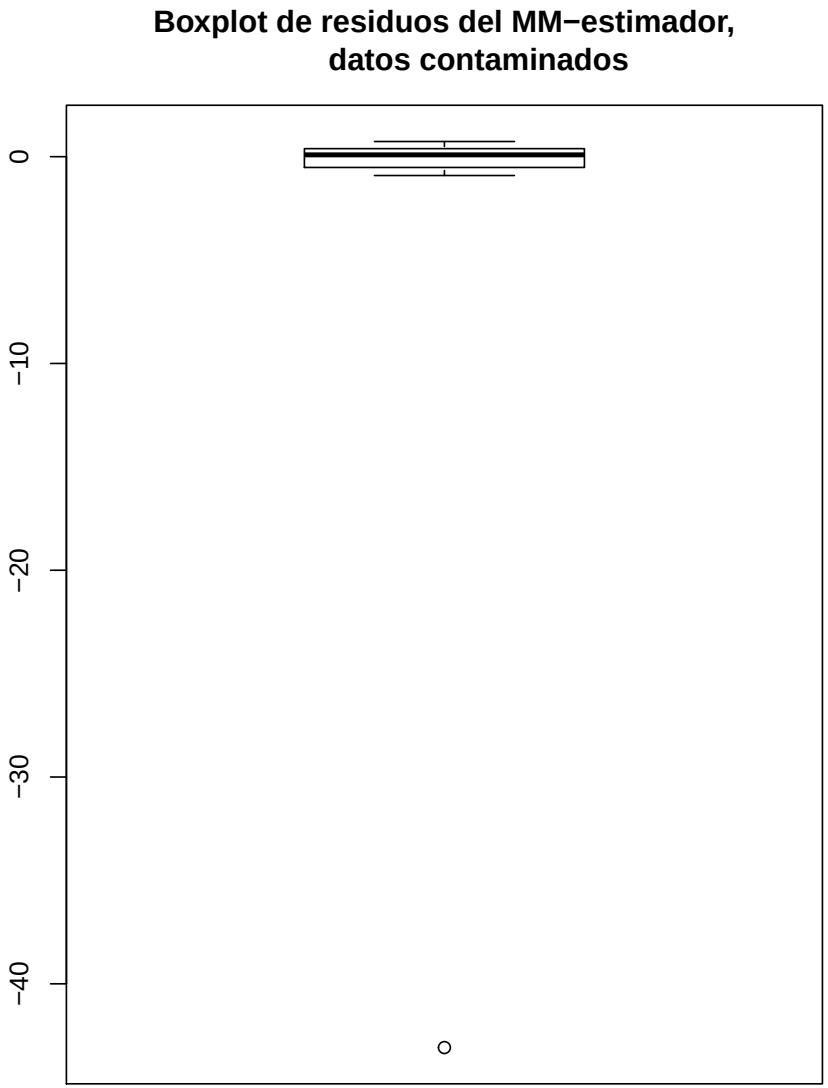
¿Qué pasa con este gráfico para el ajuste robusto? ¿Identificaríamos al outlier como la observación asociada al mayor residuo?



¡Sí!

Boxplot de los residuos del ajuste robusto

¿Detectaríamos la observación atípica estudiando los residuos robustos?



Para qué sirve un ajuste robusto

Vemos que el ajuste LS se ve completamente confundido por los outliers en este ejemplo, el LS proporciona un ajuste equivocado a los datos. Los outliers enmascaran las conclusiones del LS. Un ajuste robusto permite identificar la presencia de outliers en la muestra.

Si para un conjunto de datos las conclusiones del ajuste LS difieren mucho de las que proporciona un ajuste robusto, esto es una indicación de la presencia de observaciones atípicas.

La forma de identificar a las observaciones atípicas es a través de un ajuste robusto.

Ante un nuevo conjunto de datos, siempre conviene comparar las conclusiones de un ajuste robusto con el ajuste clásico. Si no hay diferencias, continuar con el clásico (OLS), que es más fácil de comunicar. Si las hay, identificar las observaciones atípicas y estudiarlas a fondo.

Datos contaminados, pesos obtenidos con el MM

Se prueba que los MM-estimadores se pueden calcular como los estimadores de mínimos cuadrados con pesos dados a las observaciones, que dependen del grado de atipicidad que tienen. Para eso, el ajuste robusto calcula pesos como subproducto. Veamos los pesos que el MM-estimador le da a cada observación

```
> ajusterob$rweights
1         2         3         4         5         6         7
0.9993162 0.9999608 0.9879342 0.9975451 0.8577278 0.9993229 0.99962
8         9        10        11
0.9185199 0.9936962 0.9968105 0.0000000
```

Las observación 11 recibe peso cero y no influye en la estimación.

Datos contaminados, ajuste LS con pesos robustos calculados con MM

Veamos que ajustando por mínimos cuadrados con los pesos obtenidos por el MM-estimador da lo mismo que el ajuste del MM-estimador

```
> #ajuste LS con pesos
> ajusteLSconpesos <- lm(yy ~ xx,
+                        weights = ajusterob$rweights)
> summary(ajusteLSconpesos)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.70609    0.36612   15.59 2.86e-07 ***
xx           1.01952    0.04083   24.97 7.08e-09 ***
---

Residual standard error: 0.3691 on 8 degrees of freedom
Multiple R-squared:  0.9873,Adjusted R-squared:  0.9857
F-statistic: 623.5 on 1 and 8 DF,  p-value: 7.08e-09
```